SIAM J. MATH. ANAL.
Vol. 36, No. 1, pp. 1–37

A GENERAL INTEGRAL REPRESENTATION RESULT FOR CONTINUUM LIMITS OF DISCRETE ENERGIES WITH SUPERLINEAR GROWTH*

ROBERTO ALICANDRO† AND MARCO CICALESE‡

Abstract. We study the asymptotic behavior, as the mesh size ε tends to zero, of a general class of discrete energies defined on functions $u : \alpha \in \varepsilon\mathbb{Z}^N \cap \Omega \mapsto u(\alpha) \in \mathbb{R}^d$ of the form

$$F_\varepsilon(u) = \sum_{\substack{\alpha,\beta\in\varepsilon\mathbb{Z}^N \\ [\alpha,\beta]\subset\Omega}} g_\varepsilon(\alpha,\beta,u(\alpha)-u(\beta))$$

and satisfying superlinear growth conditions. We show that all the possible variational limits are defined on $W^{1,p}(\Omega;\mathbb{R}^d)$ of the local type

$$\int_\Omega f(x,\nabla u)\,dx.$$

We show that, in general, f may be a quasi-convex nonconvex function even if very simple interactions are considered. We also treat the case of homogenization, giving a general asymptotic formula that can be simplified in many situations (e.g., in the case of nearest neighbor interactions or under convexity hypotheses).

Key words. discrete systems, homogenization, Γ-convergence

AMS subject classifications. 49J45, 74Q99

DOI. 10.1137/S0036141003426471

1. Introduction.

The energetic description of the asymptotic behavior of lattice systems when the mesh size tends to zero turns out to be useful both as a microscopical theoretical justification of theories in continuum mechanics and as a powerful means, thanks to which a great number of microscopical phenomena can be read in the macroscopical setting. In this paper we describe variational limits of discrete lattice systems in a vectorial and nonconvex setting when general "atomic" interaction energies are taken into account that lead to continuum "elastic" theories described by bulk integral energies. We will limit our analysis to square lattices, but more general geometries, e.g., hexagonal lattices, can be easily included in this framework by a change of variables (see, for instance, [10, Examples 5.1 and 5.2], for details). In mathematical terms, given a fixed open set $\Omega \subset \mathbb{R}^N$ and $\varepsilon > 0$, we consider energies defined on functions $u : \alpha \in \varepsilon\mathbb{Z}^N \cap \Omega \mapsto u(\alpha) \in \mathbb{R}^d$ of the general form

$$F_\varepsilon(u) = \sum_{\substack{\alpha,\beta\in\varepsilon\mathbb{Z}^N \\ [\alpha,\beta]\subset\Omega}} g_\varepsilon(\alpha,\beta,u(\alpha)-u(\beta)).$$

In the case $N = d = 3$ we can picture the lattice $\varepsilon\mathbb{Z}^N \cap \Omega$ as the reference configuration

*Received by the editors April 23, 2003; accepted for publication (in revised form) November 7, 2003; published electronically June 22, 2004.
http://www.siam.org/journals/sima/36-1/42647.html

†D.A.E.I.M.I, Universita' di Cassino, Via Di Biasio 03043 Cassino (FR), Italy (alicandr@unicas.it).
‡Dipartimento di Matematica "R. Caccioppoli," Universita' di Napoli, Via Cintia 80126 Napoli, Italy (cicalese@unina.it).

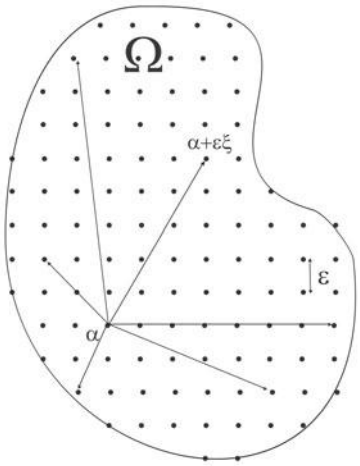

FIG. 1. *Interactions on the lattice* $\varepsilon\mathbb{Z}^N$.

of a set of interacting material points (see Figure 1). Here u is the field mapping the reference configuration into the deformed one; thus the total stored energy $F_\varepsilon(u)$ is obtained, according to the classical theory of crystalline structures in "hyperelastic" regime, by the superposition of the energy densities $g_\varepsilon(\alpha, \beta, u(\alpha) - u(\beta))$ weighing the pairwise interaction between points in positions α and β in the reference configuration lattice. Note that the only assumption we make is that g_ε depends on the displacement field in α and β through the differences $u(\alpha) - u(\beta)$. This condition, expressing the invariance under translation of our energies, arises naturally in many situations, as, for example, in frame indifferent models.

It is usually more convenient to group the energy densities as

$$F_\varepsilon(u) = \sum_{\xi\in\mathbb{Z}^N} \sum_{\alpha\in R_\varepsilon^\xi(\Omega)} g_\varepsilon(\alpha, \alpha+\varepsilon\xi, u(\alpha+\varepsilon\xi) - u(\alpha)),$$

where $R_\varepsilon^\xi(\Omega) := \{\alpha \in \varepsilon\mathbb{Z}^N : \ [\alpha, \alpha+\varepsilon\xi] \subset \Omega\}$. Setting

$$f_\varepsilon^\xi(\alpha, \zeta) = \varepsilon^{-N} g_\varepsilon(\alpha, \alpha+\varepsilon\xi, \varepsilon|\xi|\zeta)$$

we can rewrite

$$F_\varepsilon(u) = \sum_{\xi\in\mathbb{Z}^N} \sum_{\alpha\in R_\varepsilon^\xi(\Omega)} \varepsilon^N f_\varepsilon^\xi\left(\alpha, \frac{u(\alpha+\varepsilon\xi) - u(\alpha)}{\varepsilon|\xi|}\right), \tag{1.1}$$

thus highlighting the dependence of the energy on discrete difference quotients in the direction ξ.

The aim of this paper is to provide a characterization of all the possible variational limits, as the mesh size ε tends to zero, of a very general class of energies of the form (1.1). Upon identifying u with a function constant on each cell of the lattice $\varepsilon\mathbb{Z}^N$, we can make the asymptotic analysis precise, thanks to the notions and the methods of De Giorgi's Γ-convergence (see [16], [4], [15]). On the functions $f_\varepsilon^\xi(\alpha, \cdot)$ we make assumptions of two types: a growth hypothesis of superlinear type on nearest

neighbors (see 3.2) that ensures that the limit is finite only on $W^{1,p}(\Omega;\mathbb{R}^d)$ and a decay assumption as $\xi \to +\infty$ (see (3.3), (H1), (H2)) that allows us to neglect very long-range interactions. Under these conditions, a compactness theorem holds asserting that, up to passing to a subsequence, the energies F_ε have a Γ-limit energy F defined on the Sobolev space $W^{1,p}(\Omega;\mathbb{R}^d)$ and taking the form

$$F(u) = \int_\Omega f(x, Du)\, dx$$

(see Theorem 3.1). A similar compactness result for quadratic interactions in planar networks has been observed by Vogelius [23] (see also Piatnitski and Remy [20]).

Note that the decay assumption on the density energies f_ε^ξ as $|\xi| \to +\infty$ guarantees that the nonlocality of our discrete functionals disappears in the limit. If this hypothesis is lifted, then we may have nonlocal Γ-limits (see [3]). On the other hand, if growth conditions are removed, the limit may be defined on sets of functions with bounded variation where a different analytical approach is needed (see [22], [5], [14], [1], [3], [8], [10]).

To perform our analysis, we develop the discrete analogue of a localization argument used, for example, in the context of homogenization theory for multiple integrals which allows us to regard our energies and their Γ-limits as functionals defined on pairs function-set and then to prove that all the hypotheses of an integral representation theorem are fulfilled. In order to treat minimum problems with boundary data, we also derive a compactness theorem in case that our functionals are subject to Dirichlet boundary conditions (see (3.30) and Theorem 3.10).

An interesting special case is when the arrangement of the "material points" presents a periodic feature; i.e., in terms of f_ε, we have

$$f_\varepsilon^\xi(\cdot, z) = f^\xi\left(\frac{\cdot}{\varepsilon}, z\right) \qquad f^\xi(\cdot, z), \quad Q_k\text{-periodic},$$

where $Q_k = (0,k)^N$. By adapting the integral homogenization arguments to our discrete setting, we prove that the whole family F_ε Γ-converges to a limit energy of the form

$$F(u) = \int_\Omega f_{hom}(Du)\, dx.$$

Note that in this setting we also include, when $k = 1$, the situation when $f^\xi(\alpha, z)$ is independent of α. If not only nearest neighbor interactions are present, the formula for f_{hom} highlights a multiple-scale effect also in this case (see [4]). An interesting example showing the effect of nonlinearities of "geometrical" origin is contained in a work by Friesecke and Theil [18], where an interpretation in terms of the Cauchy–Born rule is given.

Here f_{hom} is given by the following homogenization formula:

$$f_{hom}(M) = \lim_{h\to+\infty} \frac{1}{h^N} \min\left\{\mathcal{F}_h(u),\ u|_{\partial Q_h} = M\alpha\right\}, \tag{1.2}$$

where

$$\mathcal{F}_h(u) = \sum_{\xi\in\mathbb{Z}^N} \sum_{\alpha\in R^\xi(Q_h)} f^\xi\left(\alpha, \frac{u(\alpha+\xi) - u(\alpha)}{|\xi|}\right),$$

and $u|_{\partial Q_h} = M\alpha$ means that "near" the boundary of Q_h the function u is the discrete interpolation of the affine function Mx (for more precise definitions see (3.29) and Theorem 4.1). This formula generalizes that obtained in [11] in a one-dimensional scalar setting.

In general, (1.2) cannot be simplified to a cell problem formula and gives rise to a quasi-convex nonconvex function even for simple interactions. Indeed, in section 7 we provide an example of quasi-convex nonconvex f_{hom} drawing inspiration from Šverák's construction of a quasi-convex function which is not polyconvex (see [21]).

In sections 5 and 6 we study some important cases when the formula for f_{hom} can be simplified. For convex interactions a periodicity cell problem formula holds: if f^ξ is a convex function in the second variable for all $\xi \in \mathbb{Z}^N$, then (1.2) can be written as

$$f_{hom}(M) = \frac{1}{k^N} \min \left\{ \mathcal{F}(u),\ u\, Q_k\text{-periodic} \right\},$$

where

$$\mathcal{F}(u) = \sum_{\xi\in\mathbb{Z}^N} \sum_{\alpha\in\{0,1,\dots,k-1\}^N} f^\xi \left(\alpha, \frac{u(\alpha+\xi)-u(\alpha)}{|\xi|} + M\cdot\frac{\xi}{|\xi|} \right)$$

(see Theorem 5.1). An analogous result for discrete quadratic forms has been obtained by Piatnitski and Remy [20]. Our result has been used by Braides and Francfort [7] as a step for the derivation of optimal bounds for composite conducting networks in the particular case of quadratic interactions (see Remarks 3.2 and 5.2).

If we consider only interactions along independent directions a reduction to the one-dimensional case occurs: if $k = 1$, that is, f^ξ does not depend on α, and

$$f^\xi \equiv 0 \quad \forall \xi \in \mathbb{Z}^N :\ \xi \neq j e_i, \quad i \in \{1,2,\dots,N\}, \quad j \in \mathbb{N}, \tag{1.3}$$

where $\{e_1, e_2, \dots, e_N\}$ is the standard orthonormal base in $\mathbb{R}^N$, then

$$f_{hom}(M) = \sum_{i=1}^{N} (\tilde{f}_i)(M^i),$$

$(\tilde{f}_i)$ being convex functions defined by a one-dimensional homogenization formula and M^i the ith column of M (see Theorem 6.3). Note that here a superposition principle holds, in the sense that the limit energy is obtained by relaxing the energies due to the interactions in every coordinate direction independently and then summing over them.

From the results obtained in the one-dimensional setting in [11] (see Theorems 6.1 and 6.2), we deduce that the limit energy density f_{hom} can be rewritten by a nonasymptotic formula only if nearest and next-to-nearest neighbor interactions along the coordinate directions are considered (see Remark 6.5). In particular, in the case of only nearest neighbor interactions, the only effect of the passage from the discrete setting to the continuum is a separate convexification process in the coordinate directions.

2. Notation and preliminaries. We denote by $\{e_1, e_2, \dots, e_N\}$ the standard basis in $\mathbb{R}^N$, by $|\cdot|$ the usual euclidean norm, and by $\langle\cdot,\cdot\rangle$ the scalar product in $\mathbb{R}^N$. We denote by $\mathcal{M}^{d\times N}$ and $\mathcal{M}^{d\times d}_{sym}$ the space of $d\times N$ matrices and symmetric $d\times d$

matrices, respectively. For $P \in \mathcal{M}^{d\times N}$, $Q \in \mathcal{M}^{N\times l}$, $P \cdot Q$ denotes the standard row by column product. For $x,\ y \in \mathbb{R}^N$, $[x, y]$ denotes the segment between x and y. If Ω is a bounded open subset of $\mathbb{R}^N$, $\mathcal{A}(\Omega)$ is the family of all open subsets of Ω, while $\mathcal{A}_0(\Omega)$ denotes the family of all open subsets of Ω whose closure is a compact subset of Ω. If $B \subset \mathbb{R}^N$ is a Borel set, we will denote by $|B|$ its Lebesgue measure. We use standard notation for L^p and Sobolev spaces.

We also recall the standard notation for slicing arguments (see [4]). Let $\xi \in S^{N-1}$, and let $\Pi_\xi = \{y \in \mathbb{R}^N : \langle y, \xi\rangle = 0\}$ be the linear hyperplane orthogonal to ξ. If $y \in \Pi_\xi$ and $E \subset \mathbb{R}^N$ we define $E^\xi = \{y \text{ s.t. } \exists t \in \mathbb{R} : y + t\xi \in E\}$ and $E^\xi_y = \{t \in \mathbb{R} : y + t\xi \in E\}$. Moreover, if $u : E \to \mathbb{R}$ we set $u_{\xi,y} : E^\xi_y \to \mathbb{R}$ by $u_{\xi,y}(t) = u(y + t\xi)$.

We also introduce a useful notation for difference quotient along any direction. Fix $\xi \in \mathbb{R}^N$; for $\varepsilon > 0$ and for every $u : \mathbb{R}^N \to \mathbb{R}^d$ we define

$$D^\xi_\varepsilon u(x) := \frac{u(x + \varepsilon\xi) - u(x)}{\varepsilon|\xi|}.$$

2.1. Γ-convergence. We recall the notion of Γ-convergence in $L^p(\Omega;\mathbb{R}^d)$ (see [16], [15], [4]). A sequence of functionals $F_j : L^p(\Omega;\mathbb{R}^d) \to [0, +\infty]$ is said to Γ-converge to a functional $F : L^p(\Omega;\mathbb{R}^d) \to [0, +\infty]$ at $u \in L^p(\Omega;\mathbb{R}^d)$ as $j \to +\infty$, and we write $F(u) = \Gamma\text{-}\lim_j F_j(u)$ if the following two conditions hold:

(i) (lower semicontinuity inequality) for all sequences (u_j) converging to u in $L^p(\Omega;\mathbb{R}^d)$ we have that $F(u) \le \liminf_j F_j(u_j)$;

(ii) (existence of a recovery sequence) there exists a sequence (u_j) converging to u in $L^p(\Omega;\mathbb{R}^d)$ such that $F(u) = \lim_j F_j(u_j)$.

We say that F_j Γ-converges to F if $F(u) = \Gamma\text{-}\lim_j F_j(u)$ at all points $u \in L^p(\Omega;\mathbb{R}^d)$ and that F is the Γ-limit of F_j. The main reason for the introduction of this convergence is the following fundamental theorem.

THEOREM 2.1. *Let $F = \Gamma\text{-}\lim_j F_j$, and let a compact set $K \subset L^p(\Omega;\mathbb{R}^d)$ exist such that $\inf_{L^p(\Omega;\mathbb{R}^d)} F_j = \inf_K F_j$ for all j. Then*

$$\exists \min_{L^p(\Omega;\mathbb{R}^d)} F = \lim_j \inf_{L^p(\Omega;\mathbb{R}^d)} F_j.$$

Moreover, if (u_j) is a converging sequence such that $\lim_j F_j(u_j) = \lim_j \inf_{L^p(\Omega;\mathbb{R}^d)} F_j$, then its limit is a minimum point for F. If (F_ε) is a family of functionals indexed by $\varepsilon > 0$, then we say that F_ε Γ-converges to F as $\varepsilon \to 0^+$ if $F = \Gamma\text{-}\lim_j F_{\varepsilon_j}$ for all (ε_j) converging to 0. If we define the *lower and upper Γ-limits* by

$$F'(u) = \Gamma\text{-}\liminf_{\varepsilon\to0^+} F_\varepsilon(u) = \inf\left\{\liminf_{\varepsilon\to0^+} F_\varepsilon(u_\varepsilon) :\ u_\varepsilon \to u\right\},$$

$$F''(u) = \Gamma\text{-}\limsup_{\varepsilon\to0^+} F_\varepsilon(u) = \inf\left\{\limsup_{\varepsilon\to0^+} F_\varepsilon(u_\varepsilon) :\ u_\varepsilon \to u\right\},$$

respectively, then F_ε Γ-converges to F as $\varepsilon \to 0^+$ if and only if $F'(u) = F''(u) = F(u)$. Note that the functions F' and F'' are lower semicontinuous (see [15, Proposition 6.8]).

2.2. Integral representation on Sobolev spaces. In this section we recall an integral representation result on Sobolev spaces for functionals defined on pairs function-sets (see [13]).

THEOREM 2.2. *Let $1 \le p < \infty$, and let $F : W^{1,p}(\Omega;\mathbb{R}^d) \times \mathcal{A}(\Omega) \to [0, +\infty]$ be a functional satisfying the following conditions:*

(i) (locality) *F is local, i.e., $F(u,A)=F(v,A)$, if $u=v$ a.e. on $A\in\mathcal{A}(\Omega)$;*
(ii) (measure property) *for all $u\in W^{1,p}(\Omega;\mathbb{R}^d)$ the set function $F(u,\cdot)$ is the restriction of a Borel measure to $\mathcal{A}(\Omega)$;*
(iii) (growth condition) *there exists $c>0$ and $a\in L^1(\Omega)$ such that*

$$F(u,A)\le c\int_A (a(x)+|Du|^p)\,dx$$

for all $u\in W^{1,p}(\Omega;\mathbb{R}^d)$ and $A\in\mathcal{A}(\Omega)$;
(iv) (translation invariance in u) *$F(u+z,A)=F(u,A)$ for all $z\in\mathbb{R}^d$, $u\in W^{1,p}(\Omega;\mathbb{R}^d)$, and $A\in\mathcal{A}(\Omega)$;*
(v) (lower semicontinuity) *for all $A\in\mathcal{A}(\Omega)$, $F(\cdot,A)$ is sequentially lower semicontinuous with respect to the weak convergence in $W^{1,p}(\Omega;\mathbb{R}^d)$.*

Then there exists a Carathéodory function $f:\Omega\times\mathbb{M}^{d\times N}\to[0,+\infty)$ satisfying the growth condition

$$0\le f(x,M)\le c(a(x)+|M|^p)$$

for all $x\in\Omega$ and $M\in M^{d\times N}$ such that

$$F(u,A)=\int_A f(x,Du(x))\,dx$$

for all $u\in W^{1,p}(\Omega;\mathbb{R}^d)$ and $A\in\mathcal{A}(\Omega)$.
If, in addition, it holds that
(vi) (translation invariance in x)

$$F(Mx,B(y,\varrho))=F(Mx,B(z,\varrho))$$

for all $M\in M^{d\times N}$, $y,z\in\Omega$, and $\varrho>0$ such that $B(y,\varrho)\cup B(z,\varrho)\subset\Omega$, then f does not depend on x.

3. Compactness and integral representation. In this section we define the class of discrete energies we are going to consider in the rest of the paper, and we prove a general compactness theorem, asserting that any sequence of energies in this class has a subsequence whose Γ-limit F is an integral functional.

In what follows, Ω will denote a bounded open set of $\mathbb{R}^N$ with Lipschitz boundary. We consider the family of functionals $F_\varepsilon: L^p(\Omega;\mathbb{R}^d)\to[0,+\infty]$ defined as

$$F_\varepsilon(u)=\begin{cases}\displaystyle\sum_{\xi\in\mathbb{Z}^N}\sum_{\alpha\in R_\varepsilon^\xi(\Omega)}\varepsilon^N f_\varepsilon^\xi\left(\alpha,D_\varepsilon^\xi u(\alpha)\right) & \text{if } u\in\mathcal{A}_\varepsilon(\Omega),\\ +\infty & \text{otherwise,}\end{cases}\tag{3.1}$$

where for any $\xi\in\mathbb{Z}^N$ and $\varepsilon>0$

$$R_\varepsilon^\xi(\Omega):=\{\alpha\in\varepsilon\mathbb{Z}^N:\ [\alpha,\alpha+\varepsilon\xi]\subset\Omega\},$$

$$\mathcal{A}_\varepsilon(\Omega):=\{u:\mathbb{R}^N\to\mathbb{R}^d\ :\ u\text{ constant on }\alpha+[0,\varepsilon)^N\text{ for any }\alpha\in\varepsilon\mathbb{Z}^N\cap\Omega\},$$

and $f_\varepsilon^\xi:(\varepsilon\mathbb{Z}^N\cap\Omega)\times\mathbb{R}^d\to[0,+\infty)$ is a given function. On f_ε^ξ we make the following assumptions:

$$f_\varepsilon^{e_i}(\alpha,z)\ge c_1(|z|^p-1)\quad\forall(\alpha,z)\in(\varepsilon\mathbb{Z}^N\cap\Omega)\times\mathbb{R}^d,\quad i\in\{1,\dots,N\},\tag{3.2}$$

$$f_\varepsilon^\xi(\alpha,z)\le C_\varepsilon^\xi(|z|^p+1)\quad\forall(\alpha,z)\in(\varepsilon\mathbb{Z}^N\cap\Omega)\times\mathbb{R}^d,\quad\xi\in\mathbb{Z}^N, \tag{3.3}$$

where $c_1>0$, and $\{C_\varepsilon^\xi\}_{\varepsilon,\xi}$ satisfies

$$\limsup_{\varepsilon\to0^+}\sum_{\xi\in\mathbb{Z}^N}C_\varepsilon^\xi<+\infty; \tag{H1}$$

$$\forall\delta>0\quad\exists\,M_\delta>0:\quad\limsup_{\varepsilon\to0^+}\sum_{|\xi|>M_\delta}C_\varepsilon^\xi<\delta. \tag{H2}$$

The main result of this section is stated in the following theorem.

THEOREM 3.1 (compactness). *Let $\{f_\varepsilon^\xi\}_{\varepsilon,\xi}$ satisfy* (3.2), (3.3), *and let* (H1)–(H2) *hold. Then for every sequence (ε_j) of positive real numbers converging to* 0, *there exists a subsequence (ε_{j_k}) and a Carathéodory function quasi-convex in the second variable $f:\Omega\times\mathbb{R}^{d\times N}$ satisfying*

$$c(|M|^p-1)\le f(x,M)\le C(|M|^p+1),$$

with $0<c<C$, such that $(F_{\varepsilon_{j_k}}(\cdot))$ Γ-converges with respect to the $L^p(\Omega;\mathbb{R}^d)$-topology to the functional $F:L^p(\Omega;\mathbb{R}^d)\to[0,+\infty]$ defined as

$$F(u)=\begin{cases}\displaystyle\int_\Omega f(x,\nabla u)\,dx & \text{if } u\in W^{1,p}(\Omega;\mathbb{R}^d),\\ +\infty & \text{otherwise.}\end{cases} \tag{3.4}$$

Remark 3.2 (quadratic forms). Under the hypotheses of Theorem 3.1, if, in addition, for any $\xi\in\mathbb{Z}^N$ and $\varepsilon>0$ $f_\varepsilon^\xi(\alpha,\cdot)$ is a positive quadratic form on $\mathbb{R}^d$, that is,

$$f_\varepsilon^\xi(\alpha,z)=\ \langle A_\varepsilon^\xi(\alpha)z,z\rangle,\quad A_\varepsilon^\xi(\alpha)\in\mathcal{M}^{d\times d}_{sym},$$

then, by the properties of Γ-convergence (see [15]), the limit energy density $f(x,\cdot)$ is a quadratic form on $\mathcal{M}^{d\times N}$, that is,

$$f(x,M)=A(x)\,(M,M)\,,\quad A(x)\in T_2\mathcal{M}^{d\times N}, \tag{3.5}$$

where $T_2\mathcal{M}^{d\times N}$ is the vectorial space of all two times covariant tensors on $\mathcal{M}^{d\times N}$.

To prove Theorem 3.1 we use a localization technique, which is a standard argument dealing with limits of integral functionals (see, for example, [6] in the context of homogenization theory). We stress the fact that here this analysis becomes more difficult to perform because of the nonlocality of our discrete energies.

The first step is to define a "localized" version of our energies: given an open set A we isolate the contributions due to interactions within A as follows. For $u\in\mathcal{A}_\varepsilon(\Omega)$, $A\in\mathcal{A}(\Omega)$, and $\xi\in\mathbb{Z}^N$, set

$$\mathcal{F}_\varepsilon^\xi(u,A):=\sum_{\alpha\in R_\varepsilon^\xi(A)}\varepsilon^N f_\varepsilon^\xi\left(\alpha,D_\varepsilon^\xi u(\alpha)\right), \tag{3.6}$$

where

$$R_\varepsilon^\xi(A):=\{\alpha\in\varepsilon\mathbb{Z}^N:\ [\alpha,\alpha+\varepsilon\xi]\subset A\}.$$

The function $\mathcal{F}_\varepsilon^\xi$ represents the energy due to the interactions within A along the direction ξ. Then the local version of the functional in (3.1) is given by

$$F_\varepsilon(u,A)=\begin{cases}\displaystyle\sum_{\xi\in\mathbb{Z}^N}\mathcal{F}_\varepsilon^\xi(u,A) & \text{if } u\in\mathcal{A}_\varepsilon(\Omega),\\ +\infty & \text{otherwise.}\end{cases} \tag{3.7}$$

We will prove also the following result.

THEOREM 3.3 (local compactness). *Let $\{f_\varepsilon^\xi\}_{\varepsilon,\xi}$ satisfy* (3.2), (3.3), *and let* (H1)–(H2) *hold. Given (ε_j), a sequence of positive real numbers converging to* 0, *let (ε_{j_k}) and f be as in Theorem* 3.1. *Then for any $u\in W^{1,p}(\Omega;\mathbb{R}^d)$ and $A\in\mathcal{A}(\Omega)$ there holds*

$$\Gamma\text{-}\lim_k F_{\varepsilon_{j_k}}(u,A)=\int_A f(x,\nabla u)\,dx.$$

We will derive the proof of Theorems 3.1 and 3.3 as a direct consequence of some propositions and lemmas which are fundamental steps to show that our limit functionals satisfy all the hypotheses of the representation theorem, Theorem 2.2.

In the next two propositions we show that, thanks to hypotheses (3.2) and (3.3), the Γ-lim inf and the Γ-lim sup of F_ε are finite only on $W^{1,p}(\Omega;\mathbb{R}^d)$ and satisfy standard p-growth conditions.

PROPOSITION 3.4. *Let $\{f_\varepsilon^{e_i}\}_{\varepsilon,i}$ satisfy* (3.2). *If $u\in L^p(\Omega;\mathbb{R}^d)$ is such that $F'(u,A)<+\infty$, then $u\in W^{1,p}(A;\mathbb{R}^d)$, and*

$$F'(u,A)\geq c\ \left(\|\nabla u\|^p_{L^p(A;\mathbb{R}^{d\times N})}-|A|\right) \tag{3.8}$$

for some positive constant c independent of u and A.

Proof. Let $\varepsilon_n\to 0^+$, and let u_n converge to u in $L^p(\Omega;\mathbb{R}^d)$ and be such that $\liminf_n F_{\varepsilon_n}(u_n,A)<+\infty$. By the growth condition (3.2) we get

$$F_{\varepsilon_n}(u_n,A)\geq c_1\sum_{i=1}^N\sum_{\alpha\in R^{e_i}_{\varepsilon_n}(A)}\varepsilon_n^N|D^{e_i}_{\varepsilon_n}u_n(\alpha)|^p-c_1N|A|.$$

For any $i\in\{1,\dots,N\}$, consider the sequence of piecewise-affine functions (v_n^i) defined as follows:

$$v_n^i(x):=u_n(\alpha)+D^{e_i}_{\varepsilon_n}u_n(\alpha)(x_i-\alpha_i),\quad x\in\left(\alpha+[0,\varepsilon_n)^N\right)\cap\Omega,\quad \alpha\in R^{e_i}_{\varepsilon_n}(A).$$

Note that v_n^i is a function of bounded variation, and we will denote by $\frac{\partial v_n^i}{\partial x_i}$ the density of the absolutely continuous part of $D_{x_i}v_n^i$ with respect to the Lebesgue measure. Moreover, for $\mathcal{H}^{N-1}$-a.e. $y\in(A)^{e_i}$ the slices $(v_n^i)_{e_i,y}\in W^{1,p}\left((A)^{e_i}_y;\mathbb{R}^d\right)$. For any $\eta>0$, set

$$A_\eta:=\{x\in A:\ \text{dist }(x,A^c)>\eta\}.$$

Then, with fixed $\eta>0$, it is easy to check that $v_n^i\to u$ in $L^p(A_\eta;\mathbb{R}^d)$ for every $i\in\{1,\dots,N\}$; moreover, since $\frac{\partial v_n^i}{\partial x_i}(x)=D^{e_i}_{\varepsilon_n}u_n(\alpha)$ for $x\in\alpha+[0,\varepsilon_n)^N$, we get

$$F_{\varepsilon_n}(u_n,A)\geq c_1\sum_{i=1}^N\int_{A_\eta}\left|\frac{\partial v_n^i}{\partial x_i}(x)\right|^p\,dx-c_1N|A|. \tag{3.9}$$

We apply now a standard slicing argument. By Fubini's theorem and Fatou's lemma for any i we get

$$\liminf_n \int_{(A_\eta)} \left|\frac{\partial v_n^i}{\partial x_i}(x)\right|^p \geq \int_{(A_\eta)^{e_i}} \liminf_n \int_{(A_\eta)_y^{e_i}} \left|(v_n^i)'_{e_i,y}(t)\right|^p \, dt \, d\mathcal{H}^{N-1}(y).$$

Since, up to passing to a subsequence, we may assume that, for $\mathcal{H}^{N-1}$-a.e. $y \in (A_\eta)^{e_i}$ $(v_n^i)_{e_i,y} \to u_{e_i,y}$ in $L^p\left((A_\eta)_y^{e_i};\mathbb{R}^d\right)$, we deduce that $u_{e_i,y} \in W^{1,p}\left((A_\eta)_y^{e_i};\mathbb{R}^d\right)$ for $\mathcal{H}^{N-1}$-a.e. $y \in (A_\eta)^{e_i}$, and

$$\liminf_n \int_{(A_\eta)} \left|\frac{\partial v_n^i}{\partial x_i}(x)\right|^p \geq \int_{(A_\eta)^{e_i}} \int_{(A_\eta)_y^{e_i}} \left|u'_{e_i,y}(t)\right|^p \, dt \, d\mathcal{H}^{N-1}(y).$$

Then, by (3.9), we have

$$\liminf_n F_{\varepsilon_n}(u_n, A) \geq c_1 \sum_{i=1}^N \int_{(A_\eta)^{e_i}} \int_{(A_\eta)_y^{e_i}} \left|u'_{e_i,y}(t)\right|^p \, dt \, d\mathcal{H}^{N-1}(y) - c_1 N|A|.$$

Since, in particular, the previous inequality implies that

$$\sum_{i=1}^N \int_{(A_\eta)^{e_i}} \int_{(A_\eta)_y^{e_i}} \left|u'_{e_i,y}(t)\right|^p \, dt \, d\mathcal{H}^{N-1}(y) < +\infty,$$

thanks to the characterization of $W^{1,p}$ by slicing, we obtain that $u \in W^{1,p}\left(A_\eta;\mathbb{R}^d\right)$, and

$$\begin{aligned}\liminf_n F_{\varepsilon_n}(u_n, A) &\geq c_1 \sum_{i=1}^N \int_{A_\eta} \left|\frac{\partial u}{\partial x_i}(x)\right|^p \, dx - c_1 N|A| \\ &\geq c\left(\int_{A_\eta} \|\nabla u(x)\|^p \, dx - |A|\right).\end{aligned}$$

Letting $\eta \to 0^+$, we get the conclusion. □

PROPOSITION 3.5. *Let $\{f_\varepsilon^\xi\}_{\varepsilon,\xi}$ satisfy* (3.3)*, and let* (H1) *hold. Then for every $u \in W^{1,p}(\Omega;\mathbb{R}^d)$ there holds*

$$F''(u, A) \leq C \, \left(\|\nabla u\|^p_{L^p(A;\mathbb{R}^{d\times N})} + |A|\right) \tag{3.10}$$

for some positive constant C independent of u and A.

Proof. We first show that inequality (3.5) holds for u smooth and then recover the proof for any $u \in W^{1,p}(\Omega;\mathbb{R}^d)$ by using a density argument.

Let $u \in C_c^\infty(\mathbb{R}^N;\mathbb{R}^d)$, and consider the family $(u_\varepsilon) \subset \mathcal{A}_\varepsilon(\Omega)$ defined as

$$u_\varepsilon(\alpha) := u(\alpha), \quad \alpha \in \varepsilon\mathbb{Z}^N.$$

Then $u_\varepsilon \to u$ in $L^p(\Omega;\mathbb{R}^d)$ as $\varepsilon \to 0^+$. Moreover, for any $\alpha \in \varepsilon\mathbb{Z}^N$, we have

$$D_\varepsilon^\xi u_\varepsilon(\alpha) = \frac{1}{|\xi|}\int_0^1 \nabla u(\alpha + \varepsilon\xi s)\xi \, ds$$

so that, by Jensen's inequality, we get

$$|D_\varepsilon^\xi u_\varepsilon(\alpha)|^p = \frac{1}{|\xi|^p}\left|\int_0^1 \nabla u(\alpha+\varepsilon\xi s)\xi\, ds\right|^p \\ \leq \frac{1}{|\xi|^p}\int_0^1 |\nabla u(\alpha+\varepsilon\xi s)\xi|^p\, ds \leq \int_0^1 |\nabla u(\alpha+\varepsilon\xi s)|^p\, ds.$$

By the regularity hypothesis on u and by Fubini's theorem, we easily obtain the following inequalities:

$$\varepsilon^N\int_0^1 |\nabla u(\alpha+\varepsilon\xi s)|^p\, ds = \int_{\alpha+[0,\varepsilon)^N}\int_0^1 |\nabla u(\alpha+\varepsilon\xi s)|^p\, ds\, dx \\ \leq \int_{\alpha+[0,\varepsilon)^N}\int_0^1 |\nabla u(x+\varepsilon\xi s)|^p\, ds\, dx + c(u)\int_{\alpha+[0,\varepsilon)^N}\int_0^1 |x-\alpha|^p\, ds\, dx \\ \leq \int_0^1\int_{\alpha+s\varepsilon\xi+[0,\varepsilon)^N} |\nabla u(x)|^p\, dx\, ds + c(u)\varepsilon^p\varepsilon^N,$$

where by $c(u)$ we denote a constant depending only on u. By (3.3) and the last inequality, we then have

$$F_\varepsilon(u_\varepsilon, A) \leq \sum_{\xi\in\mathbb{Z}^N} C_\varepsilon^\xi \sum_{\alpha\in R_\varepsilon^\xi(A)}\int_0^1\int_{\alpha+s\varepsilon\xi+[0,\varepsilon)^N} |\nabla u(x)|^p\, dx\, ds \\ + (1+c(u)\varepsilon^p)\sum_{\xi\in\mathbb{Z}^N} C_\varepsilon^\xi \sum_{\alpha\in R_\varepsilon^\xi(A)} \varepsilon^N \\ \leq \sum_{\xi\in\mathbb{Z}^N} C_\varepsilon^\xi\left(\int_{A^\varepsilon} |\nabla u(x)|^p\, dx + (1+c(u)\varepsilon^p)\,|A^\varepsilon|\right),$$

where

$$A^\varepsilon := A + [0,\varepsilon)^N.$$

Eventually, letting $\varepsilon\to 0^+$, by (H1) we get

$$\limsup_{\varepsilon\to 0^+} F_\varepsilon(u_\varepsilon, A) \leq C\left(\int_A |\nabla u(x)|^p\, dx + |A|\right),$$

and the conclusion follows by the definition of F''. Now let $u\in W^{1,p}(\Omega;\mathbb{R}^d)$, and let $(u_n)\subset C_c^\infty(\mathbb{R}^N;\mathbb{R}^d)$ converge to u in the $W^{1,p}(\Omega;\mathbb{R}^d)$-topology. Then, by the lower semicontinuity of F'', we obtain

$$F''(u,A) \leq \liminf_n F''(u_n,A) \leq \lim_n C\left(\|\nabla u_n\|^p_{L^p(A;\mathbb{R}^{d\times N})} + |A|\right) \\ = C\left(\|\nabla u\|^p_{L^p(A;\mathbb{R}^{d\times N})} + |A|\right). \qquad \square$$

The next technical lemma asserts that finite difference quotients along any direction can be controlled by finite difference quotients along the coordinate directions.

LEMMA 3.6. *Let $A\in\mathcal{A}(\Omega)$, and set $A_\varepsilon := \{x\in A : \operatorname{dist}(x,\partial A) > 2\sqrt{N}\varepsilon\}$. Then for any $\xi\in\mathbb{Z}^N$ and $u\in\mathcal{A}_\varepsilon(\Omega)$ there holds*

$$\sum_{\alpha\in R_\varepsilon^\xi(A_\varepsilon)} |D_\varepsilon^\xi u(\alpha)|^p \leq C\sum_{i=1}^N \sum_{\alpha\in R_\varepsilon^{e_i}(A)} |D_\varepsilon^{e_i} u(\alpha)|^p. \tag{3.11}$$

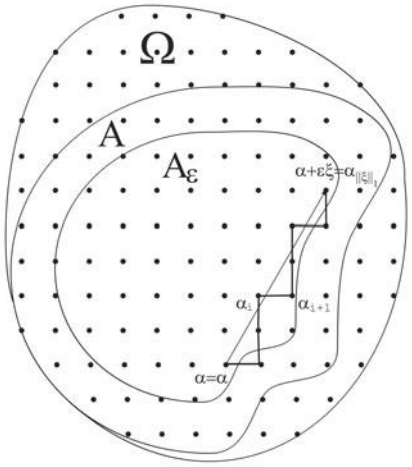

FIG. 2.

Proof. Let us fix some notations: for $\xi \in \mathbb{Z}^N$ and $\alpha \in \varepsilon\mathbb{Z}^N$, set

$$I_\varepsilon^\xi(\alpha) := \left\{\beta \in \varepsilon\mathbb{Z}^N : \ \left(\beta + [-\varepsilon,\varepsilon]^N\right) \cap [\alpha, \alpha + \varepsilon\xi] \neq \emptyset\right\};$$

moreover, we will denote by $\|\cdot\|_1$ the norm on $\mathbb{R}^N$ defined as

$$\|\xi\|_1 := \sum_{i=1}^N |\xi_i|, \quad \xi = (\xi_1, \dots, \xi_N) \in \mathbb{R}^N.$$

Let $\alpha \in R_\varepsilon^\xi(A_\varepsilon)$, and consider $\{\alpha_h\}_{h=1}^{\|\xi\|_1} \subset I_\varepsilon^\xi(\alpha)$ such that

$$\alpha_{\|\xi\|_1} = \alpha + \varepsilon\xi, \quad \alpha_1 = \alpha, \quad \alpha_h = \alpha_{h-1} + \varepsilon e_{i(h)}$$

for some $i(h) \in \{1, \dots, N\}$ (see Figure 2). Then, since

$$D_\varepsilon^\xi u(\alpha) = \frac{1}{|\xi|} \sum_{h=1}^{\|\xi\|_1} D_\varepsilon^{e_{i(h)}} u(\alpha_h)$$

by Jensen's inequality, we get

$$\begin{aligned} |D_\varepsilon^\xi u(\alpha)|^p &= \left(\frac{\|\xi\|_1}{|\xi|}\right)^p \left| \frac{1}{\|\xi\|_1} \sum_{h=1}^{\|\xi\|_1} D_\varepsilon^{e_{i(h)}} u(\alpha_h) \right|^p \\ &\le \left(\frac{\|\xi\|_1}{|\xi|}\right)^p \frac{1}{\|\xi\|_1} \sum_{h=1}^{\|\xi\|_1} |D_\varepsilon^{e_{i(h)}} u(\alpha_h)|^p. \end{aligned}$$

Since for any $h = 1, \dots, N$, $\alpha_h \in R_\varepsilon^{e_{i(h)}}(A)$ and all the norms are equivalent in a finite-dimensional space, we infer that

$$\sum_{\alpha \in R_\varepsilon^\xi(A_\varepsilon)} |D_\varepsilon^\xi u(\alpha)|^p \le C \sum_{i=1}^N \sum_{\beta \in R_\varepsilon^{e_i}(A)} \frac{\gamma_\varepsilon^\xi(\beta)}{\|\xi\|_1} |D_\varepsilon^{e_i} u(\beta)|^p,$$

where

$$\gamma_\varepsilon^\xi(\beta) := \#\{\alpha \in R_\varepsilon^\xi(A_\varepsilon) :\ \beta \in I_\varepsilon^\xi(\alpha)\}.$$

Hence, the proof is complete if we show that $\gamma_\varepsilon^\xi(\beta) \le C|\xi|$. To this aim, notice that

$$\{\alpha \in R_\varepsilon^\xi(A_\varepsilon) :\ \beta \in I_\varepsilon^\xi(\alpha)\} \subseteq \varepsilon\mathbb{Z}^N \cap Q_\varepsilon^\xi(\beta),$$

where

$$Q_\varepsilon^\xi(\beta) := \{x \in \mathbb{R}^N : x = y + t\xi,\ y \in \beta + [-\varepsilon,\varepsilon]^n,\ t \in [-\varepsilon,\varepsilon]^N\}.$$

Thus, we infer that

$$\gamma_\varepsilon^\xi(\beta) \le C\frac{\left|Q_\varepsilon^\xi(\beta)\right|}{\varepsilon^N}.$$

Now we use a slicing argument to provide an estimate of $\left|Q_\varepsilon^\xi(\beta)\right|$. By Fubini's theorem, we get

$$\begin{aligned}|Q_\varepsilon^\xi(\beta)| &= \int_{\left(Q_\varepsilon^\xi(\beta)\right)^\xi} \mathcal{H}^1\left(Q_\varepsilon^\xi(\beta)\right)_y^\xi \, d\mathcal{H}^{N-1}(y)\\ &\le \mathcal{H}^{N-1}\left((Q_\varepsilon^\xi(\beta))^\xi\right) 2(\sqrt{N}+|\xi|)\varepsilon \le c(N)|\xi|\varepsilon^N,\end{aligned}$$

where the last inequality holds, since for any $\xi \in \mathbb{Z}^N$

$$\mathcal{H}^{N-1}\left((Q_\varepsilon^\xi(\beta))^\xi\right) \le c(N)\varepsilon^{N-1}. \qquad \square$$

In the next two propositions we establish the subadditivity and the inner regularity of the set function $F''(u,\cdot)$. To this end we use a careful modification of De Giorgi's cut-off functions argument, which appears frequently in the proof of the integral representation of Γ-limits of integral functionals (see [6], [15]). We underline that the nonlocality of our energies requires a deeper analysis in which a key role is played by hypothesis (H2), which allows us to show that very long-range interactions do not lead to nonlocal terms in the limit.

PROPOSITION 3.7. *Let $\{f_\varepsilon^\xi\}_{\varepsilon,\xi}$ satisfy* (3.2),(3.3)*, and let* (H1)–(H2) *hold. Let $A, B \in \mathcal{A}(\Omega)$, and let $A', B' \in \mathcal{A}(\Omega)$ be such that $A' \subset\subset A$ and $B' \subset\subset B$. Then, for any $u \in W^{1,p}(\Omega;\mathbb{R}^d)$,*

$$F''(u, A' \cup B') \le F''(u, A) + F''(u, B).$$

Proof. Without loss of generality, we may suppose that $F''(u,A)$ and $F''(u,B)$ are finite. Let $u_\varepsilon, v_\varepsilon \in \mathcal{A}_\varepsilon(\Omega)$ both converge to u in $L^p(\Omega;\mathbb{R}^d)$ and be such that

$$\limsup_{\varepsilon\to 0^+} F_\varepsilon(u_\varepsilon, A) = F''(u,A), \quad \limsup_{\varepsilon\to 0^+} F_\varepsilon(v_\varepsilon, B) = F''(u,B).$$

By (3.2) and Lemma 3.6, we infer that

$$\sup_{\xi\in\mathbb{Z}^N}\sup_{\varepsilon>0} \sum_{\alpha\in R_\varepsilon^\xi(A_\varepsilon)} \varepsilon^N |D_\varepsilon^\xi u_\varepsilon(\alpha)|^p < +\infty, \tag{3.12}$$

$$\sup_{\xi\in\mathbb{Z}^N}\sup_{\varepsilon>0} \sum_{\alpha\in R_\varepsilon^\xi(B_\varepsilon)} \varepsilon^N |D_\varepsilon^\xi v_\varepsilon(\alpha)|^p < +\infty, \tag{3.13}$$

where A_ε and B_ε are defined as in Lemma 3.6. Moreover, since (u_ε) and (v_ε) converge to u in the $L^p(\Omega;\mathbb{R}^d)$-topology, we have

$$\sum_{\alpha\in\varepsilon\mathbb{Z}^N\cap\Omega'} \varepsilon^N \left(|u_\varepsilon(\alpha)|^p + |v_\varepsilon(\alpha)|^p\right) \le \|u_\varepsilon\|^p_{L^p(\Omega;\mathbb{R}^d)} + \|v_\varepsilon\|^p_{L^p(\Omega;\mathbb{R}^d)} \le C < +\infty, \tag{3.14}$$

$$\sum_{\alpha\in\varepsilon\mathbb{Z}^N\cap\Omega'} \varepsilon^N \left(|u_\varepsilon(\alpha) - v_\varepsilon(\alpha)|^p\right) \le \|u_\varepsilon - v_\varepsilon\|^p_{L^p(\Omega;\mathbb{R}^d)} \to 0^+ \tag{3.15}$$

for any $\Omega' \subset\subset \Omega$. Set

$$d := \operatorname{dist}(A', A^c),$$

and for any $i \in \{1,\dots,N\}$ define

$$A_i := \left\{ x \in A :\ \operatorname{dist}(x, A') < i\frac{d}{N} \right\}.$$

Let φ_i be a cut-off function between A_i and A_{i+1}, with $\|\nabla\varphi_i\|_\infty \le 2\frac{N}{d}$. Then for any $i \in \{1,\dots,N\}$ consider the family of functions $w^i_\varepsilon \in \mathcal{A}_\varepsilon(\Omega)$ still converging to u in $L^p(\Omega;\mathbb{R}^d)$ defined as

$$w^i_\varepsilon(\alpha) := \varphi_i(\alpha)u_\varepsilon(\alpha) + (1-\varphi_i(\alpha))\, v_\varepsilon(\alpha).$$

Note that, for any $\xi \in \mathbb{Z}^N$, we have

$$\begin{aligned} D^\xi_\varepsilon w^i_\varepsilon(\alpha) = \varphi_i(\alpha+\varepsilon\xi) D^\xi_\varepsilon u_\varepsilon(\alpha) + (1-\varphi_i(\alpha+\varepsilon\xi))\, D^\xi_\varepsilon v_\varepsilon(\alpha) \\ + (u_\varepsilon(\alpha) - v_\varepsilon(\alpha))\, D^\xi_\varepsilon\varphi(\alpha). \end{aligned} \tag{3.16}$$

Fix $i \in \{1,2,\dots,N-3\}$. Given $\xi \in \mathbb{Z}^N$ and $\alpha \in R^\xi_\varepsilon(A'\cup B')$, then either $\alpha \in R^\xi_\varepsilon(A_i)$ or $\alpha \in R^\xi_\varepsilon(\overline{A}^c_{i+1} \cap B')$, or

$$[\alpha, \alpha+\varepsilon\xi] \cap \left(\overline{A}_{i+1} \setminus A_i\right) \cap B' \neq \emptyset.$$

Then, if we set

$$\left(\overline{A}_{i+1} \setminus A_i\right)^{\varepsilon,\xi} := \{x = y + t\xi,\ |t| \le \varepsilon,\ y \in \overline{A}_{i+1} \setminus A_i\},$$

$$S^{\varepsilon,\xi}_i := \left(\overline{A}_{i+1} \setminus A_i\right)^{\varepsilon,\xi} \cap (A'\cup B'),$$

we get

$$R^\xi_\varepsilon(A'\cup B') \subseteq R^\xi_\varepsilon(A_i) \cup R^\xi_\varepsilon(B' \setminus \overline{A}_{i+1}) \cup R^\xi_\varepsilon\left(S^{\varepsilon,\xi}_i\right)$$

(see Figure 3). Thus, since $D^\xi_\varepsilon w^i_\varepsilon(\alpha) = D^\xi_\varepsilon u_\varepsilon(\alpha)$ if $\alpha \in R^\xi_\varepsilon(A_i)$ and $D^\xi_\varepsilon w^i_\varepsilon(\alpha) = D^\xi_\varepsilon v_\varepsilon(\alpha)$ if $\alpha \in R^\xi_\varepsilon(\overline{A}^c_{i+1} \cap B')$, we get by (3.3) and (3.16)

$$\begin{aligned} \mathcal{F}^\xi_\varepsilon\ (w^i_\varepsilon, A'\cup B') &\le \mathcal{F}^\xi_\varepsilon(u_\varepsilon, A) + \mathcal{F}^\xi_\varepsilon(v_\varepsilon, B) \\ &+ C\ C^\xi_\varepsilon \sum_{\alpha\in R^\xi_\varepsilon(S^{\varepsilon,\xi}_i)} \varepsilon^N \left(|D^\xi_\varepsilon u_\varepsilon(\alpha)|^p + |D^\xi_\varepsilon v_\varepsilon(\alpha)|^p + N^p\,|u_\varepsilon(\alpha) - v_\varepsilon(\alpha)|^p + 1\right). \end{aligned} \tag{3.17}$$

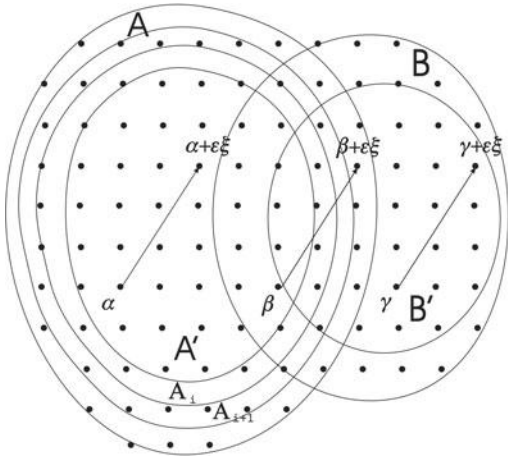

FIG. 3. $\alpha \in R^\xi_\varepsilon(A_i), \beta \in R^\xi_\varepsilon(S_i^{\varepsilon,\xi}), \gamma \in R^\xi_\varepsilon(B' \setminus A^c_{i+1})$.

If $\varepsilon|\xi| \le \frac{d}{2N}$, then

$$S_i^{\varepsilon,\xi} \subseteq \left(A_{N-1} \setminus \overline{A'}\right) \cap B' =: S_N \subset\subset A \cap B. \tag{3.18}$$

If $\varepsilon|\xi| \ge \frac{d}{2N}$, then

$$\frac{1}{\varepsilon^p|\xi|^p} \le \frac{2^p N^p}{d^p},$$

and so

$$|D^\xi_\varepsilon u_\varepsilon(\alpha)|^p \le CN^p \left(|u_\varepsilon(\alpha)|^p + |u_\varepsilon(\alpha+\varepsilon\xi)|^p\right),$$

and the same inequality holds for v_ε. Thus, in this case we get by (3.17)

$$\begin{aligned}\mathcal{F}^\xi_\varepsilon(w^i_\varepsilon, A' \cup B') &\le \mathcal{F}^\xi_\varepsilon(u_\varepsilon, A) + \mathcal{F}^\xi_\varepsilon(v_\varepsilon, B)\\ &+ CN^p\, C^\xi_\varepsilon \sum_{\alpha \in R^\xi_\varepsilon(A' \cup B')} \varepsilon^N \left(|u_\varepsilon(\alpha)|^p + |u_\varepsilon(\alpha+\varepsilon\xi)|^p + |v_\varepsilon(\alpha)|^p + |v_\varepsilon(\alpha+\varepsilon\xi)|^p + 1\right).\end{aligned} \tag{3.19}$$

Let $M_\delta > 0$ be such that $\limsup_{\varepsilon\to 0^+} \sum_{|\xi|>M_\delta} C^\xi_\varepsilon < \delta$. Then, by (3.17), (3.18), and (3.19), summing over $\xi \in \mathbb{Z}^N$, for ε small enough we get

$$\begin{aligned}F_\varepsilon(w^i_\varepsilon, A' \cup B') &\le F_\varepsilon(u_\varepsilon, A) + F_\varepsilon(v_\varepsilon, B)\\
&+ C \sum_{|\xi| \le M_\delta} C^\xi_\varepsilon \sum_{\alpha \in R^\xi_\varepsilon(S_i^{\varepsilon,\xi})} \varepsilon^N \left(|D^\xi_\varepsilon u_\varepsilon(\alpha)|^p + |D^\xi_\varepsilon v_\varepsilon(\alpha)|^p + N^p\,|u_\varepsilon(\alpha) - v_\varepsilon(\alpha)|^p + 1\right)\\
&+ C \sum_{M_\delta < |\xi| \le \frac{d}{2N\varepsilon}} C^\xi_\varepsilon \sum_{\alpha \in R^\xi_\varepsilon(S_N)} \varepsilon^N \left(|D^\xi_\varepsilon u_\varepsilon(\alpha)|^p + |D^\xi_\varepsilon v_\varepsilon(\alpha)|^p + N^p\,|u_\varepsilon(\alpha) - v_\varepsilon(\alpha)|^p + 1\right)\\
&+ CN^p \sum_{|\xi| > \frac{d}{2N\varepsilon}} C^\xi_\varepsilon \sum_{\alpha \in \varepsilon\mathbb{Z}^N \cap A' \cup B'} \varepsilon^N \left(|u_\varepsilon(\alpha)|^p + |v_\varepsilon(\alpha)|^p + 1\right).\end{aligned}$$

Note that, for ε small enough and $|\xi| \le M_\delta$, we have that $R^\xi_\varepsilon(S_i^{\varepsilon,\xi}) \cap R^\xi_\varepsilon(S_j^{\varepsilon,\xi}) \ne \emptyset$ if and only if $|i-j| = 1$, and $\bigcup_{i=1}^{N-3} R^\xi_\varepsilon(S_i^{\varepsilon,\xi}) \subseteq R^\xi_\varepsilon(A_\varepsilon \cap B_\varepsilon)$. Thus, summing over

$i \in \{1, 2, \dots, N-3\}$, averaging, and taking into account (3.12), (3.13), (3.14), and (3.15), we get

$$\begin{aligned}(3.20)\quad \frac{1}{N-3}\sum_{i=1}^{N-3} F_\varepsilon(w_\varepsilon^i, A' \cup B') \le F_\varepsilon(u_\varepsilon, A) + F_\varepsilon(v_\varepsilon, B) \\ + \frac{C}{N-3}(1 + N^p O(\varepsilon)) + C(\delta + O(\varepsilon))(1 + N^p O(\varepsilon)) \\ + C\,(\delta + O(\varepsilon))(N^p)\,.\end{aligned}$$

For any $\varepsilon > 0$ there exists $i(\varepsilon) \in \{1, \dots, N-3\}$ such that

$$(3.21)\qquad F_\varepsilon(w_\varepsilon^{i(\varepsilon)}, A' \cup B') \le \frac{1}{N-3}\sum_{i=1}^{N-3} F_\varepsilon(w_\varepsilon^i, A' \cup B').$$

Then, since $w_\varepsilon^{i(\varepsilon)}$ still converges to u in $L^p(\Omega; \mathbb{R}^d)$, by (3.20) and (3.21), letting $\varepsilon \to 0^+$, we get

$$F''(u, A' \cup B') \le F''(u, A) + F''(u, B) + \frac{C}{N-3} + C\delta\,(1 + N^p)\,.$$

Eventually, letting first $\delta \to 0^+$ and then $N \to +\infty$, we obtain the thesis. □

PROPOSITION 3.8. *Let $\{f_\varepsilon^\xi\}_{\varepsilon,\xi}$ satisfy* (3.2), (3.3)*, and let* (H1)–(H2) *hold. Then, for any $u \in W^{1,p}(\Omega; \mathbb{R}^d)$ and for any $A \in \mathcal{A}(\Omega)$, there holds*

$$\sup_{A' \subset\subset A} F''(u, A') = F''(u, A).$$

Proof. Since $F''(u, \cdot)$ is an increasing set function, it suffices to prove that

$$\sup_{A' \subset\subset A} F''(u, A') \ge F''(u, A).$$

To do this, we apply the same argument of the proof of Proposition 3.7. Given $\delta > 0$, there exists $A'' \subset\subset A$ such that

$$|A \setminus \overline{A''}| + \|\nabla u\|^p_{L^p(A \setminus \overline{A''})} \le \delta.$$

Let $\tilde\Omega \supset\supset \Omega$, and let $\tilde u \in W^{1,p}(\tilde\Omega; \mathbb{R}^d)$ be an extension of u. By reasoning as in the proof of Proposition 3.5, we may find $v_\varepsilon \in \mathcal{A}_\varepsilon(\tilde\Omega)$ such that v_ε converges to $\tilde u$ in $L^p(\tilde\Omega; \mathbb{R}^d)$ and

$$(3.22)\qquad \limsup_{\varepsilon \to 0^+} F_\varepsilon(v_\varepsilon, A \setminus \overline{A''}) \le C\left(|A \setminus \overline{A''}| + \|\nabla u\|^p_{L^p(A \setminus \overline{A''})}\right) \le C\delta.$$

We remark that this extension on $\tilde\Omega$ is just a technical tool to exploit an analogue of inequality (3.14) and obtain a control of the interactions near the boundary of Ω. Let $A' \in \mathcal{A}(\Omega)$ be such that $A'' \subset\subset A' \subset\subset A$, and let $u_\varepsilon \in \mathcal{A}_\varepsilon(\Omega)$ converge to u in $L^p(\Omega; \mathbb{R}^d)$, with

$$\limsup_{\varepsilon \to 0^+} F_\varepsilon(u_\varepsilon, A') = F''(u, A').$$

Set

$$d := \operatorname{dist}(A'', A'^c),$$

and for any $i \in \{1,\dots,N\}$ define

$$A_i := \left\{ x \in A : \ \operatorname{dist}(x, A') < i\frac{d}{N} \right\}.$$

Let φ_i be a cut-off function between A_i and A_{i+1}, with $\|\nabla\varphi_i\|_\infty \le 2\frac{N}{d}$. Then for any $i \in \{1,\dots,N\}$ consider the family of functions $w_\varepsilon^i \in \mathcal{A}_\varepsilon(\Omega)$ still converging to u in $L^p(\Omega;\mathbb{R}^d)$ defined as

$$w_\varepsilon^i(\alpha) := \varphi_i(\alpha)u_\varepsilon(\alpha) + (1-\varphi_i(\alpha))\, v_\varepsilon(\alpha).$$

Now we can set

$$S_i^{\varepsilon,\xi} := \left(\overline{A}_{i+1} \setminus A_i\right)^{\varepsilon,\xi} \cap A$$

so that

$$R_\varepsilon^\xi(A) \subseteq R_\varepsilon^\xi(A_i) \cup R_\varepsilon^\xi(A \setminus \overline{A}_{i+1}) \cup R_\varepsilon^\xi\left(S_i^{\varepsilon,\xi}\right).$$

Let $\delta > 0$, and let $M_\delta > 0$ be such that $\limsup_{\varepsilon\to 0^+} \sum_{|\xi|>M_\delta} C_\varepsilon^\xi < \delta$. Then, by reasoning as in the proof of Proposition 3.7, for ε small enough, we get

$$\begin{aligned}
&F_\varepsilon(w_\varepsilon^i, A) \le F_\varepsilon(u_\varepsilon, A') + F_\varepsilon(v_\varepsilon, A \setminus \overline{A''}) \\
&+\ C \sum_{|\xi|\le M_\delta} C_\varepsilon^\xi \sum_{\alpha\in R_\varepsilon^\xi(S_i^{\varepsilon,\xi})} \varepsilon^N \left(|D_\varepsilon^\xi u_\varepsilon(\alpha)|^p + |D_\varepsilon^\xi v_\varepsilon(\alpha)|^p + N^p\,|u_\varepsilon(\alpha) - v_\varepsilon(\alpha)|^p + 1\right) \\
&+\ C \sum_{M_\delta<|\xi|\le \frac{d}{2N\varepsilon}} C_\varepsilon^\xi \sum_{\alpha\in R_\varepsilon^\xi(S_N)} \varepsilon^N \left(|D_\varepsilon^\xi u_\varepsilon(\alpha)|^p + |D_\varepsilon^\xi v_\varepsilon(\alpha)|^p + N^p\,|u_\varepsilon(\alpha) - v_\varepsilon(\alpha)|^p + 1\right) \\
&+\ CN^p \sum_{|\xi|>\frac{d}{2N\varepsilon}} C_\varepsilon^\xi \left(\|u_\varepsilon\|^p_{L^p(\Omega;\mathbb{R}^d)} + \|v_\varepsilon\|^p_{L^p(\tilde\Omega;\mathbb{R}^d)} + 1\right).
\end{aligned}$$

Since u_ε and v_ε satisfy (3.12), (3.13), (3.14), and (3.15) with A_ε replaced by A'_ε and B_ε by $(A \setminus \overline{A''_\varepsilon})$, then we can choose $i(\varepsilon) \in \{1,\dots,N-3\}$ such that

$$\begin{aligned}
F_\varepsilon(w_\varepsilon^{i(\varepsilon)}, A) &\le \frac{1}{N-3}\sum_{i=1}^{N-3} F_\varepsilon(w_\varepsilon^i, A) \\
&\le F_\varepsilon(u_\varepsilon, A') + C\delta + \frac{C}{N-3}\left(1 + N^p O(\varepsilon)\right) \\
&+C(\delta + O(\varepsilon))\left(1 + N^p O(\varepsilon)\right) + CN^p(\delta + O(\varepsilon)).
\end{aligned} \tag{3.23}$$

Then, since $w_\varepsilon^{i(\varepsilon)}$ still converges to u in $L^p(\Omega;\mathbb{R}^d)$, by (3.23), letting $\varepsilon \to 0^+$, we get

$$F''(u, A) \le \sup_{A'\subset\subset A} F''(u, A') + C\left(\frac{1}{N-3} + \delta + \delta N^p\right).$$

Eventually, letting first $\delta \to 0^+$ and then $N \to +\infty$, we obtain the thesis. □

The following proposition asserts that $F''(\cdot,\cdot)$ satisfies hypothesis (i) of Theorem 2.2. The argument we use for the proof is still the same one exploited in the last two propositions.

PROPOSITION 3.9. *Let* $\{f_\varepsilon^\xi\}_{\varepsilon,\xi}$ *satisfy* (3.2), (3.3), *and let* (H1)–(H2) *hold. Then for any* $A \in \mathcal{A}(\Omega)$ *and for any* $u,\ v \in W^{1,p}(\Omega;\mathbb{R}^d)$ *such that* $u = v$ *a.e. there holds*

$$F''(u,A) = F''(v,A).$$

Proof. Thanks to Proposition 3.8, we may assume that $A \in \mathcal{A}_0(\Omega)$. We first prove

$$F''(u,A) \geq F''(v,A). \tag{3.24}$$

Once more we apply the argument used in the previous proposition. Given $\delta > 0$, there exists $A_\delta \subset\subset A$ such that

$$|A \setminus \overline{A_\delta}| + \|\nabla u\|^p_{L^p(A\setminus\overline{A_\delta})} \leq \delta.$$

Let $v_\varepsilon \in \mathcal{A}_\varepsilon(\Omega)$ and $u_\varepsilon \in \mathcal{A}_\varepsilon(\Omega)$ be such that

$$v_\varepsilon \to v \text{ in } L^p(\Omega;\mathbb{R}^d), \tag{3.25}$$

$$u_\varepsilon \to u \text{ in } L^p(\Omega;\mathbb{R}^d), \tag{3.26}$$

and

$$\limsup_{\varepsilon\to 0^+} F_\varepsilon(u_\varepsilon, A) = F''(u,A),$$

$$\limsup_{\varepsilon\to 0^+} F_\varepsilon(v_\varepsilon, A\setminus\overline{A_\delta}) = F''(v, A\setminus\overline{A_\delta}) \leq C\left(|A\setminus\overline{A_\delta}| + \|\nabla u\|^p_{L^p(A\setminus\overline{A_\delta})}\right) \leq C\delta. \tag{3.27}$$

Set

$$d := \operatorname{dist}(A_\delta, A^c),$$

and for any $i \in \{1,\dots,N\}$ define

$$A_i := \left\{x \in A :\ \operatorname{dist}(x,A_\delta) < i\frac{d}{N}\right\}.$$

Let φ_i be a cut-off function between A_i and A_{i+1}, with $\|\nabla\varphi_i\|_\infty \leq 2\frac{N}{d}$. Then for any $i \in \{1,\dots,N\}$ consider the family of functions $w_\varepsilon^i \in \mathcal{A}_\varepsilon(\Omega)$ converging to v in $L^p(\Omega;\mathbb{R}^d)$ defined as

$$w_\varepsilon^i(\alpha) := \varphi_i(\alpha)u_\varepsilon(\alpha) + (1-\varphi_i(\alpha))\,v_\varepsilon(\alpha).$$

Then, following the same steps as in the proofs of Propositions 3.7 and 3.8, we can choose $i(\varepsilon) \in \{1,\dots,N-3\}$ such that

$$\begin{aligned} F_\varepsilon(w_\varepsilon^{i(\varepsilon)}, A) &\leq \frac{1}{N-3}\sum_{i=1}^{N-3} F_\varepsilon(w_\varepsilon^i, A) \\ &\leq F_\varepsilon(u_\varepsilon, A) + C\delta + \frac{C}{N-3}\left(1 + N^p O(\varepsilon)\right) \\ &\quad + C(\delta + O(\varepsilon))\left(1 + N^p O(\varepsilon)\right) + C(\delta + O(\varepsilon))N^p. \end{aligned} \tag{3.28}$$

Then, since $w_\varepsilon^{i(\varepsilon)}$ still converges to v in $L^p(\Omega;\mathbb{R}^d)$, by (3.28), letting $\varepsilon \to 0^+$, we get

$$F''(v,A) \le F''(u,A) + C\left(\frac{1}{N-3} + \delta + \delta N^p\right).$$

Eventually, letting first $\delta \to 0^+$ and then $N \to +\infty$, we obtain (3.24). Reversing the roles of u and v we obtain the thesis. □

Proof of Theorems 3.1 *and* 3.3. By the compactness property of the Γ-convergence and by Proposition 3.8, there exists a subsequence (ε_{j_k}) such that, for any $(u,A) \in W^{1,p}(\Omega;\mathbb{R}^d) \times \mathcal{A}(\Omega)$, there holds

$$\Gamma(L^p)\text{-}\lim_k F_{\varepsilon_{j_k}}(u,A) := F(u,A)$$

(see [6, Theorem 10.3]). Moreover, by Proposition 3.4,

$$\Gamma(L^p)\text{-}\lim_k F_{\varepsilon_{j_k}}(u) = +\infty$$

for $u \in L^p(\Omega;\mathbb{R}^d) \setminus W^{1,p}(\Omega;\mathbb{R}^d)$. So far, it suffices to check that, for every $(u,A) \in W^{1,p}(\Omega;\mathbb{R}^d) \times \mathcal{A}(\Omega)$, $F(u,A)$ satisfies all the hypotheses of Theorem 2.2. In fact, it can be easily seen that the superadditivity property of $F_\varepsilon(u,\cdot)$ is conserved in the limit. Thus, as an easy consequence of Propositions 3.5, 3.7, 3.8, and 3.9 and thanks to the De Giorgi–Letta criterion (see [17], [6]), hypotheses (i), (ii), and (iii) hold true. Moreover, as $F_\varepsilon(u,A)$ depends on u only through its difference quotients, hypothesis (iv) is satisfied, and, finally, by the lower semicontinuity property of the Γ-limit, also hypothesis (v) is fulfilled. □

3.1. Convergence of minimum problems. In order to treat minimum problems with boundary data, we also derive a compactness theorem in case that our functionals are subject to Dirichlet boundary conditions.

Given $\varphi \in Lip\,(\mathbb{R}^N)$ and $l \in \mathbb{N}$, set, for any $\varepsilon > 0$ and $A \in \mathcal{A}(\Omega)$,

$$\mathcal{A}^l_{\varepsilon,\varphi}(A) := \{u \in \mathcal{A}_\varepsilon(\mathbb{R}^N) : u(\alpha) = \varphi(\alpha) \text{ if } \big(\alpha + [-l\varepsilon, l\varepsilon]^N\big) \cap A^c \neq \emptyset\big)\}. \tag{3.29}$$

Then define $F^{\varphi,l}_\varepsilon : L^p(\Omega;\mathbb{R}^d) \times \mathcal{A}(\Omega) \to [0,+\infty]$ as

$$F^{\varphi,l}_\varepsilon(u,A) = \begin{cases} F_\varepsilon(u,A) & \text{if } u \in \mathcal{A}^l_{\varepsilon,\varphi}(A), \\ +\infty & \text{otherwise.} \end{cases} \tag{3.30}$$

By simplicity of notation we set $\mathcal{A}_{\varepsilon,\varphi}(A) := \mathcal{A}^1_{\varepsilon,\varphi}(A)$ and $F^\varphi_\varepsilon := F^{\varphi,1}_\varepsilon$.

THEOREM 3.10. *Let* $\{f^\xi_\varepsilon\}_{\varepsilon,\xi}$ *satisfy* (3.2), (3.3), *and let* (H1)–(H2) *hold. Given* (ε_j), *a sequence of positive real numbers converging to* 0, *let* (ε_{j_k}) *and* f *be as in Theorem* 3.1. *For any* $\varphi \in Lip\,(\mathbb{R}^N)$, *let* $F^\varphi : L^p(\Omega;\mathbb{R}^d) \times \mathcal{A}(\Omega) \to [0,+\infty]$ *be defined as*

$$F^\varphi(u,A) = \begin{cases} \displaystyle\int_A f(x,\nabla u)\,dx & \text{if } u - \varphi \in W^{1,p}_0(A;\mathbb{R}^d), \\ +\infty & \text{otherwise.} \end{cases}$$

Then, for any $A \in \mathcal{A}(\Omega)$ *with Lipschitz boundary and* $l \in \mathbb{N}$, $(F^{\varphi,l}_{\varepsilon_{j_k}}(\cdot,A))$ Γ-*converges with respect to the* $L^p(\Omega;\mathbb{R}^d)$-*topology to the functional* $F^\varphi(\cdot,A)$.

Proof. For the sake of simplicity we prove the theorem with $l = 1$, the proof being the same in the other cases. Let us first prove the Γ-liminf inequality. Let (u_k) be a sequence of functions belonging to $\mathcal{A}_{\varepsilon_{j_k},\varphi}(A)$ converging to u in the L^p-topology such that

$$\liminf_k F^{\varphi}_{\varepsilon_{j_k}}(u_k, A) = \lim_k F^{\varphi}_{\varepsilon_{j_k}}(u_k, A) < +\infty.$$

Then, from (3.2), we get in particular that

$$\sup_k \sum_{i=1}^{N} \sum_{\alpha \in R^{e_i}_{\varepsilon_{j_k}}(A)} \varepsilon_{j_k}^N |D^{e_i}_{\varepsilon_{j_k}} u_n(\alpha)|^p < +\infty. \tag{3.31}$$

Thanks to the boundary conditions on u_k it is easy to deduce that

$$\sup_k \sum_{i=1}^{N} \sum_{\alpha \in R^{e_i}_{\varepsilon_{j_k}}(\Omega)} \varepsilon^N |D^{e_i}_{\varepsilon_{j_k}} u_n(\alpha)|^p < +\infty.$$

Then, by reasoning as in the proof of Proposition 3.4, we can prove that $u \in W^{1,p}(\Omega; \mathbb{R}^d)$, and, since (u_k) converge to φ in $L^p(\Omega \setminus A; \mathbb{R}^d)$, we get that $u - \varphi \in W^{1,p}_0(A; \mathbb{R}^d)$. By Theorem 3.3 one has

$$\liminf_k F^{\varphi}_{\varepsilon_{j_k}}(u_k, A) = \liminf_k F_{\varepsilon_{j_k}}(u_k, A) \geq F^{\varphi}(u, A).$$

To prove the Γ-limsup inequality, let us first consider $u \in W^{1,p}(\Omega; \mathbb{R}^d)$ such that supp $(u - \varphi) \subset\subset A$. Let $u_k \in \mathcal{A}_{\varepsilon_{j_k}}(\Omega)$ be such that (u_k) converges to u in $L^p(\Omega; \mathbb{R}^d)$, and

$$\limsup_k F_{\varepsilon_{j_k}}(u_k, A) = F^{\varphi}(u, A).$$

Then, by reasoning as in the proof of Proposition 3.8, given $\delta > 0$, we can find suitable cut-off functions ϕ_k with supp $(u - \varphi) \subset\subset$ supp $\phi_k \subset\subset A$ such that if we set

$$v_k(\alpha) := \phi_k(\alpha) u_k(\alpha) + (1 - \phi_k(\alpha))\varphi(\alpha),$$

then (v_k) still converges to u in $L^p(\Omega; \mathbb{R}^d)$, $v_k \in \mathcal{A}_{\varepsilon_{j_k},\varphi}(\Omega)$ for k large enough, and

$$\limsup_k F_{\varepsilon_{j_k}}(v_k, A) \leq \limsup F_{\varepsilon_{j_k}}(u_k, A) + \delta.$$

Thus, thanks to the definition of Γ-limsup, we have

$$\Gamma\text{-limsup} F^{\varphi}_{\varepsilon_{j_k}}(u, A) \leq F^{\varphi}(u, A) + \delta.$$

By the arbitrariness of δ, we obtain the required inequality. In the general case the thesis follows by a density argument, thanks to the lower semicontinuity of Γ-limsup and to the continuity of F with respect to the strong convergence in $W^{1,p}(\Omega; \mathbb{R}^d)$. □

As a consequence of the previous theorem we derive the following result about the convergence of minimum problems with boundary data.

COROLLARY 3.11. *Under the hypotheses of Theorem* 3.10 *we get that, for any* $\varphi \in Lip\ (\mathbb{R}^N)$, $l \in \mathbb{N}$ *and* $A \in \mathcal{A}(\Omega)$ *with Lipschitz boundary,*

$$\liminf_k \{F_{\varepsilon_{j_k}}(u,A):\ u \in \mathcal{A}^l_{\varepsilon_{j_k},\varphi}\} = \min\{F(u,A):\ u-\varphi \in W^{1,p}_0(A;\mathbb{R}^d)\}.$$

Moreover, if (u_k) *is a converging sequence such that*

$$\lim_k F_{\varepsilon_{j_k}}(u_k,A) = \liminf_k \{F_{\varepsilon_{j_k}}(u,A):\ u \in \mathcal{A}^l_{\varepsilon_{j_k},\varphi}\},$$

then its limit is a minimizer for $\min\{F(u,A):\ u-\varphi \in W^{1,p}_0(A;\mathbb{R}^d)\}$.

Proof. Let (u_k) be a sequence such that $F_{\varepsilon_{j_k}}(u_k,A) < +\infty$. Then, by (3.2) and by the boundary conditions on u_k, it is easy to show that

$$\sup_n \sum_{i=1}^N \sum_{\alpha \in \varepsilon_n \mathbb{Z}^N \cap K} \varepsilon^N |D^{e_i}_{\varepsilon_{j_k}} u_k(\alpha)|^p < +\infty$$

for any compact set K of $\mathbb{R}^N$. By virtue of this property, up to passing to a continuous extension of u_k vanishing outside a bounded open set containing Ω, we get

$$\lim_{|h|\to 0} \sup_k \|\tau_h u_k - u_k\|_{L^p(\mathbb{R}^N;\mathbb{R}^d)} = 0,$$

where we have set

$$(\tau_h u)(x) := u(x+h), \quad x \in \mathbb{R}^N, \quad h \in \mathbb{R}^N.$$

Then, by the Frechét–Kolmogorov theorem, there exists a subsequence (u_{k_n}) converging in $L^p(\Omega;\mathbb{R}^d)$ to a function $u \in L^p(\Omega;\mathbb{R}^d)$. Arguing as in the previous proof it is easy to show that $u-\varphi \in W^{1,p}_0(\Omega)$. The thesis follows, thanks to Theorem 3.10 and Theorem 2.1. □

We can also derive the analogue of Theorem 3.10 and Corollary 3.11 about the convergence of minimum problems with periodic conditions.

Let $\mathcal{Q}(\Omega)$ be the family of all open N-cubes contained in Ω. For any $\varepsilon > 0$, $r > 0$, $Q = (x_0, x_0+r)^N \in \mathcal{Q}(\Omega)$, and $\varphi \in Lip\ (\mathbb{R}^N)$, set

$$r_\varepsilon = \varepsilon\left(\left[\frac{r}{\varepsilon}\right] - 2\right),$$

$$\mathcal{A}^{\#}_{\varepsilon,\varphi}(Q) = \left\{u \in \mathcal{A}_\varepsilon(\mathbb{R}^N):\ u-\hat{\varphi}\ \ r_\varepsilon - \text{periodic}\right\},$$

where $\hat{\varphi} \in \mathcal{A}_\varepsilon(\mathbb{R}^N)$, $\hat{\varphi}(\alpha) = \varphi(\alpha)$ for any $\alpha \in \varepsilon\mathbb{Z}^N$. Then define $F^{\varphi,\#}_\varepsilon : L^p(\Omega;\mathbb{R}^d) \times \mathcal{Q}(\Omega) \to [0,+\infty]$ as

$$F^{\varphi,\#}_\varepsilon(u,Q) = \begin{cases} F_\varepsilon(u,Q) & \text{if } u \in \mathcal{A}^{\#}_{\varepsilon,\varphi}(Q), \\ +\infty & \text{otherwise.} \end{cases} \tag{3.32}$$

THEOREM 3.12. *Let* $\{f^\xi_\varepsilon\}_{\varepsilon,\xi}$ *satisfy* (3.2), (3.3), *and let* (H1)–(H2) *hold. Given* (ε_j), *a sequence of positive real numbers converging to* 0, *let* (ε_{j_k}) *and* f *be as in*

Theorem 3.1. *Then, for any* $\varphi \in Lip\ (\mathbb{R}^N)$, *let* $F^{\#} : L^p(\Omega;\mathbb{R}^d) \times \mathcal{Q}(\Omega) \to [0,+\infty]$ *be defined as*

$$F^{\varphi,\#}(u,Q) = \begin{cases} \displaystyle\int_Q f(x,\nabla u)\,dx & \text{if } u \in W^{1,p}_{\#}(Q;\mathbb{R}^d), \\ +\infty & \text{otherwise.} \end{cases}$$

Then, for any $Q \in \mathcal{Q}(\Omega)$, $(F^{\varphi,\#}_{\varepsilon_{j_k}}(\cdot,Q))$ Γ-*converges with respect to the* $L^p(\Omega;\mathbb{R}^d)$-*topology to the functional* $F^{\varphi,\#}(u,Q)$.

Proof. To prove the Γ-liminf inequality, let (u_k) be a sequence of functions belonging to $\mathcal{A}^{\#}_{\varepsilon_{j_k},\varphi}(Q)$ converging to u in the L^p-topology such that

$$\liminf_k F^{\varphi,\#}_{\varepsilon_{j_k}}(u_k,Q) = \lim_k F^{\varphi,\#}_{\varepsilon_{j_k}}(u_k,Q) < +\infty.$$

Then, arguing as in the proof of Theorem 3.10 and observing that $r_\varepsilon \to r$, we can conclude that $u-\varphi \in W^{1,p}_{\#}(Q;\mathbb{R}^d)$, and

$$\liminf_k F^{\varphi,\#}_{\varepsilon_{j_k}}(u_k,Q) \geq F^{\varphi,\#}(u,Q).$$

By a density argument it suffices to prove the Γ-limsup inequality for u such that $u-\varphi \in W^{1,\infty}_{\#}(Q';\mathbb{R}^d)$ for any open N-cube Q' such that $(x_0+\delta, x_0+r-\delta) \subseteq Q' \subseteq Q$ for some $\delta > 0$. Note that, for such a u, $\mathcal{A}_{\varepsilon_{j_k},u} \subseteq \mathcal{A}^{\#}_{\varepsilon_{j_k},\varphi}$ for k large enough. Then the existence of a recovery sequence is ensured by Theorem 3.10. □

As a consequence of the previous theorem, by reasoning as in the proof of Corollary 3.11 one can prove the following result.

Corollary 3.13. *Under the hypotheses of Theorem* 3.12 *we get that, for any* $\varphi \in Lip\ (\mathbb{R}^N)$ *and* $Q \in \mathcal{Q}(\Omega)$,

$$\liminf_k \{F_{\varepsilon_{j_k}}(u,Q) :\ u \in \mathcal{A}^{\#}_{\varepsilon_{j_k},\varphi}(Q)\} = \min\{F(u,Q) :\ u-\varphi \in W^{1,p}_{\#}(Q;\mathbb{R}^d)\}.$$

Moreover, if (u_k) *is a converging sequence such that*

$$\lim_k F_{\varepsilon_{j_k}}(u_k,Q) = \liminf_k \{F_{\varepsilon_{j_k}}(u,Q) :\ u \in \mathcal{A}^{\#}_{\varepsilon_{j_k},\varphi}\},$$

then its limit is a minimizer for $\min\{F(u,Q) :\ u-\varphi \in W^{1,p}_{\#}(Q;\mathbb{R}^d)\}$.

4. Homogenization. In this section we will show that if the functions f^ξ_ε are obtained by rescaling by ε functions f^ξ periodic in the space variable, then a Γ-convergence result holds true. This models the case when the arrangement of the "material points" presents a periodic feature (see Figure 4).

Let $\mathbf{k} = (k_1,\dots,k_N) \in \mathbf{Z}^N$ be given, and set

$$\mathcal{R}_{\mathbf{k}} := (0,k_1) \times \cdots \times (0,k_N).$$

For any $\xi \in \mathbf{Z}^N$, let $f^\xi : \mathbf{Z}^N \times \mathbf{R}^d \to [0,+\infty)$ be such that $f^\xi(\cdot,z)$ is $\mathcal{R}_{\mathbf{k}}$-periodic for any $z \in \mathbf{R}^d$. Then we consider f^ξ_ε of the following form:

$$f^\xi_\varepsilon(\alpha,z) := f^\xi\left(\frac{\alpha}{\varepsilon},z\right). \tag{4.1}$$

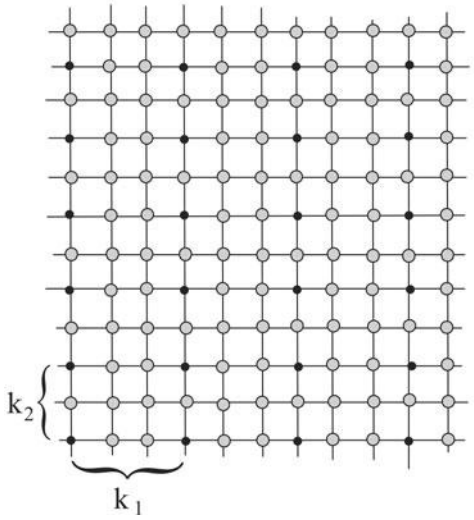

FIG. 4. *Example of periodic structure.*

In this case, the growth conditions (3.2) and (3.3) and hypotheses (H1) and (H2) can be rewritten as follows:

$$f^{e_i}(\alpha, z) \geq c_1(|z|^p - 1) \quad \forall i \in \{1, \dots, N\}, \tag{4.2}$$

$$f^{\xi}(\alpha, z) \leq C^{\xi}(|z|^p + 1), \tag{4.3}$$

where

$$\sum_{\xi \in \mathbf{Z}^N} C^{\xi} < +\infty. \tag{H3}$$

In what follows we will use the following notation: for any $x = (x_1, \dots, x_N) \in \mathbf{R}^N$ define

$$[x]_{\mathbf{k}} := \left(\left[\frac{x_1}{k_1} \right] k_1, \dots, \left[\frac{x_N}{k_N} \right] k_N \right).$$

Moreover, for any $A \in \mathcal{A}(\Omega)$, $\varepsilon > 0$, $l \in \mathbf{N}$, and $M \in \mathcal{M}^{d \times N}$ we denote by $\mathcal{A}^l_{\varepsilon,M}(A)$ the set defined in formula (3.29) with $\varphi(x) = Mx$. By simplicity of notation, we set $\mathcal{A}^1_{\varepsilon,M}(A) := \mathcal{A}_{\varepsilon,M}(A)$. Finally, for every $r > 0$ we set $Q_r := (0, r)^N$.

The following theorem is the main result of this section, and its proof is obtained by adapting a homogenization argument to the discrete setting. We remark that a central role is played by Theorems 3.1 and 3.3 and by the convergence of minimum problems with boundary data stated in Corollary 3.11. Moreover, we recall that the following result has been already proven in [11] in the one-dimensional case, where a more straightforward proof is possible.

THEOREM 4.1. *Let* $\{f^{\xi}_{\varepsilon}\}_{\varepsilon,\xi}$ *satisfy* (4.1)–(4.3)*, and let* (H3) *hold. Then* (F_ε) Γ*-converges with respect to the* $L^p(\Omega; \mathbf{R}^d)$*-topology to the functional* $F : L^p(\Omega; \mathbf{R}^d) \to [0, +\infty]$ *defined as*

$$F(u) = \begin{cases} \displaystyle\int_\Omega f_{hom}(\nabla u)\, dx & \text{if } u \in W^{1,p}(\Omega; \mathbf{R}^d), \\ +\infty & \text{otherwise,} \end{cases} \tag{4.4}$$

where $f_{hom} : \mathcal{M}^{d\times N} \to [0,+\infty)$ is given by the following homogenization formula:

$$(4.5)\quad f_{hom}(M) := \lim_{h\to+\infty} \frac{1}{h^N} \min\left\{ \sum_{\xi\in\mathbf{Z}^N} \sum_{\beta\in R_1^\xi(Q_h)} f^\xi(\beta, D_1^\xi v(\beta)),\quad v \in \mathcal{A}_{1,M}(Q_h) \right\}.$$

Proof. Let (ε_n) be a sequence of positive numbers converging to 0. Then, by Theorems 3.1 and 3.3, we can extract a subsequence (not relabelled) such that (F_{ε_n}) Γ-converges to a functional F defined as in (3.4) and such that, for any $u \in W^{1,p}(\Omega;\mathbf{R}^d)$, $A \in \mathcal{A}(\Omega)$,

$$\Gamma\text{-}\lim_n F_{\varepsilon_n}(u,A) = \int_A f(x,\nabla u)\,dx.$$

The theorem is proved if we show that f does not depend on the space variable x and $f \equiv f_{hom}$. To prove the first claim, by Theorem 2.2, it suffices to show that if we set

$$F(u,A) = \int_A f(x,\nabla u)\,dx,$$

then

$$F(Mx, B(y,\rho)) = F(Mx, B(z,\rho))$$

for all $M \in \mathcal{M}^{d\times N}$, $y,z \in \Omega$ and $\rho > 0$ such that $B(y,\rho)\cup B(z,\rho) \subset \Omega$. We will prove that

$$F(Mx, B(y,\rho)) \le F(Mx, B(z,\rho)),$$

the proof of the opposite inequality being analogous. By the inner regularity of $F(Mx,\cdot)$, given by Proposition 3.8, it suffices to show that for any $\rho' < \rho$ we get

$$F(Mx, B(y,\rho')) \le F(Mx, B(z,\rho)). \tag{4.6}$$

Then let $v_n \in \mathcal{A}_{\varepsilon_n}(\Omega)$ be such that (v_n) converges to Mx in $L^p(\Omega;\mathbf{R}^d)$, and

$$\lim_n F_{\varepsilon_n}(v_n, B(z,\rho)) = F(Mx, B(z,\rho)). \tag{4.7}$$

For $n \in \mathbf{N}$, define $u_n \in \mathcal{A}_{\varepsilon_n}(\Omega)$ as

$$u_n(\alpha) := \begin{cases} v_n\left(\alpha - \varepsilon_n\left[\frac{y-z}{\varepsilon_n}\right]_{\mathbf{k}}\right) + \varepsilon_n M\left[\frac{y-z}{\varepsilon_n}\right]_{\mathbf{k}} & \text{if } \alpha \in \varepsilon_n\mathbf{Z}^N \cap B(y,\rho'),\\ M\alpha & \text{otherwise.}\end{cases}$$

Then it is easy to verify that (u_n) converges to Mx in $L^p(\Omega;\mathbf{R}^d)$. Moreover, for n large enough

$$R^\xi_{\varepsilon_n}(B(y,\rho')) - \varepsilon_n\left[\frac{y-z}{\varepsilon_n}\right]_{\mathbf{k}} \subseteq R^\xi_{\varepsilon_n}(B(z,\rho)).$$

Thus, since, by the periodicity hypothesis, $f^\xi(\alpha - \varepsilon_n[\frac{y-z}{\varepsilon_n}]_{\mathbf{k}}, z) = f^\xi(\alpha,z)$ and $D^\xi_\varepsilon u_n(\alpha)$ $= D^\xi_\varepsilon v_n(\alpha - \varepsilon_n[\frac{y-z}{\varepsilon_n}]_{\mathbf{k}})$, we get for n large enough

$$F_{\varepsilon_n}(u_n, B(y,\rho')) \le F_{\varepsilon_n}(v_n, B(z,\rho)).$$

Eventually, by (4.7), we obtain

$$\begin{aligned} F(Mx, B(y,\rho')) &\le \liminf_{n\to+\infty} F_{\varepsilon_n}(u_n, B(y,\rho')) \\ &\le \lim_{n\to+\infty} F_{\varepsilon_n}(v_n, B(z,\rho)) = F(Mx, B(z,\rho)). \end{aligned}$$

In order to prove that $f \equiv f_{hom}$, first note that, by the lower semicontinuity of F in $W^{1,p}(\Omega;\mathbf{R}^d)$, f is quasi-convex so that, by the p-growth properties of f, for any $A \in \mathcal{A}(\Omega)$ with Lipschitz boundary and for any $M \in \mathcal{M}^{d\times N}$ there holds

$$\begin{aligned} f(M) &= \frac{1}{|A|}\min\left\{\int_A f(\nabla u)\,dx : \quad u - Mx \in W_0^{1,p}(A;\mathbf{R}^d)\right\} \\ &= \frac{1}{|A|}\min\left\{F(u,A) : \quad u - Mx \in W_0^{1,p}(A;\mathbf{R}^d)\right\} \\ &= \frac{1}{|A|}\liminf_n \left\{F_{\varepsilon_n}(u,A) : \quad u \in \mathcal{A}_{\varepsilon_n,M}(A)\right\}, \end{aligned}$$

where the last equality follows by Corollary 3.11. In particular, if $x_0 \in \Omega$ and $r > 0$ are such that $Q_r(x_0) := (x_0, x_0 + r)^N \subseteq \Omega$, then

$$f(M) = \lim_n \frac{1}{r^N}\inf\{F_{\varepsilon_n}(u, Q_r(x_0)) :\ u \in \mathcal{A}_{\varepsilon_n,M}(Q_r(x_0))\}.$$

Without loss of generality, we may suppose $x_0 = 0$. If we set

$$T_n := \left[\frac{r}{\varepsilon_n}\right] + 1,$$

then it is easy to show that $\mathcal{A}_{\varepsilon_n,M}(Q_r) = \mathcal{A}_{\varepsilon_n,M}(Q_{\varepsilon_n T_n})$ and that for $\xi \in \mathbf{Z}^N$ $R^\xi_{\varepsilon_n}(Q_r) = R^\xi_{\varepsilon_n}(Q_{\varepsilon_n T_n})$. Thus

$$f(M) = \lim_n \frac{1}{r^N}\inf\{F_{\varepsilon_n}(u, Q(0,\varepsilon_n T_n)) :\ u \in \mathcal{A}_{\varepsilon_n,M}(Q_{\varepsilon_n T_n})\}.$$

Eventually, through the change of variable

$$\beta = \frac{\alpha}{\varepsilon}, \quad v(\beta) = \frac{1}{\varepsilon}u(\varepsilon\beta), \tag{4.8}$$

we get

$$\begin{aligned} f(M) &= \lim_n \left(\frac{\varepsilon_n}{r}\right)^N \inf\left\{\sum_{\xi\in\mathbf{Z}^N}\sum_{\beta\in R_1^\xi(Q_{T_n})} f^\xi(\beta, D_1^\xi v(\beta)), \quad v \in \mathcal{A}_{1,M}(Q_{T_n})\right\} \\ &= \lim_n \frac{1}{T_n^N}\inf\left\{\sum_{\xi\in\mathbf{Z}^N}\sum_{\beta\in R_1^\xi(Q_{T_n})} f^\xi(\beta, D_1^\xi v(\beta)), \quad v \in \mathcal{A}_{1,M}(Q_{T_n})\right\}, \end{aligned}$$

where the last equality holds since

$$\lim_n T_n \frac{\varepsilon_n}{r} = 1.$$

Then the thesis will follow by the next proposition.

PROPOSITION 4.2. *Let f^ξ satisfy* (4.2), (4.3), *and* (H3) *for any $\xi \in \mathbf{Z}^N$. Then the limit*

$$\lim_{h\to+\infty} \frac{1}{h^N} \inf \left\{ \sum_{\xi\in\mathbf{Z}^N} \sum_{\beta\in R_1^\xi(Q_h)} f^\xi(\beta, D_1^\xi v(\beta)), \quad v \in \mathcal{A}_{1,M}(Q_h) \right\}$$

exists for all $M \in \mathcal{M}^{d\times N}$.

Proof. Let $M \in \mathcal{M}^{d\times N}$ be fixed, and set

$$F_1(v, A) := \sum_{\xi\in\mathbf{Z}^N} \sum_{\beta\in R_1^\xi(A)} f^\xi(\beta, D_1^\xi v(\beta)),$$

$$f_h(M) := \frac{1}{h^N} \inf \{F_1(v, Q_h), \quad v \in \mathcal{A}_{1,M}(Q_h)\}.$$

Moreover, for any $R > 0$, set

$$F_1^R(v, A) := \sum_{|\xi|\le R} \sum_{\beta\in R_1^\xi(A)} f^\xi(\beta, D_1^\xi v(\beta)),$$

$$f_h^R(M) := \frac{1}{h^N} \inf \left\{F_1^R(v, Q_h), \quad v \in \mathcal{A}_{1,M}(Q_h)\right\}.$$

We prove that

$$\lim_{R\to+\infty} \sup_h |f_h^R(M) - f_h(M)| = 0. \tag{4.9}$$

To this end, since $f_h^R(M) \le f_h(M)$ for any $h \in \mathbf{N}$ and $R > 0$, it suffices to prove that for any $\delta > 0$, there exist $R_\delta > 0$ such that

$$f_h(M) \le f_h^R(M) + \delta \quad \forall R > R_\delta, \quad h \in \mathbf{N}.$$

Fix $\delta > 0$, and let $v_h^R \in \mathcal{A}_{1,M}(Q_h)$ be such that

$$\frac{1}{h^N} F_h^R(v_h^R, Q_h) \le f_h^R(M) + \frac{1}{R}. \tag{4.10}$$

By testing the minimum problem defining $f_h^R(M)$ with $v(\alpha) = M\alpha$, we get, by (4.3) and (H3), that

$$f_h^R(M) \le \frac{1}{h^N} F_h^R(M\alpha, Q_h) \le C|M|^p.$$

Thus, by (4.10) and (4.2), we obtain that

$$\sup_{h,R} \frac{1}{h^N} \sum_{i=1}^N \sum_{\beta\in R_1^{e_i}(Q_h)} |D_1^{e_i} v_h^R(\beta)|^p < +\infty.$$

Then, by arguing as in the proof of Lemma 3.6 and thanks to the particular geometry of the sets Q_h, we deduce that

$$\sup_{h,R} \frac{1}{h^N} \sup_{\xi\in\mathbf{Z}^N} \sum_{\beta\in R_1^\xi(Q_h)} |D_1^\xi v_h^R(\beta)|^p < +\infty.$$

Eventually, we have

$$f_h(M) \le \frac{1}{h^N} F_h(v_h^R, Q_h) \le \frac{1}{h^N} F_h^R(v_h^R, Q_h) + \frac{1}{h^N} \sum_{|\xi|>R} C^\xi \sum_{\beta \in R_1^\xi(Q_h)} |D_1^\xi v_h^R(\beta)|^p$$
$$\le f_h^R(M) + \frac{1}{R} + C \sum_{|\xi|>R} C^\xi.$$

Thus, it suffices to choose $R_\delta > 0$ such that for $R > R_\delta$

$$\frac{1}{R} + C \sum_{|\xi|>R} C^\xi \le \delta.$$

So far, in order to prove the thesis, it suffices to show that for any $R > 0$ there exists the limit

$$\lim_h f_h^R(M).$$

Set

$$f_h^{R,R}(M) := \frac{1}{h^N} \inf \left\{ F_1^R(v, Q_h), \quad v \in \mathcal{A}_{1,M}^{[R]}(Q_h) \right\}.$$

Using backward the scaling argument exploited in the proof of the previous proposition and thanks to Theorem 3.10 and Corollary 3.11, one can show that, for any subsequence $(h_n) \subset \mathbf{N}$, it is possible to extract a further subsequence (not relabelled) such that

$$\lim_n f_{h_n}^R(M) = \lim_n f_{h_n}^{R,R}(M). \tag{4.11}$$

Thus, to complete the proof, it is sufficient to prove that there exists the limit

$$\lim_h f_h^{R,R}(M).$$

Let $h \in \mathbf{N}$, and let $v_h \in \mathcal{A}_{1,M}^{[R]}(Q_h)$ be such that

$$\frac{1}{h^N} F_1^R(v_h, Q_h) \le f_h^{R,R}(M) + \frac{1}{h}.$$

For any $k > h$ define a function $u_k \in \mathcal{A}_{1,M}^{[R]}(Q_k)$ as follows:

$$u_k(\alpha) = \begin{cases} v_h(\alpha - h\mathbf{i}) + hM\mathbf{i} & \text{if } \alpha \in h\mathbf{i} + Q_h,\ \mathbf{i} \in \left\{0, \dots, \left[\frac{k}{h}\right] - 1\right\}^N, \\ M\alpha & \text{otherwise.} \end{cases}$$

Note that for any $\xi \in \mathbf{Z}^N$, $|\xi| \le R$ we have

$$R_1^\xi(Q_k) \subseteq \left(\bigcup_{\mathbf{i} \in \left\{0, \dots, \left[\frac{k}{h}\right]-1\right\}^N} R_1^\xi(h\mathbf{i} + Q_h) \right) \cup R_1^\xi \left(Q_k \setminus \bigcup_{\mathbf{i} \in \left\{0, \dots, \left[\frac{k}{h}\right]-1\right\}^N} (h\mathbf{i} + Q_h) \right)$$
$$\cup \left(\bigcup_{\mathbf{i} \in \left\{0, \dots, \left[\frac{k}{h}\right]-1\right\}^N} \left(h\mathbf{i} + \left(\{0, \dots, h+R\}^N \setminus \{0, \dots, h-R\}^N \right) \right) \right).$$

Moreover, $D_1^\xi u_k(\alpha) = M\frac{\xi}{|\xi|}$ if $\alpha \in R_1^\xi(Q_k \setminus \bigcup_{\mathbf{i}\in\{0,\dots,[\frac{k}{h}]-1\}^N}(h\mathbf{i}+Q_h))$ or $\alpha \in \bigcup_{\mathbf{i}\in\{0,\dots,[\frac{k}{h}]-1\}^N}\left(h\mathbf{i}+\left(\{0,\dots,h+R\}^N \setminus \{0,\dots,h-R\}^N\right)\right)$, and

$$\#\left(R_1^\xi\left(Q_k \setminus \bigcup_{\mathbf{i}\in\{0,\dots,[\frac{k}{h}]-1\}^N}(h\mathbf{i}+Q_h)\right)\right) \le k^N - \left[\frac{k}{h}\right]^N h^N,$$

$$\#\left(\{0,\dots,h+R\}^N \setminus \{0,\dots,h-R\}^N\right) \le (h+R)^N - (h-R)^N.$$

Then, by (4.3) and (H3), we get

$$\begin{aligned} f_k^{R,R}(M) &\le \frac{1}{k^N}F_1^R(u_k,Q_k) \le \left[\frac{k}{h}\right]^N \frac{1}{k^N}F_1^R(v_h,Q_h) \\ &+ C|M|^P\frac{1}{k^N}\left(k^N - \left[\frac{k}{h}\right]^N h^N + \left[\frac{k}{h}\right]^N\left((h+R)^N-(h-R)^N\right)\right) \\ &\le \left[\frac{k}{h}\right]^N \frac{h^N}{k^N}\left(f_h^{R,R}(M)+\frac{1}{h}\right) \\ &+ C|M|^P\frac{1}{k^N}\left(k^N - \left[\frac{k}{h}\right]^N h^N + \left[\frac{k}{h}\right]^N\left((h+R)^N-(h-R)^N\right)\right). \end{aligned}$$

By letting k tend to $+\infty$, we then get

$$\limsup_k f_k^{R,R}(M) \le f_h^{R,R}(M) + \frac{1}{h} + C|M|^P\frac{1}{h^N}\left((h+R)^N-(h-R)^N\right).$$

Eventually, letting h tend to $+\infty$, we obtain

$$\limsup_k f_k^{R,R}(M) \le \liminf_h f_h^{R,R}(M),$$

that is, the conclusion. □

Remark 4.3. In formula (4.5) we can replace $\mathcal{A}_{1,M}(Q_h)$ by $\mathcal{A}^l_{1,M}(Q_h)$ for any fixed $l \in \mathbf{N}$, the proof being exactly the same.

Remark 4.4. The function f_{hom} in Theorem 4.1 also satisfies

$$\begin{aligned} f_{hom}(M) = \lim_{h\to+\infty}\frac{1}{h^N}\inf\Bigg\{ &\sum_{\xi\in\mathbf{Z}^N}\sum_{\beta\in R_1^\xi(Q_h)} f^\xi\left(\beta, M\frac{\xi}{|\xi|}+D_1^\xi v(\beta)\right), \\ &v \in \mathcal{A}_{1,\#}(Q_{h-2})\Bigg\}, \end{aligned} \tag{4.12}$$

where, for every $k \in \mathbf{R}$,

$$\mathcal{A}_{1,\#}(Q_k) := \{v \in \mathcal{A}_1(\mathbf{R}^N) : \ v \ k\text{-periodic}\}.$$

This characterization can be proved by arguing as in the proof of Theorem 4.1 and Proposition 4.2, taking into account Corollary 3.13 and recalling that, since f_{hom} is

quasi-convex, there holds

$$f_{hom}(M) = \frac{1}{r^N}\min\left\{\int_{Q_r} f_{hom}(M+\nabla\psi)\,dx : \psi \in W^{1,p}_{\#}(Q_r;\mathbf{R}^d)\right\}$$
$$= \frac{1}{r^N}\min\left\{F(M\alpha+\psi,Q_r) : \psi \in W^{1,p}_{\#}(Q_r;\mathbf{R}^d)\right\}.$$

As a consequence of Theorem 3.10, Corollary 3.11, and Theorem 4.1 we immediately derive the following result about Γ-convergence and convergence of minimum problems for homogeneous functionals subject to Dirichlet boundary conditions.

THEOREM 4.5. *For any $\varphi \in Lip\,(\mathbf{R}^N)$ and $l \in \mathbf{N}$ let $F^{\varphi,l}_\varepsilon$ be defined by* (3.30)*, and let $F^\varphi : L^p(\Omega;\mathbf{R}^d)\times\mathcal{A}(\Omega)\to[0,+\infty]$ be defined as*

$$F^\varphi(u,A) = \begin{cases} \displaystyle\int_A f_{hom}(\nabla u)\,dx & \text{if } u-\varphi \in W^{1,p}_0(A;\mathbf{R}^d),\\ +\infty & \text{otherwise.}\end{cases} \tag{4.13}$$

Under the hypotheses of Theorem 4.1*, $F^\varphi_\varepsilon(\cdot,A)$ Γ-converges with respect to the $L^p(\Omega;\mathbf{R}^d)$-topology to $F^\varphi(\cdot,A)$ for any $A\in\mathcal{A}$.*

COROLLARY 4.6. *Under the hypotheses of Theorem* 4.5*, for any $\varphi \in Lip\,(\mathbf{R}^N)$, $l\in\mathbf{N}$, and $A\in\mathcal{A}(\Omega)$,*

$$\liminf_{\varepsilon\to0}\{F_\varepsilon(u,A) :\ u\in\mathcal{A}^l_{\varepsilon,\varphi}\} = \min\{F(u,A) :\ u-\varphi\in W^{1,p}_0(A;\mathbf{R}^d)\}.$$

Moreover, for any (ε_j) converging to zero as j tends to infinity, if (u_j) is a converging sequence such that

$$\lim_j F_{\varepsilon_j}(u_j,A) = \liminf_j\{F_{\varepsilon_j}(u,A) :\ u\in\mathcal{A}^l_{\varepsilon_j,\varphi}\},$$

then its limit is a minimizer for $\min\{F(u,A) :\ u-\varphi\in W^{1,p}_0(A;\mathbf{R}^d)\}$.

An analogous result about the convergence of minimum problems with periodic conditions follows by Theorem 3.12 and Corollary 3.13.

5. The convex case: A cell problem formula. In this section we will see that in the convex case the function f_{hom} can be rewritten by a single periodic minimization problem on the periodic cell $\mathcal{R}_{\mathbf{k}}$. Set

$$\hat{k} := \prod_{i=1}^N k_i,$$

$$I_{\mathbf{k}} := \prod_{i=1}^N\{0,\dots,k_i-1\},$$

and

$$\mathcal{A}_{1,\#}(\mathcal{R}_{\mathbf{k}}) := \{u\in\mathcal{A}_1(\mathbf{R}^N)\ :\ u \text{ is } \mathcal{R}_{\mathbf{k}}\text{-periodic}\}.$$

THEOREM 5.1. *Let $(f^\xi_\varepsilon)_{\varepsilon,\xi}$ satisfy all the assumptions of Theorem* 4.1*, and in, addition, let $f^\xi_\varepsilon(\alpha,\cdot)$ be convex for all $\alpha\in\varepsilon\mathbf{Z}^N$, $\varepsilon>0$, and $\xi\in\mathbf{Z}^N$. Then the conclusion of Theorem* 4.1 *holds with f_{hom} satisfying*

$$f_{hom}(M) = \frac{1}{\hat{k}}\inf\left\{\sum_{\xi\in\mathbf{Z}^N}\sum_{\beta\in I_{\mathbf{k}}} f^\xi\left(\beta, M\frac{\xi}{|\xi|}+D^\xi_1 v(\beta)\right),\quad v\in\mathcal{A}_{1,\#}\left(\mathcal{R}_{\mathbf{k}}\right)\right\}$$

for all $M \in \mathcal{M}^{d\times N}$.

Proof. Set

$$\overline{f}(M) := \frac{1}{\hat{k}} \inf \left\{ \sum_{\xi\in\mathbf{Z}^N} \sum_{\beta\in I_{\mathbf{k}}} f^{\xi}\left(\beta, M\frac{\xi}{|\xi|} + D_1^{\xi} v(\beta)\right), \quad v \in \mathcal{A}_{1,\#}\left(\mathcal{R}_{\mathbf{k}}\right) \right\}.$$

We first prove that

$$f_{hom}(M) \le \overline{f}(M). \tag{5.1}$$

With fixed $\delta > 0$, let $v \in \mathcal{A}_{1,\#}\left(\mathcal{R}_k\right)$ be such that

$$\frac{1}{\hat{k}} \sum_{\xi\in\mathbf{Z}^N} \sum_{\beta\in I_{\mathbf{k}}} f^{\xi}\left(\beta, M\frac{\xi}{|\xi|} + D_1^{\xi} v(\beta)\right) \le \overline{f}(M) + \delta.$$

$$f_h^{\#}(M) := \inf \left\{ \sum_{xi\in\mathbf{Z}^N} \sum_{\beta\in R_1^{\xi}(Q_h)} f^{\xi}\left(\beta, M\frac{\xi}{|\xi|} + D_1^{\xi} v(\beta)\right), \quad v \in \mathcal{A}_{1,\#}(Q_{h-2}) \right\}.$$

For $n \in \mathbf{N}$, since in particular $v \in \mathcal{A}_{1,\#}(Q_{n\hat{k}})$, we get

$$\begin{aligned} f_{n\hat{k}+2}^{\#}(M) &\le \sum_{\xi\in\mathbf{Z}^N} \sum_{\beta\in R_1^{\xi}(Q_{n\hat{k}})} f^{\xi}\left(\beta, M\frac{\xi}{|\xi|} + D_1^{\xi} v(\beta)\right) \\ &\le n^N \hat{k}^{N-1} \sum_{\xi\in\mathbf{Z}^N} \sum_{\beta\in I_{\mathbf{k}}} f^{\xi}\left(\beta, M\frac{\xi}{|\xi|} + D_1^{\xi} v(\beta)\right), \end{aligned}$$

where the last inequality follows by the periodicity of v, $(f(\cdot, z))$ and by the fact that $Q_{n\hat{k}}$ is the union of $n^N \hat{k}^{N-1}$ periodicity cells. Eventually, by Remark 4.4, we get

$$\begin{aligned} f_{hom}(M) &\le \limsup_n \frac{1}{(n\hat{k}+2)^N} f_{n\hat{k}+2}^{\#}(M) \\ &\le \frac{1}{\hat{k}} \sum_{\xi\in\mathbf{Z}^N} \sum_{\beta\in I_{\mathbf{k}}} f^{\xi}\left(\beta, M\frac{\xi}{|\xi|} + D_1^{\xi} v(\beta)\right) \le \overline{f}(M) + \delta, \end{aligned}$$

and inequality (5.1) follows by letting δ tend to 0. Let us prove that

$$f_{hom}(M) \ge \overline{f}(M).$$

For any $R > 0$, set

$$f_{hom}^R(M) := \lim_{h\to+\infty} \frac{1}{h^N} \inf \left\{ \sum_{|\xi|\le R} \sum_{\beta\in R_1^{\xi}(Q_h)} f^{\xi}(\beta, D_1^{\xi} v(\beta)), \quad v \in \mathcal{A}_{1,M}^{[R]}\left(Q_h\right) \right\},$$

$$\overline{f}^R(M) := \frac{1}{\hat{k}} \inf \left\{ \sum_{|\xi|\le R} \sum_{\beta\in I_{\mathbf{k}}} f^{\xi}\left(\beta, M\frac{\xi}{|\xi|} + D_1^{\xi} v(\beta)\right), \quad v \in \mathcal{A}_{1,\#}\left(\mathcal{R}_{\mathbf{k}}\right) \right\}.$$

By (4.9) and (4.11)we easily derive that

$$\lim_{R\to+\infty} f^R_{hom}(M) = f_{hom}(M).$$

Analogously one can prove that

$$\lim_{R\to+\infty} \overline{f}^R(M) = \overline{f}(M).$$

Thus it suffices to prove that for any $R>0$

$$f^R_{hom}(M) \geq \overline{f}^R(M). \tag{5.2}$$

For $n\in\mathbf{N}$, let $u\in\mathcal{A}^{[R]}_{1,M}(Q_{n\hat{k}})$, and let $v\in\mathcal{A}_{1,\#}(Q_{n\hat{k}})$ be such that

$$v(\alpha) = u(\alpha) - M\alpha \quad \forall \alpha\in Q_{n\hat{k}}.$$

Moreover, set

$$I^n_{\mathbf{k}} := \prod_{i=1}^N \left\{0,\dots,n\prod_{j\neq i}k_j - 1\right\}.$$

Then we get

$$\begin{aligned}
&\frac{1}{(n\hat{k})^N}\sum_{|\xi|\leq R}\sum_{\beta\in R^\xi_1(Q_{n\hat{k}})} f^\xi\left(\beta, M\frac{\xi}{|\xi|}+D^\xi_1 v(\beta)\right)\\
&= \frac{1}{(n\hat{k})^N}\sum_{|\xi|\leq R}\sum_{\beta\in\{0,\dots,n\hat{k}\}^N} f^\xi\left(\beta, M\frac{\xi}{|\xi|}+D^\xi_1 v(\beta)\right) - O\left(\frac{1}{n}\right)\\
&= \frac{1}{\hat{k}}\sum_{|\xi|\leq R}\sum_{\beta\in I_k}\frac{1}{\hat{k}^{N-1}n^N}\sum_{\gamma\in I^n_{\mathbf{k}}} f^\xi\left(\beta, M\frac{\xi}{|\xi|}+D^\xi_1 v\left(\beta+\sum_{i=1}^N\gamma_i k_i e_i\right)\right) - O\left(\frac{1}{n}\right)\\
&\geq \frac{1}{\hat{k}}\sum_{|\xi|\leq R}\sum_{\beta\in I_k} f^\xi\left(\beta, M\frac{\xi}{|\xi|}+\frac{1}{\hat{k}^{N-1}n^N}\sum_{\gamma\in I^n_{\mathbf{k}}} D^\xi_1 v\left(\beta+\sum_{i=1}^N\gamma_i k_i e_i\right)\right) - O\left(\frac{1}{n}\right),
\end{aligned}$$

where in the last inequality we have used the convexity hypothesis on f^ξ. Eventually, set

$$v_n(\beta) := \frac{1}{\hat{k}^{N-1}n^N}\sum_{\gamma\in I^n_{\mathbf{k}}} v\left(\beta+\sum_{i=1}^N\gamma_i k_i e_i\right).$$

It is easy to show that $v_n\in\mathcal{A}_{1,\#}(\mathcal{R}_k)$, and so, by the previous inequality, we get

$$\begin{aligned}
&\frac{1}{(n\hat{k})^N}\sum_{|\xi|\leq R}\sum_{\beta\in R^\xi_1(Q_{n\hat{k}})} f^\xi\left(\beta, M\frac{\xi}{|\xi|}+D^\xi_1 v(\beta)\right)\\
&\quad\geq \frac{1}{\hat{k}}\sum_{|\xi|\leq R}\sum_{\beta\in I_k} f^\xi\left(\beta, M\frac{\xi}{|\xi|}+D^\xi_1 v_n(\beta)\right) - O\left(\frac{1}{n}\right)\\
&\quad\geq \overline{f}^R(M) - O\left(\frac{1}{n}\right).
\end{aligned}$$

Passing to the inf with respect to $u \in \mathcal{A}^R_{1,\#}(Q_{n\hat{k}})$, we get

$$f^{R,R}_{n\hat{k}+2}(M) \geq \overline{f}^R(M) - O\left(\frac{1}{n}\right),$$

and then, letting n tend to $+\infty$, we obtain (5.2). □

Remark 5.2 (quadratic forms). Under the hypotheses of Theorem 5.1, if, in addition, for any $\xi \in \mathbf{Z}^N$ $f^\xi(\alpha, \cdot)$ is a positive quadratic form on $\mathbf{R}^d$, that is,

$$f^\xi(\alpha, z) = \langle A^\xi(\alpha)z, z\rangle, \quad A^\xi(\alpha) \in \mathcal{M}^{d\times d}_{sym},$$

then, thanks to Remark 3.2, the limit energy density $f_{hom}(\cdot)$ is a homogeneous quadratic form on $\mathcal{M}^{d\times N}$, and formula (3.5) becomes

$$\begin{aligned} f_{hom}(M) &= A_{hom}(M, M) \\ &= \frac{1}{\hat{k}} \inf \left\{ \sum_{\xi\in\mathbf{Z}^N} \sum_{\beta\in I_{\mathbf{k}}} \left\langle A^\xi(\beta) \cdot \left(M\frac{\xi}{|\xi|} + D^\xi_1 v(\beta)\right), \left(M\frac{\xi}{|\xi|} + D^\xi_1 v(\beta)\right)\right\rangle, \right. \\ &\qquad \left. v \in \mathcal{A}_{1,\#}(\mathcal{R}_{\mathbf{k}}) \right\} \end{aligned}$$

with $A_{hom} \in T_2(\mathcal{M}^{d\times N})$.

If $N = d = 1$ and only nearest-neighbor interactions are taken into account, that is,

$$f^\xi \equiv 0 \text{ if } \xi \neq e_1, \qquad f^{e_1}(\alpha, z) = a(\alpha)z^2,$$

with $a : \mathbf{Z}^N \to (0, +\infty)$ k-periodic, the previous minimum problem can be easily solved (see [9]), giving the analogue in the discrete setting of a well-known homogenization result for integral functionals (see [6]). In fact, in this case

$$A_{hom} = \frac{1}{k}\left(\sum_{\beta=0}^{k-1} \frac{1}{a(\beta)}\right)^{-1}$$

is the harmonic mean of $a(\cdot)$.

Remark 5.3. Note that if $\mathcal{R}_{\mathbf{k}} = (0, 1)^N$, that is, f^ξ does not depend on the space variable α, in Theorem 5.1 we obtain

$$f_{hom}(M) = \sum_{\xi\in\mathbf{Z}^N} f^\xi\left(M\frac{\xi}{|\xi|}\right).$$

6. Interactions along independent directions and reduction to the one-dimensional case. In this section we first recall some results proven in the one-dimensional setting in [11], where a nonasymptotic formula defining the limit energy density f_{hom} is provided when only nearest and next-to-nearest neighbor interactions are considered.

Then in Theorem 6.3 we will show that if only interactions along the coordinate directions are taken into account, the N-dimensional problem can be reduced to a one-dimensional one.

The following two theorems have been proven in [11] in the case $d = 1$. Their proof in the case $d > 1$ is the same.

THEOREM 6.1 (nearest-neighbor interactions). *Let $\Omega = (0, l) \subset \mathbf{R}$, and let $F_\varepsilon : L^p(\Omega; \mathbf{R}^d) \to [0, +\infty)$ be defined as*

$$F_\varepsilon(u) := \begin{cases} \displaystyle\sum_{i=1}^{l-2} \varepsilon f\left(\frac{u(\varepsilon(i+1)) - u(\varepsilon i)}{\varepsilon}\right) & \text{if } u \in \mathcal{A}_\varepsilon(\Omega), \\ +\infty & \text{otherwise,} \end{cases}$$

with $f : \mathbf{R}^d \to [0, +\infty)$ satisfying $f(z) \geq C(|z|^p - 1)$. Then the conclusions of Theorem 4.1 hold with

$$f_{hom}(z) = f^{**}(z).$$

THEOREM 6.2 (next-to-nearest neighbor interactions). *Let $\Omega = (0, l) \subset \mathbf{R}$, and let $F_\varepsilon : L^p(\Omega; \mathbf{R}^d) \to [0, +\infty)$ be defined as*

$$F_\varepsilon(u) := \begin{cases} \displaystyle\sum_{i=1}^{l-2} \varepsilon f^1\left(\frac{u(\varepsilon(i+1)) - u(\varepsilon i)}{\varepsilon}\right) + \sum_{i=1}^{l-3} \varepsilon f^2\left(\frac{u(\varepsilon(i+2)) - u(\varepsilon i)}{2\varepsilon}\right) & \\ \qquad\qquad \text{if } u \in \mathcal{A}_\varepsilon(\Omega), & \\ +\infty \qquad\qquad \text{otherwise,} & \end{cases}$$

with $f^1, f^2 : \mathbf{R}^d \to [0, +\infty)$ satisfying $f^1(z) \geq C(|z|^p - 1)$. Then the conclusions of Theorem 4.1 hold with

$$f_{hom}(z) = \tilde{f}^{**}(z),$$

where $\tilde{f}(z) = f^2(z) + \frac{1}{2} \inf\{f^1(z_1) + f^1(z_2),\ z_1 + z_2 = 2z\}$. Back to the general N-dimensional setting, we consider now energies of the form

$$F_\varepsilon(u) = \begin{cases} \displaystyle\sum_{i=1}^{N} \mathcal{F}_\varepsilon^i(u, \Omega) & \text{if } u \in \mathcal{A}_\varepsilon(\Omega), \\ +\infty & \text{otherwise,} \end{cases} \tag{6.1}$$

where, for any $i \in \{1, \dots, N\}$, $\mathcal{F}_\varepsilon^i : \mathcal{A}_\varepsilon(\Omega) \times \mathcal{A}(\Omega) \to [0, +\infty]$ is defined as

$$\mathcal{F}_\varepsilon^i(u, A) := \sum_{k=1}^{+\infty} \sum_{\alpha \in R_\varepsilon^{ke_i}(A)} \varepsilon^N f_i^k(D_\varepsilon^{ke_i} u(\alpha)), \tag{6.2}$$

with $f_i^k : \mathbf{R}^d \to [0, +\infty)$ satisfying

$$f_i^1(z) \geq c(|z|^p - 1), \qquad f_i^k(z) \leq C_i^k(|z|^p + 1),$$

and

$$\sum_{i=1}^{N} \sum_{k=1}^{+\infty} C_i^k < +\infty.$$

This is a particular case of the model considered in section 4, with $f^\xi \equiv 0$ if $\xi \neq ke_i$, $f^{ke_i}(0,z) = f_i^k(z)$, $i \in \{1,\dots,N\}$, $k \in \mathbf{N}$, and $\mathcal{R}_{\mathbf{k}} = (0,1)^N$.

The following theorem shows that, in this case, the homogenization formula defining f_{hom} can be rewritten as a sum of N one-dimensional homogenization formulas.

THEOREM 6.3. *Let F_ε be defined by* (6.2). *Then the Γ-convergence result stated in Theorem* 4.1 *holds with f_{hom} satisfying*

$$f_{hom}(M) = \sum_{i=1}^{N} \tilde{f}_i(M^i) \tag{6.3}$$

for any $M = (M^1,\dots,M^N) \in \mathcal{M}^{d\times N}$, where $\tilde{f}_i : \mathbf{R}^d \to \mathbf{R}$, $i \in \{1,\dots,N\}$, is defined by the following one-dimensional homogenization formula:

$$\tilde{f}_i(z) := \lim_{h\to+\infty} \frac{1}{h} \inf \left\{ \sum_{k=1}^{+\infty} \sum_{j=1}^{h-k-1} f_i^k\left(\frac{v(j+k)-v(j)}{k}\right),\ v \in \mathcal{A}_{1,z}((0,h)) \right\}.$$

Proof. We first prove that

$$f_{hom}(M) \geq \sum_{i=1}^{N} \tilde{f}_i(M^i).$$

To do this, by the definition of $f_{hom}(M)$, it suffices to show that for any $i \in \{1,\dots,N\}$, $u \in \mathcal{A}_{1,M}(Q_h)$ we have

$$\frac{1}{h^N} \mathcal{F}_1^i(u,Q_h) \geq \tilde{f}_i(M^i) + O(h). \tag{6.4}$$

We use a slicing argument. For $i \in \{1,\dots,N\}$, set

$$m_h^i(z) := \frac{1}{h} \inf \left\{ \sum_{k=1}^{+\infty} \sum_{j=1}^{h-k-1} f_i^k\left(\frac{v(j+k)-v(j)}{k}\right),\ v \in \mathcal{A}_{1,z}((0,h)) \right\}.$$

By simplicity of notation, we prove (6.4) for $i = 1$. Given $u \in \mathcal{A}_{1,M}(Q_h)$, we may write

$$\mathcal{F}_1^1(u,Q_h) = \sum_{\beta\in\{1,\dots,h-1\}^{N-1}} \sum_{k=1}^{+\infty} \sum_{j=1}^{h-k-1} f_1^k\left(\frac{u(j+k,\beta)-u(j,\beta)}{k}\right). \tag{6.5}$$

Since for any $\beta \in \{1,\dots,h-1\}^{N-1}$ the function $v(j) := u(j,\beta) - \tilde{M}\beta$ belongs to $\mathcal{A}_{1,M^1}(0,h)$, where $\tilde{M} := \left(M^2,\dots,M^N\right)$, from (6.5) we get

$$\frac{1}{h^N} \mathcal{F}_1^1(u,Q_h) \geq \frac{1}{h^{N-1}} \#\left(\{1,\dots,h-1\}^{N-1}\right) m_h^1(M^1) \geq m_h^1(M^1).$$

We then easily infer inequality (6.4).

We now prove that

$$f_{hom}(M) \leq \sum_{i=1}^{N} \tilde{f}_i(M^i). \tag{6.6}$$

With fixed $\eta > 0$, for any $i \in \{1, \dots N\}$ let $v_h^i \in \mathcal{A}^2_{1,M^i}(0,h)$ be such that

$$\frac{1}{h}\sum_{k=1}^{+\infty}\sum_{j=1}^{h-k-1} f_i^k\left(\frac{v_h^i(j+k)-v_h^i(j)}{k}\right) \le m_h^i(M^i)+\eta, \tag{6.7}$$

and set

$$u_h(\alpha) := \sum_{i=1}^N v_h^i(\alpha_i), \quad \alpha = (\alpha_1, \dots, \alpha_N).$$

Note that $u_h \in M\alpha + \mathcal{A}_{1,\#}(Q_{h-2})$. Moreover, by the analogue of (6.5) applied to $\mathcal{F}_1^i(u,Q_h)$ for any $i \in \{1,\dots,N\}$ and by (6.7), we easily deduce that

$$\frac{1}{h^N}\sum_{i=1}^N \mathcal{F}_1(u_h, Q_h) \le \sum_{i=1}^N m_h^i(M^i) + N\eta.$$

Eventually, by the characterization of f_{hom} given by formula (4.12), letting first h tend to $+\infty$ and then η tend to 0, we get (6.6). □

Remark 6.4. Note that formula (6.3) highlights that a superposition principle holds, in the sense that the limit energy is obtained by relaxing the energies due to the interactions in every coordinate direction independently and then summing over them.

Remark 6.5. (a) (nearest-neighbors) by Theorem 6.1, if $f_i^k = 0$ for all $k \neq 1$, then formula (6.3) can be rewritten as

$$f_{hom}(M) = \sum_{i=1}^N (f_i^1)^{**}(M^i);$$

(b) (next-to-nearest neighbors) by Theorem 6.2, if $f_i^k = 0$ for all $k \neq 1, 2$, then formula (6.3) can be rewritten as

$$f_{hom}(M) = \sum_{i=1}^N (\tilde{f}_i)^{**}(M^i),$$

with

$$\tilde{f}_i(z) = f_i^2(z) + \frac{1}{2}\inf\{f_i^1(z_1) + f_i^1(z_2),\ z_1 + z_2 = 2z\}.$$

7. An example of quasi-convex nonconvex limit energy density. In the following we provide an example of vector-valued discrete interaction energies defined in the plane whose continuous counterpart has an energy density which is a quasi-convex (nonpolyconvex) function. Our example draws inspiration from Šverák's construction of a quasi-convex function which is not polyconvex (see [21]). Let $N = d = 2$, $p > 1$, and define $f_i : \mathbf{R}^2 \to [0, +\infty)$, $i = 1, 2, 3$, as

$$f_i(z) = \begin{cases} 1 + |z|^p & \text{if } z \neq \pm\frac{\xi_i}{|\xi_i|}, \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\xi_1 = e_1$, $\xi_2 = e_2$, $\xi_3 = e_1 + e_2$. Let F_ε be defined as

$$F_\varepsilon(u) = \sum_{i=1}^{3} \sum_{\alpha \in R_\varepsilon^{\xi_i}} \varepsilon^2 f_i\left(D_\varepsilon^{\xi_i} u(\alpha)\right);$$

then the conclusions of Theorems 4.1 and 4.5 and Corollary 4.6 hold with f_{hom} given by

$$f_{hom}(M) = \lim_{h \to +\infty} \frac{1}{h^N} \min \left\{ \sum_{i=1}^{3} \sum_{\beta \in R_1^{\xi_i}(Q_h)} f_i(D_1^{\xi_i} v(\beta)), \quad v \in \mathcal{A}_{1,M}\left(Q_h\right) \right\}.$$

THEOREM 7.1. *f_{hom} is not convex.*

Proof. By testing the minimum problem defining f_{hom} with the identity function and its opposite, we immediately obtain that

$$f_{hom}(I) = f_{hom}(-I) = 0,$$

where I is the identity matrix in $M^{2\times 2}$. The claim is proven if we show that $f_{hom}(0) > 0$. We argue by contradiction. Without loss of generality we may assume that Theorem 4.5 holds with $A = Q_1$. If $f_{hom}(0)$ were zero, there should exist a sequence $u_n \in \mathcal{A}_{\varepsilon_n,0}(Q_1)$ such that $u_n \to 0$ in $L^p(Q_1; \mathbf{R}^2)$ and

$$\lim_n F_{\varepsilon_n}(u_n) = 0. \tag{7.1}$$

Set

$$\begin{aligned} T^+ &:= \{(x_1, x_2) \in \mathbf{R}^2 : \ 0 \le x_1 \le 1, \quad x_1 \le x_2 \le 1\}, \\ T^- &:= \{(x_1, x_2) \in \mathbf{R}^2 : \ 0 \le x_1 \le 1, \quad 0 \le x_2 \le x_1\}, \end{aligned}$$

and consider the family of piecewise affine functions $v_n : Q_1 \to \mathbf{R}^2$ defined as follows:

$$v_n(x) = \begin{cases} u_n(\alpha) + D_{\varepsilon_n}^{e_1} u_n(\alpha)(x_1 - \alpha_1) \\ +D_{\varepsilon_n}^{e_2} u_n(\alpha + \varepsilon_n e_1)(x_2 - \alpha_2) & \text{if } x \in \alpha + \varepsilon_n T^-, \\ \\ u_n(\alpha) + D_{\varepsilon_n}^{e_1} u_n(\alpha + \varepsilon_n e_2)(x_1 - \alpha_1) \\ +D_{\varepsilon_n}^{e_2} u_n(\alpha)(x_2 - \alpha_2) & \text{if } x \in \alpha + \varepsilon_n T^+. \end{cases}$$

Note that $v_n|_{\partial Q_1} = 0$. Moreover, it is easy to check that

$$F_{\varepsilon_n}(u_n) = \int_{Q_1} \tilde{f}(\nabla v_n)\, dx, \tag{7.2}$$

where $\tilde{f} : M^{2\times 2} \to [0, +\infty)$ is defined as

$$\tilde{f}(\zeta) := f_1(\zeta_1) + f_2(\zeta_2) + f_3\left(\frac{\zeta_1 + \zeta_2}{\sqrt{2}}\right), \quad \zeta = (\zeta_1, \zeta_2) \in M^{2\times 2}.$$

In particular, by (7.1)

$$\lim_n \int_{Q_1} \tilde{f}(\nabla v_n)\, dx = 0. \tag{7.3}$$

Since we have

$$\tilde{f}(\zeta) \geq c(|\zeta_{11} - \zeta_{22}|^p + |\zeta_{12} + \zeta_{21}|^p),$$

by (7.1) and (7.2) we obtain

$$\lim_n \int_{Q_1} (|\nabla_1 v_n^1 - \nabla_2 v_n^2|^p + |\nabla_1 v_n^2 + \nabla_2 v_n^1|^p)\ dx = 0. \tag{7.4}$$

Since

$$\Delta v_n^1 = \operatorname{div}(\nabla_1 v_n^1 - \nabla_2 v_n^2, \nabla_1 v_n^2 + \nabla_2 v_n^1),$$

$$\Delta v_n^2 = \operatorname{div}(\nabla_1 v_n^2 + \nabla_2 v_n^1, -\nabla_1 v_n^1 + \nabla_2 v_n^2),$$

using the L^p estimates for the Laplace operator (see [19]) we obtain that

$$\begin{aligned}\|\nabla v_n^i\|^p_{L^p(Q_1;\mathbf{R}^2)} &\leq \|\Delta v_n^i\|^p_{W^{-1,p}(Q_1;\mathbf{R}^2)} \\ &\leq \int_{Q_1} (|\nabla_1 v_n^1 - \nabla_2 v_n^2|^p + |\nabla_1 v_n^2 + \nabla_2 v_n^1|^p)\ dx\end{aligned}$$

for $i = 1, 2$. Then, by (7.4) and the previous estimates, ∇v_n converges to 0 strongly in $L^p(Q_1; M^{2\times 2})$, so that

$$\lim_n \int_{Q_1} \tilde{f}(\nabla v_n)\ dx = \tilde{f}(0)\ |Q_1| > 0.$$

Hence we reach a contradiction. □

Remark 7.2. In the particular case $1 < p < 2$, thanks to the growth hypotheses on f_i, f_{hom} is a quasi-convex nonpolyconvex function (see [6, Remark 6.9]).

Acknowledgment. Our attention to this problem was drawn by Andrea Braides. We thank him for fruitful discussions and useful remarks.

REFERENCES

[1] R. Alicandro, M. Focardi, and M. S. Gelli, *Finite difference approximation of energies in fracture mechanics*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 29 (2000), pp. 671–709.
[2] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, Oxford University Press, Oxford, UK, 2000.
[3] A. Braides, *Non local variational limits of discrete systems*, Commun. Contemp. Math., 2 (2000), pp. 285–297.
[4] A. Braides, Γ-*Convergence for Beginners*, Oxford University Press, Oxford, UK, 2002.
[5] A. Braides, G. Dal Maso, and A. Garroni, *Variational formulation of softening phenomena in fracture mechanics: The one-dimensional case*, Arch. Ration. Mech. Anal., 146 (1999), pp. 23–58.
[6] A. Braides and A. Defranceschi, *Homogenization of Multiple Integrals*, Oxford University Press, Oxford, UK, 1998.
[7] A. Braides and G. Francfort, *Bounds on the effective behaviour of a square conducting lattice*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., to appear.
[8] A. Braides and M. S. Gelli, *Continuum limits of discrete systems without convexity hypotheses*, Math. Mech. Solids, 6 (2002), pp. 395–414.
[9] A. Braides and M. S. Gelli, *From Discrete to Continuum: A Variational Approach*, Lecture Notes SISSA, Trieste, 2000.
[10] A. Braides and M. S. Gelli, *Limits of discrete systems with long range interactions*, J. Convex Anal., 9 (2002), pp. 363–399.

[11] A. Braides, M. S. Gelli, and M. Sigalotti, *The passage from non-convex discrete systems to variational problem in Sobolev spaces: The one-dimensional case*, Proc. Steklov Inst. Math., 236 (2002), pp. 395–414.
[12] H. Brezis, *Analyse Fonctionelle*, Masson, Paris, 1983.
[13] G. Buttazzo, *Semicontinuity, Relaxation and Integral Representation in the Calculus of Variations*, Pitman, London, 1989.
[14] A. Chambolle, *Finite differences discretizations of the Mumford-Shah functional*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 261–288.
[15] G. Dal Maso, *An Introduction to* Γ*-Convergence*, Birkhäuser Boston, Boston, 1993.
[16] E. De Giorgi and T. Franzoni, *Su un tipo di convergenza variazionale*, Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur. (8), 58 (1975), pp. 842–850.
[17] E. De Giorgi and G. Letta, *Une notion générale de convergence faible pour des fonctions croissantes d'ensamble*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 4 (1977), pp. 61–99.
[18] G. Friesecke and F. Theil, *Validity and failure of the Cauchy-Born hypothesis in a* $2D$ *mass-spring lattice*, J. Nonlinear Sci., 12 (2002), pp. 445–478.
[19] C. B. Morrey, *Multiple Integrals of the Calculus of Variations*, Springer-Verlag, Berlin, 1966.
[20] A. Piatnitski and E. Remy, *Homogenization of elliptic difference operators*, SIAM J. Math. Anal., 33 (2001), pp. 53–83.
[21] V. Šverák, *Quasiconvex functions with subquadratic growth*, Proc. Roy. Soc. London Ser. A, 433 (1991), pp. 723–725.
[22] L. Truskinovsky, *Fracture as phase transition*, in Contemporary Research in the Mechanics and Mathematics of Materials, R. C. Batra and M. F. Beatty, eds., CIMNE, Barcelona, 1996, pp. 322–332.
[23] M. Vogelius, *A homogenization for planar polygonal networks*, RAIRO Modél. Math. Anal. Numér., 25 (1991), pp. 483–514.

SIAM J. MATH. ANAL.
Vol. 36, No. 1, pp. 38–68

TRAVELING WAVE SOLUTIONS OF FOURTH ORDER PDEs FOR IMAGE PROCESSING*

J. B. GREER† AND A. L. BERTOZZI‡

Abstract. The authors introduce two nonlinear advection-diffusion equations, each of which combines Burgers's convection with a fourth order nonlinear diffusion previously designed for image denoising. One equation uses the L^2-curvature diminishing diffusion of You and Kaveh [*IEEE Trans. Image Process.*, 9 (2000), pp. 1723–1730], and the other uses the "low curvature image simplifiers" diffusion of Tumblin and Turk [*Proceedings of the* 26*th Annual Conference on Computer Graphics*, ACM Press/Addison-Wesley, New York, 1999, pp. 83–90]. The new PDEs are compared with a third advection-diffusion equation that combines Burgers's convection with a second order diffusion recommended by Perona and Malik for denoising and edge detection [*IEEE Trans. Pattern Anal. Machine Intell.*, 12 (1990), pp. 629–639]. We prove results regarding the existence and nonexistence of traveling wave solutions of each PDE. Visualizations of each ODE's phase space show qualitative differences between the two fourth order problems. The combined work gives insight into the existence of finite time singularities in solutions of the diffusion equations.

Key words. fourth order diffusion, image denoising, traveling waves, edge detection, Conley index, advection diffusion, dynamical systems, nonlinear partial differential equations

AMS subject classifications. 35G25, 74J30, 68U10, 37B30, 37C29

DOI. 10.1137/S0036141003427373

1. Introduction.

We introduce two nonlinear advection-diffusion equations that each combine Burgers's convection with a fourth order nonlinear diffusion intended for image processing:

$$u_t + \left(\frac{1}{2}u^2\right)_x = -(g(u_{xx})u_{xx})_{xx} \tag{YK}$$

and

$$u_t + \left(\frac{1}{2}u^2\right)_x = -(g(u_{xx})u_{xxx})_x, \tag{TT}$$

with $g(s) = \frac{1}{1+s^2}$. Very little is known about the fourth order diffusions, despite recent demonstrations of their effectiveness for image denoising [43, 51]. The combined advection-diffusion equations have the possibility of smooth traveling wave solutions approximating Burgers's shocks. We prove rigorously that such smooth traveling wave solutions of (YK) do not exist for sufficiently large jumps, whereas smooth traveling wave solutions of (TT) exist for all jump values. These results suggest very different behavior of the fourth order nonlinear imaging equations introduced by You and Kaveh [51] and Tumblin and Turk [43].

*Received by the editors May 6, 2003; accepted for publication (in revised form) October 10, 2003; published electronically June 22, 2004. This work was supported by ONR grant N000140110290, NSF grants DMS-0074049 and DMS-0244498, and ARO grant DAAD19-02-1-0055. NSF grant DMS-0345602 provided support under the Approaches to Combat Terrorism Program, jointly funded by the Intelligence Community and the NSF Directorate for Mathematical and Physical Sciences.

http://www.siam.org/journals/sima/36-1/42737.html

†Department of Mathematics, Duke University, Durham, NC 27708 (jbg33@math.duke.edu) and Department of Mathematics, UCLA, Los Angeles, CA 90095.

‡Department of Mathematics and Physics, Duke University, Durham, NC 27708 (bertozzi@math.duke.edu) and Department of Mathematics, UCLA, Los Angeles, CA 90095.

1.1. Nonlinear PDEs for image denoising. Nonlinear PDEs are now commonly used in image processing for issues ranging from edge detection, denoising, and image inpainting to texture decomposition. Before the development of nonlinear PDE-based methods, the problem of noise reduction in images was treated through linear filtering, in which the image intensity function is convolved with a Gaussian. The method of linear filtering was introduced by Marr and Hildreth [34] and then further developed by Witkin [50], Koenderink [28], and Canny [15]. It is equivalent to solving the heat equation with initial data given by the noisy image intensity function. Although this technique quickly damps out any noise in the image, it also badly blurs edges, often leaving objects in the image unrecognizable.

Nonlinear second order PDEs were introduced with the intention of smoothing while preserving edges. Examples of second order nonlinear PDEs for image processing date back to the seminal works of Perona and Malik [36] and Rudin, Osher, and Fatemi [38]. Their methods are based on a nonlinear version of the heat equation,

$$u_t = \nabla \cdot ((g(|\nabla u|)\nabla u), \tag{1.1}$$

in which the "thresholding function" g is small in regions of sharp gradients. A number of mathematical issues arise with these equations and their use. For example, Perona and Malik suggest using a smooth, positive, and even function g that decays fast enough for large ∇u so that significant diffusion takes place only in regions away from image edges. Specifically, Perona and Malik required the existence of some $K > 0$ such that

$$\frac{d}{ds}\left(g(s)s\right) > 0 \text{ for } 0 < s < K \tag{1.2}$$

and

$$\frac{d}{ds}\left(g(s)s\right) < 0 \text{ for } s > K. \tag{1.3}$$

However, the nonmonotonicity of $g(s)s$ causes (1.1) to be ill-posed in regions of high gradients, and the ensuing dynamics result in a characteristic "staircase" instability. Following [1] and [26], the cause of this ill-posedness can be seen by rewriting the Laplacian locally in terms of $\nu = \frac{\nabla u}{|\nabla u|}$ and a direction η perpendicular to ν. Letting $F(s) = g(s)s$, (1.1) can be rewritten as

$$u_t = F'(|\nabla u|)u_{\nu\nu} + g(|\nabla u|)u_{\eta\eta}. \tag{1.4}$$

Requirement (1.3) then implies that in regions where $|\nabla u| > K$, (1.4) (and therefore (1.1)) is backwards parabolic in the direction of the gradient.

A typical thresholding function g is

$$g(s) = \frac{1}{1 + \left(\frac{s}{k}\right)^2}, \tag{1.5}$$

where k is a parameter used to establish a standard edge size for the image [21, 27, 45, 46]. Figure 1 shows (1.5) for $k = 1$. We note that degenerate parabolic equations which have structure similar to those of Perona and Malik, and which exhibit the same "staircasing" effect, arise in simplified models for the velocity field of a sheared granular medium [49].

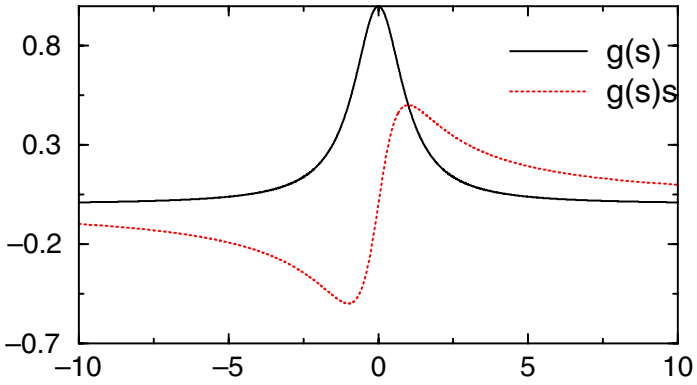

FIG. 1. *An example thresholding function.* $g(s)$ *and* $g(s)s$ *are shown for* $g(s) = \frac{1}{1+s^2}$.

In the past few years, a number of authors have proposed analogous fourth order PDEs for edge detection and image denoising with the hope that these methods would perform better than their second order analogues [17, 18, 32, 33, 43, 44, 51]. Indeed there are good reasons to consider fourth order equations. First, fourth order linear diffusion damps oscillations at high frequencies (i.e., noise) much faster than second order diffusion. Second, there is the possibility of having schemes that include effects of curvature (i.e., the second derivatives of the image) in the dynamics, thus creating a richer set of functional behaviors. On the other hand, the theory of fourth order nonlinear PDEs is far less developed than that of their second order analogues. Also, such equations often do not possess a maximum principle or comparison principle, and implementation of the equations could thus introduce artificial singularities or other undesirable behavior.

Some examples of fourth order equations include the L^2-*curvature* gradient flow method of You and Kaveh [51],

$$u_t = -\Delta(g(\Delta u)\Delta u), \tag{1.6}$$

the Perona–Malik analogue by Wei [44],

$$u_t = -\nabla \cdot (g(|\nabla u|)\nabla \Delta u), \tag{1.7}$$

and Tumblin and Turk's "low curvature image simplifiers" [43],

$$u_t + \nabla \cdot (g(D_{ij}u)\nabla \Delta u) = 0. \tag{1.8}$$

In (1.8), g is a function of the second derivatives of the image intensity function u. Although application of these PDEs to images as demonstrated in [43], [44], and [51] give similar results, it is unclear how the dynamics of the equations compare to each other and to the more established second order methods. Rigorous analysis is thus needed to better understand the new PDEs. One immediate observation is that (1.6) is linearly ill-posed in regions of high curvature, while (1.8) is not. Further insight into the two equations is gained by again defining $F(s) = g(s)s$ and noticing that (1.6) can be rewritten as

$$u_t = -F'(\Delta u)\,\Delta^2 u - F''(\Delta u)\,|\nabla \Delta u|^2, \tag{1.9}$$

while picking $D_{ij}u = \Delta u$ allows us to rewrite (1.8) as

$$u_t = -g\left(\Delta u\right)\Delta^2 u - g'\left(\Delta u\right)\left|\nabla\Delta u\right|^2. \tag{1.10}$$

We see that (1.9), like (1.1), is ill-posed in regions where $F'(s)$ is negative. On the other hand, (1.10) is always linearly well-posed. Also note that (1.9) becomes unstable in all directions when $F'(s)$ changes sign, whereas (1.4) has the instability only in the direction of ∇u.

A class of equations including (1.7) and (1.8) was studied in [24] by the authors, who proved global existence of H^1 solutions when the argument of g, in the form of derivatives of the intensity u, is convolved with a standard mollifier kernel. However, as is well known for some second order equations, as in (1.1), such mollification can turn an ill-posed problem into a well-posed problem [16]. The resulting numerical methods for the equations with mollification appear to smooth out, but not remove, undesirable artifacts of the method without mollification, such as the staircase instability of the Perona–Malik method.

1.2. The model equations. We introduce two model problems designed for studying the dynamics of these new image processing equations without mollification. Both are convection-diffusion equations which can be studied by a combination of analytical and computational methods. We introduce a Burgers convection into the dynamics of the fourth order diffusions (1.6) and (1.8) in order to instigate shock or jump-type behavior typical of edges in images. Such convective motion has real application in image processing. One area in particular is *image inpainting* [2, 3], for which image information is convectively flowed into a region where the image content is unknown. Thus our study gives insight into the behavior of hybrid imaging methods that combine diffusion and convection.

The two fourth order equations are compared with a second order convection diffusion equation that was introduced in [23] and [29]. This equation combines a Burgers convection term with the second order diffusion of (1.1). The authors of [23] and [29] share our motivation of using these equations as tools for understanding the diffusion dynamics.

The three model equations that we consider are

$$u_t + \left(\frac{1}{2}u^2\right)_x = (g(u_x)u_x)_x, \tag{PM}$$

$$u_t + \left(\frac{1}{2}u^2\right)_x = -(g(u_{xx})u_{xx})_{xx}, \tag{YK}$$

and

$$u_t + \left(\frac{1}{2}u^2\right)_x = -(g(u_{xx})u_{xxx})_x. \tag{TT}$$

In each equation, we use the thresholding function

$$g(s) = \frac{1}{1+s^2}, \tag{1.11}$$

as in [36]. Many of our results can be easily generalized to thresholding functions g which satisfy the properties stated in [36]. Remarks are made regarding possible generalizations of our results.

The idea of creating a simplified model PDE in order to understand more complex dynamics is an approach in applied analysis that has met with tremendous success in recent decades. A few examples include applications in combustion [12], singularities [25], aggregation in bacterial colonies [13], surface tension driven interfaces [4, 6, 11], shockwaves [19], vortex dynamics [20], and solidification [39]. In imaging, a very relevant problem is the interaction of higher order diffusion with jump discontinuities. Thus it is very natural to consider a model problem combining Burgers's equation, which produces shocks, with higher order diffusion.

We are interested in one overarching question for all three problems: When do the equations have smooth solutions, and when do they develop singularities (jumps in u or its derivatives)? This fundamental question arises when using such methods for image processing. Moreover, if a singularity forms, it is unclear whether a solution to the equation will continue to exist, perhaps as a weak or distribution solution, as is the case with shock dynamics.

We focus on a special class of similarity solutions—*traveling waves* of the form $u(x-ct)$. This traveling wave ansatz reduces the fourth order PDEs (YK) and (TT) to third order ODEs, to which we apply phase plane analysis from dynamical systems theory, as well as rigorous analysis using Conley index theory and estimates involving Lyapunov functions. Analyzing the simpler Perona–Malik equation (PM) is much more straightforward; however, it gives some insight and provides a standard for comparison with the more complicated fourth order equations.

Our approach in this paper has been successfully used for other fourth order nonlinear equations that model physical systems. A mathematically similar family of PDEs are the lubrication equations used to model thin liquid films under the influence of surface tension. These equations take the form

$$u_t + \nabla \cdot (m(u)\nabla\Delta u) = 0,$$

where $m(u)$ is typically degenerate (i.e., f vanishes when u vanishes). Convection in thin films can arise due to body forces such as gravity or surface stresses involving gradients of surface tension. Recent analysis of traveling waves for the PDE

$$u_t + (f(u))_x = -(u^3 u_{xxx})_x$$

has led to an understanding of compressive and undercompressive shock dynamics in driven films [8, 10, 9, 14]. Similar work has also been done to study the convective Cahn–Hilliard equation [47, 48]. We consider some of the analytical methods for these problems in our study of traveling waves for image processing.

1.3. Organization. We derive traveling wave ODEs for all three PDEs in section 2. By restricting ourselves to traveling wave solutions, the problems simplify to nonlinear ODEs. Sections 3–5 each contain an analysis of one of the three traveling wave ODEs. We first consider the simpler problem (PMODE) in section 3 and use it as a standard for comparing (YKODE), discussed in section 4, and (TTODE), considered in section 5. The three sections share the same outline. We first prove analytic results for the considered ODE. These results are then illustrated with phase plane visualizations which also provide strong evidence for ODE properties that are not proved here. We close each section with a numerical demonstration of the PDE's behavior and its relationship with the corresponding ODE.

2. Traveling wave solutions to PDEs. Traveling waves are similarity solutions of the form

$$u(x,t) = \phi(x - ct), \tag{2.1}$$

where $c \in \mathbb{R}$ is the wave speed. By substituting (2.1) into the PDE, we reduce the problem to an ODE in the variable $\xi = x - ct$. ODEs are typically easier to study, as there are many well-understood analytical and numerical methods for examining their qualitative behavior.

In this paper we consider traveling wave solutions that satisfy

$$\lim_{\xi \to -\infty} \phi(\xi) = u_L \quad \text{and} \quad \lim_{\xi \to +\infty} \phi(\xi) = u_R. \tag{2.2}$$

Such solutions correspond to trajectories connecting $\phi = u_L$ to $\phi = u_R$ in the phase space of the traveling wave ODE. They give diffusive shocks, similar to those for the viscous Burgers equation [31]. The values of u_L and u_R determine the viscous shock's wave speed, c.

2.1. ODEs resulting from (PM), (YK), and (TT). Assume

$$u(x,t) = \phi(x - ct) = \phi(\xi) \tag{2.3}$$

for some real number c to be determined. Using the notation $\phi' := \frac{d}{d\xi}\phi$ and substituting (2.3) into (PM), (YK), and (TT), we derive the ODEs

$$\phi'(\phi - c) = (g(\phi')\phi')', \tag{2.4}$$

$$\phi'(\phi - c) = -(g(\phi'')\phi'')'', \tag{2.5}$$

and

$$\phi'(\phi - c) = -(g(\phi'')\phi''')', \tag{2.6}$$

respectively. Assuming (2.2) and that all of the derivatives of ϕ decay at infinity, integrating each ODE yields

$$r(\phi) = g(\phi')\phi', \tag{PMODE}$$

$$r(\phi) = -(g(\phi'')\phi'')', \tag{YKODE}$$

and

$$r(\phi) = -g(\phi'')\phi''', \tag{TTODE}$$

where

$$r(\phi) := \frac{1}{2}\phi^2 - c\phi + \frac{1}{2}u_L u_R, \tag{2.7}$$

with wave speed

$$c = \frac{1}{2}(u_L + u_R). \tag{2.8}$$

For reference, we call (PMODE) the Perona–Malik ODE, (YKODE) the You–Kaveh ODE, and (TTODE) the Tumblin–Turk ODE. Each ODE has two equilibrium points: L, where $\phi = u_L$, and R, where $\phi = u_R$. A trajectory of one of the given ODEs is a traveling wave solution of the respective PDE if and only if that trajectory is a heteroclinic orbit connecting L and R. Each equation also has an entropy condition (which we derive) requiring $u_L > u_R$ for such an orbit to exist. This entropy condition is analogous to that of the viscous Burgers equation [31].

2.2. Reducing the number of parameters. Consider (PMODE), for a given pair u_L and u_R, and corresponding wave speed, $c = \frac{1}{2}(u_L + u_R)$. Letting $\Phi = \phi - c$, (PMODE) becomes

$$\frac{1}{2}\left(\Phi^2 - \frac{1}{4}(u_R - u_L)^2\right) = g(\Phi')\Phi'. \tag{2.9}$$

The dynamics of (PMODE) and (2.9) are affected solely by the difference between u_L and u_R. Changing their average, which gives the wave speed c, alters ϕ by only an added constant. The same holds true for (YKODE) and (TTODE).

For simplicity, we consider only the case $c = 0$, and we do so without loss of generality. All of our computational examples are done with $\phi(0) = c = 0$. These ODE solutions correspond to PDE solutions that travel with zero speed. We study the full range of behavior of the traveling wave ODEs by adjusting only one parameter, $\gamma := u_L > 0$. Insisting $c = 0$ forces $u_R = -\gamma$. With these conditions $r(\phi) = \frac{1}{2}\left(\phi^2 - \gamma^2\right)$. For both fourth order equations, $L = (0, 0, \gamma)$ and $R = (0, 0, -\gamma)$. For (PMODE), L corresponds to $\phi = \gamma$, and R corresponds to $\phi = -\gamma$.

2.3. Comparing the traveling wave ODEs. In [29], Kurganov, Levy, and Rosenau proved the existence of traveling wave solutions of (PM) for the case $g(s) = \frac{1}{1+s^2}$. Traveling wave solutions exist for only a small range of left and right states. In particular, if u_L is much larger than u_R, the ODE will not have a solution connecting L to R. We generalize the results of [23] and [30] in section 3, which contains a proof of the existence of solutions of (PMODE) for the general class of functions g satisfying the properties listed by Perona and Malik. By studying (PM) and (PMODE), we develop a framework for analyzing the higher order equations. In section 3.4, we compare solutions of (PMODE) with the PDE (PM). Numerical experiments show a one-to-one correspondence between heteroclinic orbits of the ODE and attracting steady state solutions of the PDE. When there is no trajectory connecting L to R in the ODE, a jump discontinuity forms in the PDE. We show that this restriction of left and right states stems from a singularity in the ODE which is caused by the lack of monotonicity of $g(s)s$. The same dilemma also occurs in (YKODE), and we establish results in section 3 that parallel the higher order problem.

The higher order diffusion makes analytical results more difficult to obtain for (YKODE) and (TTODE). However, in section 4 we prove that (YKODE) does not have a smooth solution connecting L and R for large γ. By studying the ODE phase plane with the method introduced by [8], we discover that the unstable manifold of the left state intersects the stable manifold of the right state only when γ is small enough—just as in the second order case. We conclude the section by comparing the ODE solutions with the PDE (YK).

The Tumblin–Turk ODE is remarkably different from the other two ODEs. In section 5, we use a topological argument to prove that (TTODE) has smooth solutions connecting L and R for all $\gamma > 0$. Cross-sections of its phase plane illustrate the key differences between the phase plane geometries of (YKODE) and (TTODE). Once again, we follow the discussion with numerical computations of the PDE.

3. Perona–Malik with advection. Equation (PM) is carefully studied in [23] and [29]. We review and expand upon those results here, as they provide an excellent foundation for our analysis of (YK) and (TT). We first prove that (PMODE) has an orbit corresponding to a traveling wave solution of (PM) only when $\gamma > 0$ is smaller than a critical value, γ_c. This result is followed with a numerical and asymptotic description of solutions of (PMODE) for $\gamma > \gamma_c$.

3.1. The traveling wave ODE. We consider a general thresholding function g, as described in the introduction. Define

$$F(s) = g(s)s \tag{3.1}$$

so that (PMODE) can be written as

$$r(\phi) = F(\phi').$$

Since $g(s)s$ is bounded, we can only define F^{-1} on a subset of $\mathbb{R}$. F^{-1} has three branches that depend on the unique K satisfying

$$g'(K)K + g(K) = 0. \tag{3.2}$$

Two of these branches correspond to the regions $|s| > K$, where $\frac{d}{ds}(g(s)s) < 0$. The third is an interior branch with its range centered around zero and corresponds to the interval $|s| < K$, where $\frac{d}{ds}(g(s)s) > 0$. We define F^{-1} on the interior branch, since our traveling waves have $\phi' \to 0$ as $\xi \to \pm\infty$. With this definition, we rewrite (PMODE) as

$$\phi' = F^{-1}(r(\phi)),$$

with the requirement

$$|r(\phi)| \le F(K) = g(K)K. \tag{3.3}$$

This condition is satisfied if and only if

$$0 \le \gamma \le \sqrt{2g(K)K} \tag{3.4}$$

and is essential to proving the following theorem, which is proved in [29] for the specific case $g(s) = \frac{1}{1+s^2}$.

THEOREM 3.1. *Let g be a smooth, positive, and nonincreasing function of $|s|$, with some $K > 0$ satisfying*

$$\frac{d}{ds}(g(s)s) > 0 \text{ for } |s| < K \quad \text{and} \quad \frac{d}{ds}(g(s)s) < 0 \text{ for } |s| > K.$$

Then the ODE (PMODE) has a continuous solution $\phi(\xi)$ satisfying

$$\lim_{x\to-\infty} \phi(\xi) = \gamma \quad \text{and} \quad \lim_{x\to+\infty} \phi(\xi) = -\gamma \tag{3.5}$$

if and only if

$$0 \le \gamma \le \sqrt{2g(K)K}. \tag{3.6}$$

Proof. Any traveling wave solution of (PM) satisfying (2.2) corresponds to a trajectory of (PMODE) connecting L, the point $\phi = \gamma$, to R, the point $\phi = -\gamma$. Such a trajectory can only exist when $\gamma > 0$, since $F^{-1}(r(\phi)) < 0$ for $|\phi| < |\gamma|$. This is analogous to the Lax–Oleinik entropy condition for Burgers's equation [31]. If $\gamma \le 1$, $r(\phi) \le \sqrt{2g(K)K}$ for all $\phi \in (-\gamma, \gamma)$, so the existence of an orbit connecting L to R is obvious.

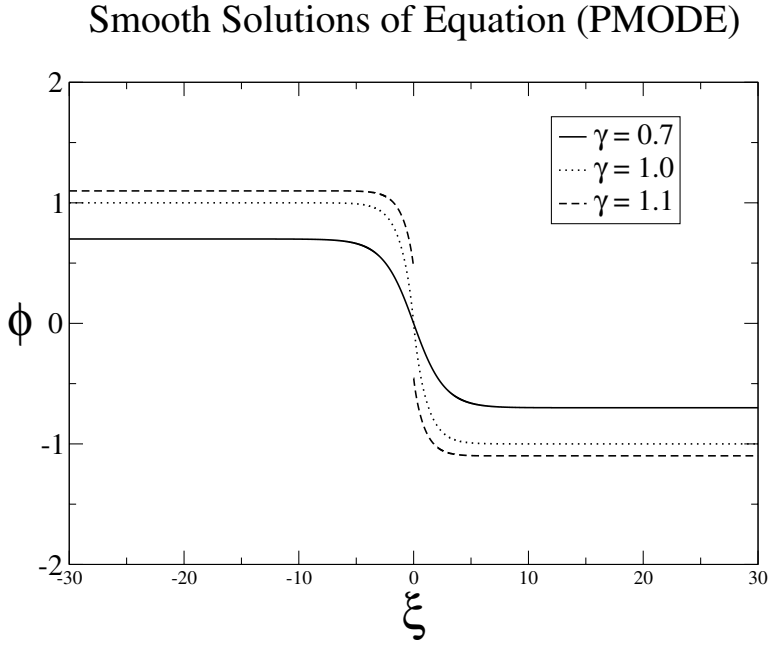

FIG. 2. *Heteroclinic orbits of* (PMODE) *for different values of* γ. *Solutions connecting* L *to* R *exist only for* $\gamma \leq 1$. *For* $\gamma = 1.1$, *we show two trajectories—one starting near* L *and one approaching* R.

Suppose $\gamma > \sqrt{2g(K)K}$. Any continuous heteroclinic orbit, ϕ, connecting L to R must have $\phi(\xi_0) = 0$ for some ξ_0. We calculate $|r(0)| = \frac{1}{2}\gamma^2 > g(K)K$ and remember that $g(s)s \leq g(K)K$ for all s, implying that ϕ cannot possibly satisfy (PMODE). ∎

Remark. For the remainder of the paper, we restrict the main part of our discussion to $g(s) = \frac{1}{1+s^2}$, for which $K = 1$, and $|g(s)s| \leq \frac{1}{2}$. Comments regarding generalizing our results to other thresholding functions will be made throughout the paper.

Figure 2 shows solutions of (PMODE) for $g(s) = \frac{1}{1+s^2}$ and various values of γ. Equation (PMODE) has a trajectory connecting L to R only when $\gamma \leq 1$. When $\gamma > 1$, (PMODE) has only a solution near the equilibrium points. Starting with ϕ slightly smaller than γ, we integrate forward in time until $|r(\phi)| = \frac{1}{2} = \max\{g(s)s\}$. We then start with ϕ slightly larger than $-\gamma$ and integrate backward in time until $|r(\phi)| = \frac{1}{2}$. Figure 2 shows $\phi(\xi)$ for $\gamma = 1.1$ in the regions of ξ, where $F^{-1}(r(\phi(\xi)))$ is defined.

3.2. Second order version of (PMODE). Expanding the right side of (2.4) yields a second order form of the traveling wave ODE for (PM):

$$\phi' = (g'(\phi')\phi' + g(\phi'))\,\phi''. \tag{3.7}$$

Unlike (PMODE), (3.7) does not depend on the choice of γ. Due to the properties of g, (3.7) becomes singular as $|\phi'| \to 1$. We rewrite (3.7) as a system of two ODEs:

$$\phi' = v, \quad v' = \frac{\phi v}{g'(v)v + g(v)}. \tag{3.8}$$

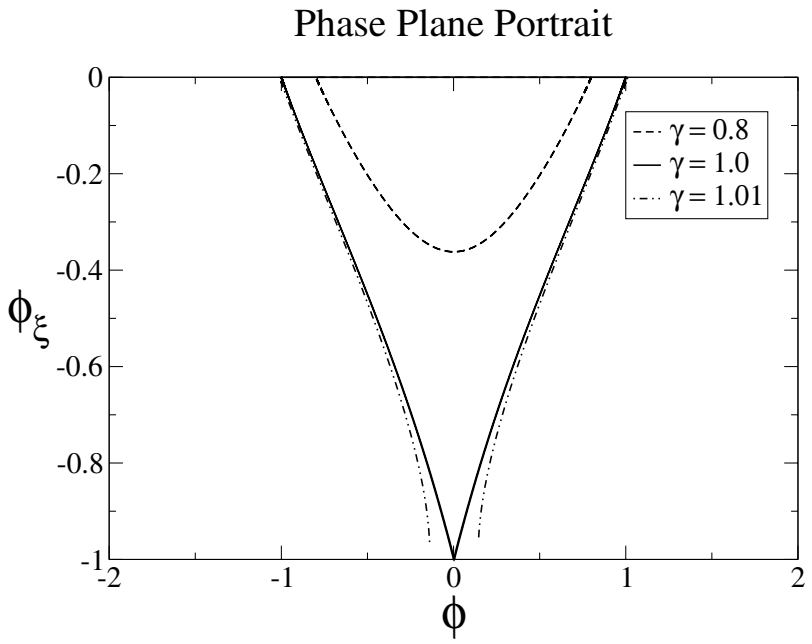

FIG. 3. *Phase plane of ODE system* (3.8). *A series of trajectories are plotted for different values of* γ. *If* $\gamma > 1$, *any connection from* $(\gamma, 0)$ *to* $(-\gamma, 0)$ *would need to pass through the line of singularity,* $\phi' = -1$.

System (3.8) has a line of equilibrium points at $v = 0$. Figure 3 shows integral curves where $\phi \to -\gamma$ as $\xi \to \infty$ and $\phi \to \gamma$ as $\xi \to -\infty$. Each integral curve coincides with a particular value of γ. As γ increases, the integral curves move toward the singular line $v = -1$, clearly illustrating the results of section 3 and showing why heteroclinic orbits of (PMODE) do not exist for large γ. Such traveling waves would require ϕ' to pass through the singular value $\phi' = -1$.

3.3. Singularities in solutions of (PMODE). We now consider the behavior of singular solutions of (PMODE). We examine two cases: $\gamma > 1$ and $\gamma = 1$. When $\gamma > 1$, there is no traveling wave solution. We consider a trajectory $\phi(\xi)$ starting near L and moving toward R and examine ξ_0 satisfying

$$\lim_{\xi \to \xi_0^-} \phi'(\xi) = -1 \text{ and } \lim_{\xi \to \xi_0^-} \phi(\xi) = \phi^*$$

for some $\phi^* > 0$. We have $\phi'\phi \to -\phi^*$ as $\xi \to \xi_0$. Near $\phi' = -1$,

$$g'(\phi')\phi' + g(\phi') \sim \frac{1}{2}(\phi' + 1),$$

so

$$\int_\xi^{\xi_0} -\phi^* \sim \int_\xi^{\xi_0} (\phi' + 1)\phi'',$$

and

$$\phi'(\xi) \sim \sqrt{4\phi^*(\xi_0 - \xi)} - 1. \tag{3.9}$$

When $\gamma = 1$, there is a nonsmooth traveling wave solution. In this case, $\phi^* = 0$ and $\phi(\xi) \sim -(\xi_0 - \xi)$ near $\xi = \xi_0$, so

$$\phi'(\xi) \sim \sqrt{2}|\xi - \xi_0| - 1. \tag{3.10}$$

This singular behavior is demonstrated by the solid line trajectory in Figure 3.

3.4. Second order PDE computations. We test the stability of each traveling wave solution found from (PMODE) by choosing an initial condition near the traveling wave and numerically integrating the PDE (PM). We use centered differences in space and backward Euler in time with an adaptive time step. The Burgers term is computed with a centered difference in flux form. We use Newton's method to approximate solutions of the nonlinear system, and the time step is adjusted to expedite convergence of Newton's method. If convergence requires more than three iterations, the time step is decreased by 10%.

Figure 4 shows computations for $\gamma = 1$ and $\gamma = 1.1$. When $0 \le \gamma \le 1$ (PMODE) has a heteroclinic orbit between L and R. The case $\gamma = 1$ is discussed in section 3.3. This traveling wave, ϕ, is continuous but nonsmooth. ϕ' behaves like (3.10) near $\phi = 0$. Given an initial condition near this traveling wave, the PDE solution converges to the traveling wave solution, as long as the gradient of the initial condition is not too large (for large gradients, (PM) becomes ill-posed, and a jump discontinuity occurs). There is no traveling wave solution for $\gamma > 1$, as seen in the computations for $\gamma = 1.1$; although the initial condition is smooth with small gradient, a discontinuity develops in finite time, and the long time solution has a jump discontinuity.

4. You–Kaveh with advection. Equation (YK) shares many of the properties of (PM). We prove that orbits of (YKODE) corresponding to traveling wave solutions of (YK) do not exist when γ is too large. This nonexistence follows from a singularity in (YKODE) that is analogous to that of (PMODE). We study the phase space of (YKODE) for evidence of the existence of traveling wave solutions when γ is small. For simplicity, we assume $g(s) = \frac{1}{1+s^2}$, which is the thresholding function chosen by You and Kaveh in [51]. However, our results generalize to other thresholding functions as described in section 1.1.

4.1. The traveling wave ODE. Equation (YKODE) can be expanded to

$$r(\phi) = -\left(g'(\phi'')\phi'' + g(\phi'')\right)\phi'''. \tag{4.1}$$

Since $g'(s)s + g(s) = 0$ for $s = \pm 1$, we immediately see a similarity to (PMODE): a solution ϕ of (YKODE) becomes singular in ϕ''' when $|\phi''| \to \pm 1$, just as a solution ϕ of (PMODE) becomes singular in ϕ'' when $|\phi'| \to \pm 1$.

Remark. For general functions g as described in [36], there exists a $K > 0$ satisfying (3.2), so (4.1) is singular at $\phi'' = \pm K$. A solution ϕ of (PMODE) becomes singular in ϕ'' when $|\phi'| \to K$, and a solution ϕ of (YKODE) becomes singular in ϕ''' when $|\phi''| \to K$.

4.2. Lyapunov function for the You–Kaveh ODE. Equation (YKODE) has a Lyapunov function. Multiplying (YKODE) by ϕ' and integrating, we have

$$\int_{-\infty}^{\xi} r(\phi(y))\phi'(y)dy + g(\phi''(\xi))\phi'(\xi)\phi''(\xi) = \int_{-\infty}^{\xi} g(\phi''(y))(\phi''(y))^2 dy. \tag{4.2}$$

Define

$$\mathcal{R}(s) = \int^{s} r(\alpha)d\alpha.$$

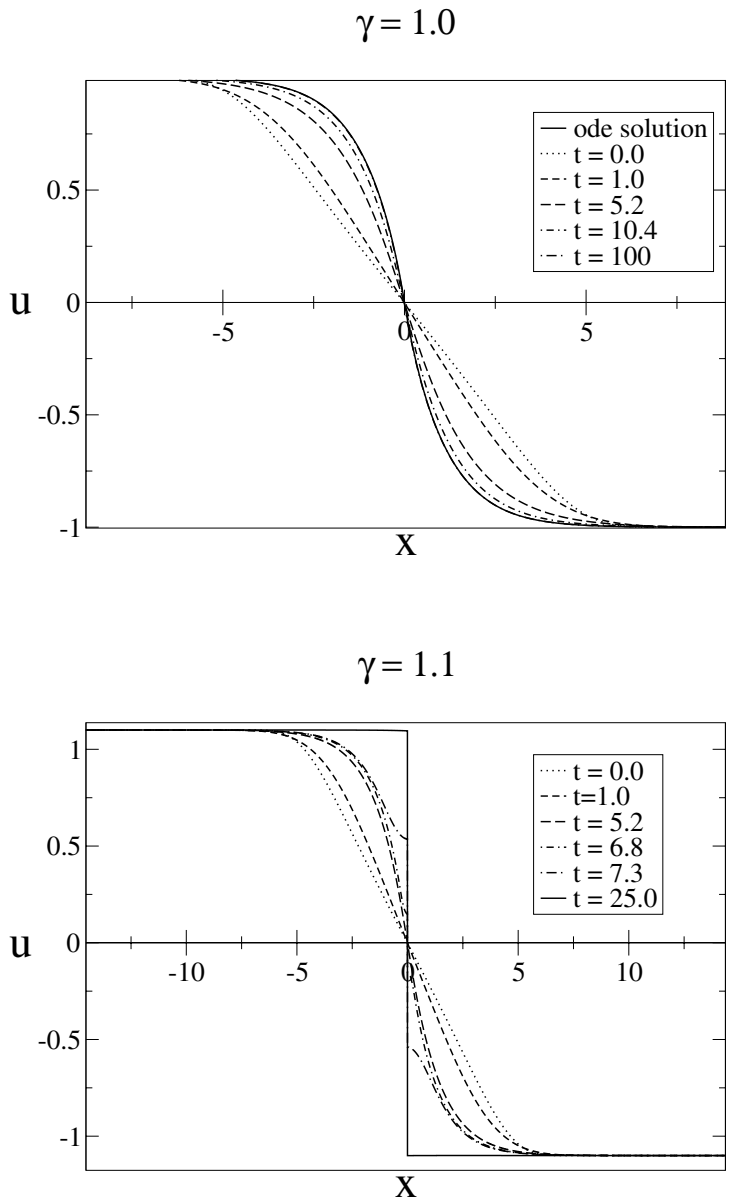

FIG. 4. *PDE* (PM) *solution,* u*, for* $\gamma = 1.0$ *and* $\gamma = 1.1$*. When* $\gamma = 1.0$*,* u *approaches the corresponding traveling wave ODE solution as* t *increases.* $\gamma = 1.0$ *is the maximum value for which the PDE has a traveling wave connecting* γ *to* $-\gamma$*. When* $\gamma = 1.1$*,* u *forms a jump discontinuity in finite time.*

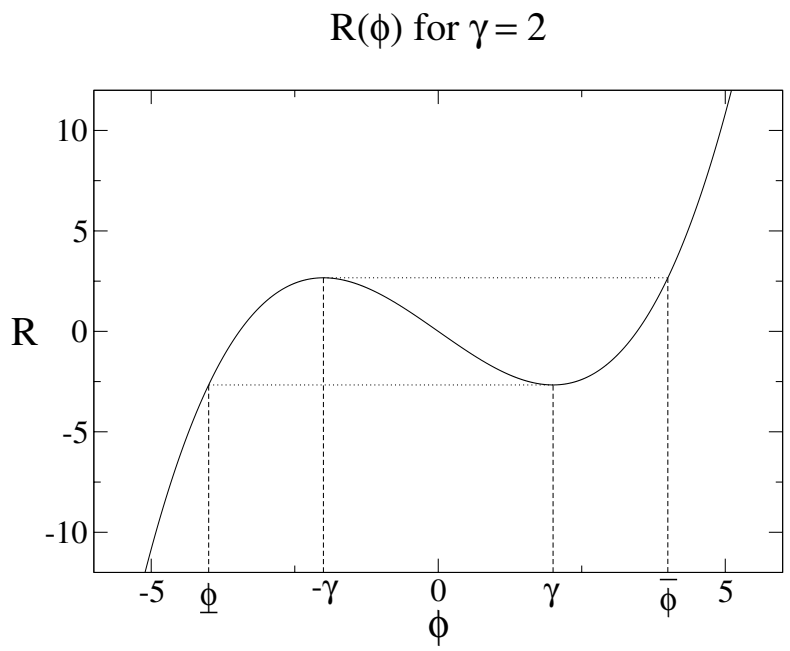

FIG. 5. $\mathcal{R}(\phi)$ *for* $\gamma = 2$. *We mark the maximum and minimum values,* $\overline{\phi}$ *and* $\underline{\phi}$, *for a bounded solution of* (YKODE).

We see that

$$\mathcal{L}_1(\xi) = \mathcal{R}(\phi(\xi)) + g(\phi''(\xi))\phi'(\xi)\phi''(\xi) \tag{4.3}$$

is nondecreasing, since

$$\frac{d}{d\xi}\mathcal{L}_1(\xi) = g(\phi''(\xi))(\phi''(\xi))^2 \geq 0. \tag{4.4}$$

Since $\mathcal{L}_1(\xi) = \mathcal{R}(\phi(\xi))$ at extrema of ϕ, the structure of $\mathcal{R}(\phi) = \frac{1}{6}\phi^3 - \frac{1}{2}\gamma^2\phi$ has a tremendous effect on solutions of (YKODE). Figure 5 shows $\mathcal{R}$ for a particular γ. The structure of $\mathcal{R}$ implies the entropy condition $\gamma > 0$. If $\gamma < 0$, then $\mathcal{R}(\gamma) > \mathcal{R}(-\gamma)$, so there could not be a heteroclinic orbit traveling from $L = (0, 0, \gamma)$ to $R = (0, 0, -\gamma)$. $\mathcal{R}$'s essential behavior remains the same for different values of γ. Assuming $\gamma > 0$, $\mathcal{R}$ is a cubic polynomial with a local maximum at $-\gamma$ and a local minimum at γ. $\mathcal{R}(\phi)$ strictly increases for $\phi < -\gamma$ and for $\phi > \gamma$, while it strictly decreases for $-\gamma < \phi < \gamma$.

Let

$$\overline{\phi} = 2\gamma \quad \text{and} \quad \underline{\phi} = -2\gamma. \tag{4.5}$$

A simple calculation shows

$$\mathcal{R}(\overline{\phi}) = \mathcal{R}(-\gamma) \quad \text{and} \quad \mathcal{R}(\underline{\phi}) = \mathcal{R}(\gamma). \tag{4.6}$$

The following lemmas are essential for proving that (YKODE) does not have a smooth heteroclinic orbit connecting L and R when γ is large. Lemma 4.1 is merely a tool for proving Lemma 4.2.

LEMMA 4.1. *Let $\overline{\phi}$ and $\underline{\phi}$ be defined by* (4.5). *Let $\xi_* \in \mathbb{R}$ be given. Suppose $\phi(\xi)$ is a bounded solution of* (YKODE) *that is defined for all $\xi \in \mathbb{R}$. Then there exists a $\xi+ > \xi_*$ satisfying*

$$\underline{\phi} \ < \ \phi(\xi_+) \ < \ \overline{\phi}.$$

Similarly, there is a $\xi_- < \xi_*$ satisfying the same

$$\underline{\phi} \ < \ \phi(\xi_-) \ < \ \overline{\phi}.$$

Proof. Consider any bounded solution ϕ of (YKODE). We first show that given any $\xi_* \in \mathbb{R}$, there exists a $\xi_+ > \xi_*$ such that $\underline{\phi} < \phi(\xi_+) < \overline{\phi}$. If this were not the case, $|r(\phi(\xi))| \geq |r\left(\overline{\phi}\right)| = |r\left(2\gamma\right)| > 0$ for all $\xi > \xi_*$, thus implying $\left|(g(\phi'')\phi'')'\right| \geq |r(2\gamma)| > 0$ for all $\xi > \xi*$, contradicting the fact that $g(s)s$ is bounded. Similarly, given $\xi_* \in \mathbb{R}$, there exists a $\xi_- < \xi_*$ with $\underline{\phi} < \phi(\xi_-) < \overline{\phi}$. ∎

LEMMA 4.2. *Let $\overline{\phi}$ and $\underline{\phi}$ be defined by* (4.5). *Any bounded solution $\phi(\xi)$ of* (YKODE) *that is defined for all $\xi \in \mathbb{R}$ must satisfy*

$$\underline{\phi} \ \leq \ \phi(\xi) \ \leq \ \overline{\phi} \tag{4.7}$$

for all $\xi \in \mathbb{R}$.

Proof. For the sake of contradiction, suppose there exists a ξ_1 with $\phi(\xi_1) > \overline{\phi}$. With the knowledge of Lemma 4.1, we choose some $\xi_0 < \xi_1$ with $\underline{\phi} < \phi(\xi_0) < \overline{\phi}$. By the same lemma, there also exists a $\xi_2 > \xi_1$ with $\phi(\xi_1) < \overline{\phi}$, so ϕ has a local maximum, $\phi(\xi_M) = \phi_M > \overline{\phi}$ with $\xi_0 < \xi_M < \xi_2$. Since $\phi' = 0$ at extrema,

$$\mathcal{L}_1(\xi_M) \ = \ \mathcal{R}(\phi(\xi_M)) \ > \ \mathcal{R}(-\gamma) \ > \ \mathcal{R}(\gamma).$$

Since ϕ is smooth, there exists some $\overline{\xi}$ with $\xi_M < \overline{\xi} < \xi_2$ satisfying $\phi(\overline{\xi}) = \overline{\phi}$ and $\phi(\xi) < \overline{\phi}$ for all $\xi \in (\overline{\xi}, \xi_1]$. There are two possible behaviors of $\phi(\xi)$ for $\xi > \overline{\xi}$. Either ϕ has an extrema $\phi(\xi_*) < \overline{\phi}$ or ϕ is monotonically decreasing for $\xi > \overline{\xi}$. Consider the first case. Since $\phi(\xi_*)$ is an extrema,

$$\mathcal{L}_1(\xi_*) \ = \ \mathcal{R}(\phi(\xi_*)) < \ \mathcal{R}(\phi(\xi_M)) \ = \ \mathcal{L}_1(\xi_M).$$

This contradicts the fact that $\mathcal{L}_1$ is strictly increasing, since $\xi_M < \xi_*$. Now suppose $\phi(\xi)$ decreases monotonically for $\xi > \overline{\xi}$. Since ϕ is bounded, it approaches a limit as $\xi \to \infty$. This limit must be either γ or $-\gamma$, or $g(\phi'')\phi''$ would blow up as argued in the proof of Lemma 4.1. For each limit,

$$\lim_{\xi \to \infty} \mathcal{L}_1(\xi) < \mathcal{L}_1(\xi_M),$$

contradicting the fact that $\mathcal{L}_1$ increases. It follows that $\phi(\xi) < \overline{\phi}$ for all ξ. A similar argument shows that ϕ is bounded below by $\underline{\phi}$. ∎

As already noted, $|\phi''| = 1$ is a singular value for (YKODE). We use the Lyapunov function to prove the following lemma, which shows that smooth heteroclinic orbits are forbidden from crossing this value. Lemma 4.3 is essential for showing that (YKODE) does not have a smooth heteroclinic orbit connecting L to R when γ is too large.

LEMMA 4.3. *Let $\phi(\xi)$ be a smooth heteroclinic orbit connecting L to R. Then for all ξ, $|\phi''(\xi)| \leq 1$.*

Proof. We show $\phi''(\xi) \leq 1$. Proving $\phi''(\xi) \geq -1$ follows the same line of argument. Suppose that ϕ is a smooth trajectory for which there exists a ξ_* such that $\phi''(\xi_*) > 1$.

We show that ϕ cannot connect L to R. Our argument follows directly from the ODE and its Lyapunov function. Since ϕ is a smooth heteroclinic orbit,

$$\lim_{\xi\to-\infty}\phi''(\xi)=0,$$

so we can find some ξ_c so that $\phi''(\xi_c)=1$ and $\phi''(\xi)>1$ for $\xi\in(\xi_c,\xi_*]$. Since ϕ is smooth, and $g'(\phi'')\phi''+g(\phi'')=0$ for $\phi''=1$, we must have $r(\phi(\xi_c))=0$, and therefore $\phi(\xi_c)=\pm\gamma$. Suppose $\phi(\xi_c)=\gamma$. Then the Lyapunov function implies $\phi'(\xi_c)>0$, so there is some $\epsilon>0$ so that $\phi(\xi)>\gamma$ for $\xi\in(\xi_c,\xi_c+\epsilon)$. The ODE then implies $\phi'''(\xi)>0$ for $\xi\in(\xi_c,\xi_c+\epsilon)$, and since both $\phi'(\xi)$ and $\phi''(\xi)$ are positive on the same interval, ϕ will continue to grow without bound, prohibiting it from being a heteroclinic orbit.

Now suppose $\phi(\xi_c)=-\gamma$. Then the Lyapunov function implies $\phi'(\xi_c)<0$. We can pick a new $\epsilon>0$ such that $\phi(\xi)<-\gamma$ and $\phi''(\xi)>1$ for $\xi\in(\xi_c,\xi_c+\epsilon)$. The ODE then implies $\phi'''(\xi)>0$ on the same interval. In fact, the ODE ensures that this interval can be extended and ϕ'' will continue to increase until ϕ' becomes positive and ϕ once again intersects $-\gamma$. So there is some $\xi'>\xi_c$ with $\phi(\xi')=-\gamma$, $\phi'(\xi')>0$, and $\phi''(\xi')>1$. So $\mathcal{L}_1(\xi')>\mathcal{R}(-\gamma)$, and ϕ cannot be a heteroclinic orbit connecting L to R. ∎

4.3. Nonexistence of traveling waves for (YKODE). Integrating (YKODE) on an arbitrary interval $[\xi_1,\xi_2]$, we see

$$g(\phi''(\xi_2))\phi''(\xi_2)-g(\phi''(\xi_1))\phi''(\xi_1)=\int_{\xi_1}^{\xi_2}r(\phi(y))dy. \tag{4.8}$$

Since $|g(s)s|\le\frac{1}{2}$, smooth solutions of (YKODE) are restricted by

$$\left|\int_{\xi_1}^{\xi_2}r(\phi(y))dy\right|\le 1 \tag{4.9}$$

on any interval $[\xi_1,\xi_2]$. We now use (4.9) to show that when γ is too large, the You–Kaveh ODE does not have a smooth heteroclinic orbit between L and R.

THEOREM 4.4. *There exists a finite $C>0$ such that* (YKODE) *has no smooth solution satisfying*

$$\lim_{\xi\to-\infty}\phi(\xi)=\gamma \quad \textit{and} \quad \lim_{\xi\to+\infty}\phi(\xi)=-\gamma \tag{4.10}$$

when $\gamma>C$.

Proof. Suppose ϕ is a smooth solution of (YKODE) that satisfies (4.10). Then ϕ must be a heteroclinic orbit connecting L to R, and there exists at least one ξ with $\phi(\xi)=0$. Let ξ_0 be the minimum of all points ξ satisfying $\phi(\xi)=0$. Let ξ_- be the largest number satisfying both $\xi_-<\xi_0$ and $\phi(\xi_-)=\gamma$. Since $\phi'''>0$ when $-\gamma<\phi<\gamma$, $\phi'(\xi)\le 0$ for all $\xi\in[\xi_-,\xi_0]$. Otherwise both ϕ' and ϕ'' would become positive in (ξ_-,ξ_0). ϕ' would have to become negative again so that $\phi(\xi_0)=0$, but this would require that ϕ become larger than γ, contradicting the assumptions on ξ_-.

Let μ denote the minimum of ϕ' on $[\xi_-,\xi_c]$. Then restriction (4.9) implies

$$\begin{aligned}1\ge\int_{\xi_-}^{\xi_c}-r(\phi(s))ds&=\int_{\xi_-}^{\xi_c}-r(\phi(s))\frac{\phi'(s)}{\phi'(s)}ds\ge\frac{1}{\mu}\int_{\xi_-}^{\xi_c}-r(\phi(s))\phi'(s)ds\\&=\frac{1}{\mu}(\mathcal{R}(\gamma)-\mathcal{R}(0)).\end{aligned}$$

Since $\mu < 0$ and $\mathcal{R}(\gamma) < \mathcal{R}(0)$, the above gives

$$\mathcal{R}(0) - \mathcal{R}(\gamma) \leq |\mu| \,. \tag{4.11}$$

From the bounds on ϕ and ϕ'' given by Lemmas 4.2 and 4.3, we see that

$$|\mu| \leq 2\sqrt{2\gamma} \tag{4.12}$$

as a result of the following interpolation lemma.

LEMMA 4.5. *Suppose $f \in C^2(\mathbb{R})$ satisfies $|f| \leq M$ and $|f''| \leq C$. Then*

$$|f'| \leq 2\sqrt{CM}.$$

Proof. Given $x \in \mathbb{R}$, Taylor's theorem shows

$$f'(x) = \frac{f(x+2h) - f(x)}{2h} - f''(\xi)h \tag{4.13}$$

for all $h > 0$ and some $\xi \in [-h, h]$. The bounds on f and f'' give us

$$|f'(x)| \leq \frac{M}{h} + Ch.$$

Choosing $h = \sqrt{\frac{M}{C}}$ gives

$$|f'(x)|^2 \leq \left(\frac{M}{h} + Ch\right)^2 = 4MC. \qquad \square$$

Calculating

$$\mathcal{R}(0) - \mathcal{R}(\gamma) = \frac{1}{3}(\gamma)^3$$

and combining (4.11) with (4.12) proves Theorem 4.4. $\square$

Remark. Theorem 4.4 does not depend on the choice $g = \frac{1}{1+s^2}$. It relies only on the properties of thresholding functions as explained in [36] and in section 1.1. In particular, the nonexistence follows mainly from the nonmonotonicity of $g(s)s$.

4.4. The (YKODE) phase space. We rewrite (4.1) as a system of first order ODEs:

$$\phi' = v, \quad v' = w, \quad w' = -\frac{r(\phi)}{g'(w)w + g(w)}. \tag{4.14}$$

System (4.14) has two equilibrium points, $L = (\gamma, 0, 0)$ and $R = (-\gamma, 0, 0)$. A traveling wave solution of (YK) satisfying (2.2) corresponds to a heteroclinic orbit connecting L to R. Let $W^s(L)$ and $W^u(L)$ denote, respectively, the stable and unstable manifolds of L, and define $W^s(R)$ and $W^u(R)$ in the same way.

Since $\gamma > 0$, $W^u(L)$ and $W^s(R)$ are both two-dimensional with complex eigenvalues, while $W^u(R)$ and $W^s(L)$ are one-dimensional manifolds. We follow the method used in [8] and [14]. We illustrate the unstable manifold of L by considering a set of initial values near L and integrating (4.14) forward in time. Each trajectory will approach $W^u(L)$. To visualize the manifold, we mark the intersections of each computed trajectory with a two-dimensional plane (a Poincaré section) in the phase space.

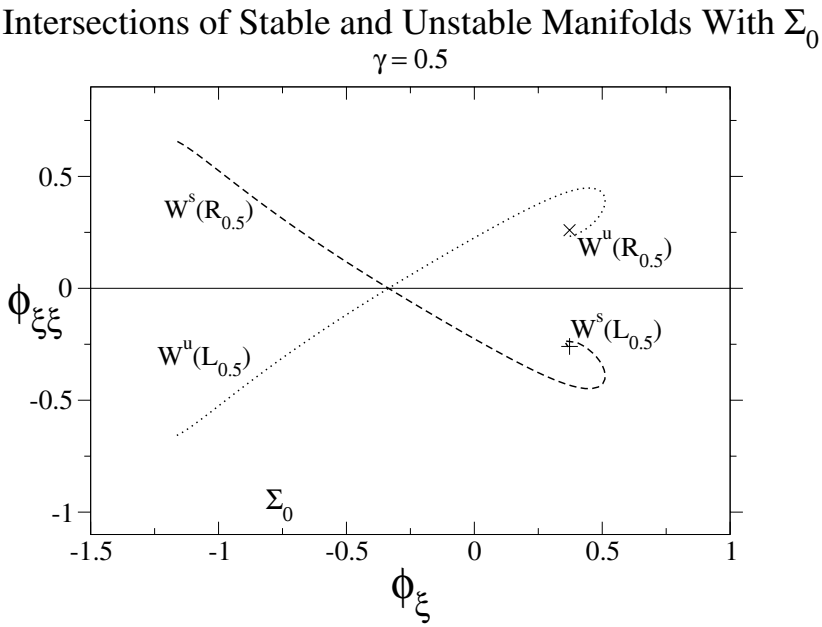

FIG. 6. *Cross-section of the phase plane of* (YKODE) *with* $\gamma = 0.5$. *We show the intersections of the stable and unstable manifolds of both equilibrium points with the plane* $\phi = 0$ *(denoted* Σ_0*).*

This plane is chosen so that all trajectories intersect the plane transversely. Any two-dimensional manifold intersects the plane on a curve, and any one-dimensional manifold intersects at a point. Picking initial points near R and integrating the ODE backward in time produces trajectories approaching $W^s(R)$. Traveling wave solutions of (4.14) correspond to intersections of $W^u(L)$ with $W^s(R)$.

In each figure, initial values are taken at a distance of 10^{-7} to 10^{-5} from the corresponding equilibrium point. We consider the plane $\phi = 0$, denoted by Σ_0. Any intersection of $W^u(L)$ with $W^s(R)$ must appear on Σ_0. The symmetry of (4.14) implies that the restriction of $W^u(L) \bigcap W^s(R)$ to Σ_0 occurs on the line $w = 0$.

Figure 6 shows the intersection of stable and unstable manifolds of u_L and u_R with Σ_0 for $\gamma = 0.5$. Since $W^u(L)$ and $W^s(R)$ intersect each other, there is a heteroclinic orbit connecting L to R. One end of $W^u(L)$ spirals around the one-dimensional manifold, $W^u(R)$. Symmetry gives the same relationship between $W^s(R)$ and $W^s(L)$. As γ is increased, the spiral structure of $W^u(L)$ shifts toward the line $w = 1$, while $W^s(R)$ shifts toward $w = -1$. Figure 7 demonstrates that the manifolds do not have this spiral structure on Σ_0 when γ is too large. The one-dimensional manifolds $W^s(L)$ and $W^s(R)$ no longer intersect Σ_0 when these spiral structures disappear. Further increasing γ moves $W^u(L)$ and $W^s(R)$ away from each other. For large enough γ, $W^u(L)$ and $W^s(R)$ do not intersect each other, as seen in Figure 7, where $\gamma = 1.3$.

In Figure 8, we draw $W^u(L)$ for a sequence of γ values. $W^s(R)$ is not shown, since it can be deduced by reflecting $W^u(L)$ across the line $w = 0$. The two manifolds intersect only when the restriction of $W^u(L)$ to Σ_0 intersects the line $w = 0$. $W^u(L)$ (and consequently $W^s(R)$) shifts away from the line $w = 0$ as γ increases. For large enough γ, $W^u(L)$ does not intersect the line $w = 0$ at Σ_0. As proved in Theorem 4.4, there is a value γ_c such that $W^u(L)$ and $W^s(L)$ do not intersect when $\gamma > \gamma_c$. Our numerical experiments suggest that $1.16 < \gamma_c < 1.17$.

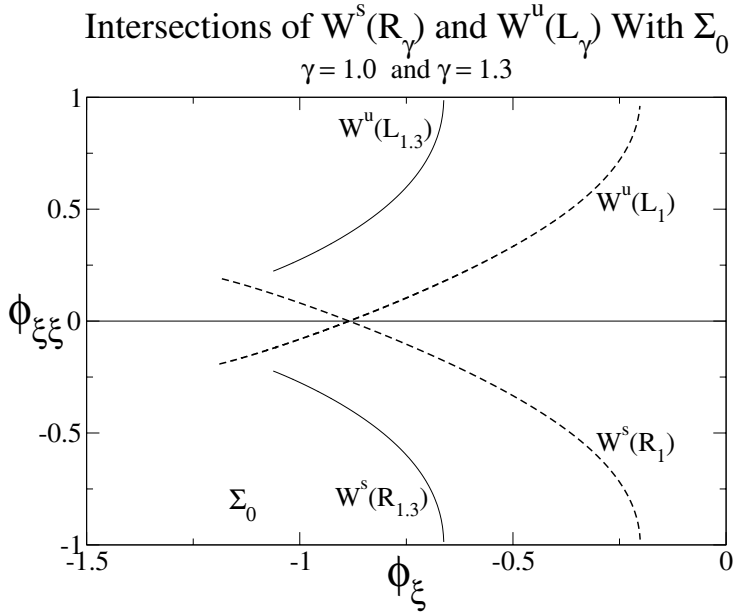

FIG. 7. *Changing manifolds of* (YKODE) *with increasing* γ. *The intersections of* $W^u(L)$ *and* $W^s(R)$ *with* Σ_0 *are shown for* $\gamma = 1.0$ *and* $\gamma = 1.3$. *In both cases,* $W^s(L)$ *and* $W^u(R)$ *do not intersect* Σ_0. *When* $\gamma = 1.3$, $W^u(L)$ *does not intersect* $W^s(R)$, *so there can be no traveling wave solution of the PDE.*

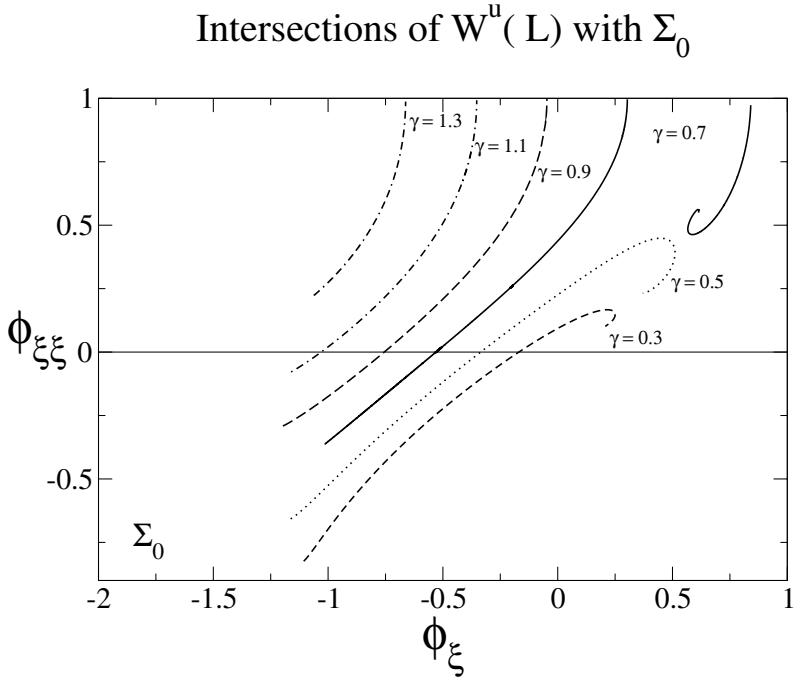

FIG. 8. $W^u(L) \bigcap \Sigma_0$ *for* (YKODE) *with different values of* γ. *A traveling wave solution exists when* $W^u(L) \bigcap \Sigma_0$ *intersects the line* $\phi'' = 0$. *We see that no such intersection exists for large enough* γ.

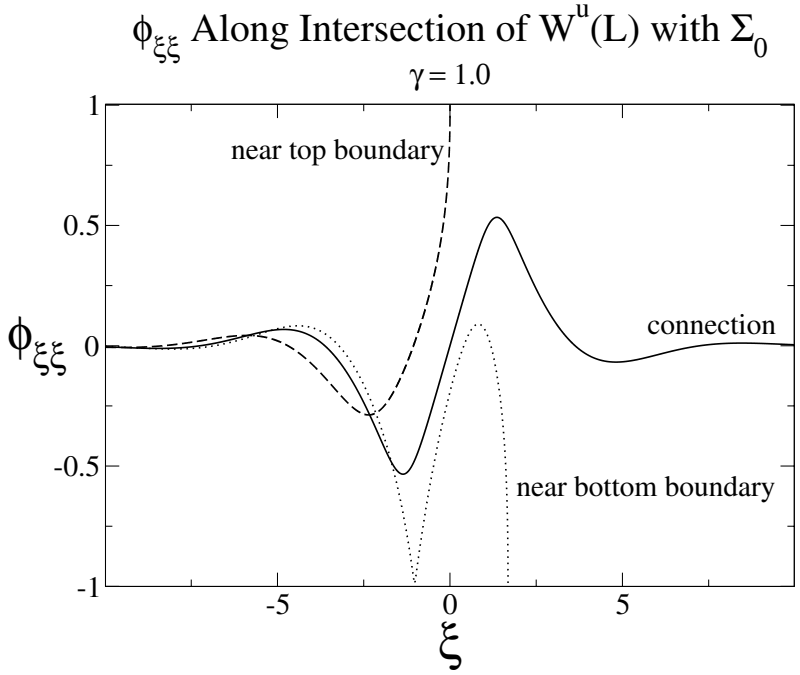

FIG. 9. *Trajectories of* (YKODE) *near the boundaries of* $W^u(L)$ *for* $\gamma = 1.0$. *We show the second derivatives of trajectories that pass near the top and bottom boundaries of* $W^u(L) \bigcap \Sigma_0$. *Trajectories near the top boundary have second derivatives approaching the singular value* $w = 1$ *as* ϕ *approaches* 0, *as can be seen from Figure* 7. *Trajectories near the bottom boundary have a second derivative near* $\phi'' = -1$ *but not where* $\phi = 0$. *The traveling wave solution's second derivative is shown for comparison.*

4.5. Manifold boundaries caused by singularities in solutions of (YKODE). $W^u(L)$ and $W^s(R)$ have boundaries caused by the ODE's singularity. Consider $\gamma = 1.0$, for which $W^u(L) \bigcap \Sigma_0$ is bounded above by $w = 1$. Certainly the manifold cannot extend past $w = 1$, since (4.14) is singular there, but there is also a boundary on the opposite end of $W^u(L) \bigcap \Sigma_0$. This boundary is far from either line of singularity, $w = \pm 1$. Figure 9 shows the second derivative of trajectories near these top and bottom boundaries of $W^u(L) \bigcap \Sigma_0$. Let ξ_0 denote the value of ξ for which a given trajectory $\phi(\xi)$ intersects Σ_0 (ξ_0 could be different for each trajectory). Near the top boundary, $\phi''(\xi_0)$ gets arbitrarily close to $\phi''(\xi_0) = 1$. Trajectories near the bottom boundary approach $\phi''(\xi_c) = -1$ for some $\xi_c < \xi_0$.

The singularities of solutions to (YKODE) are similar to those of (PMODE), but they occur in higher derivatives. Consider a trajectory ϕ with second derivative approaching -1 (the case $\phi'' \to 1$ is very similar). Assume there is some ξ_* with

$$\lim_{\xi \to \xi_*} \phi''(\xi) = -1 \text{ and } \lim_{\xi \to \xi_*} \phi(\xi) = \phi^*.$$

Again we have multiple cases, but this time they depend on the zeros of $r(\phi)$.

Case 1. $r(\phi^*) \neq 0$. This corresponds to the case $\gamma > 1$ for (PMODE). But now the singularity occurs in ϕ''' as $\xi \to \xi_*$:

$$\phi''(\xi) \sim 2\sqrt{r(\phi^*)(\xi_* - \xi)} - 1. \tag{4.15}$$

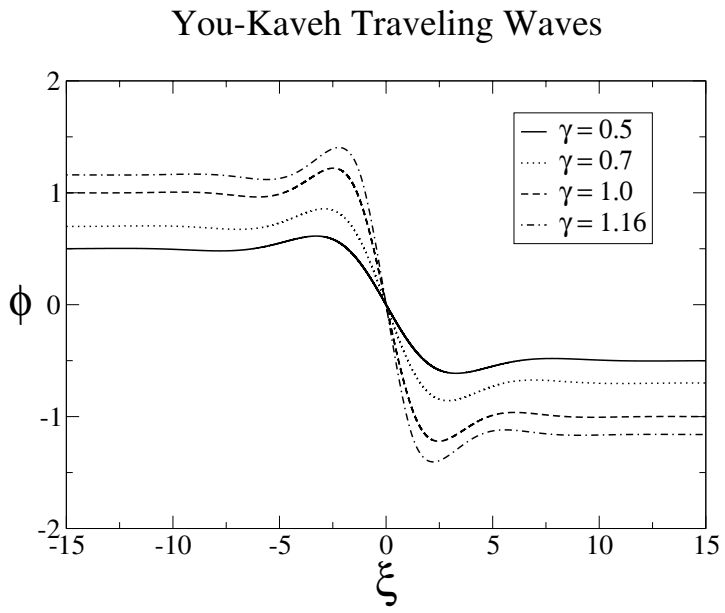

FIG. 10. *Traveling wave solutions of* (YK). *We show them for different values of* γ.

This singularity is demonstrated by trajectory near the top boundary of $W^u(L)$, drawn in Figure 9.

Case 2. $r(\phi^*) = 0$. Either $\phi^* = \gamma$ or $\phi^* = -\gamma$. It is easy to check that

$$\phi''(\xi) \sim \sqrt{2}|\xi - \xi_*| - 1. \tag{4.16}$$

Case 2 is demonstrated by the trajectory near the bottom boundary of $W^u(L)$, as seen in Figure 9. It also corresponds to a critical case for traveling wave solutions of (YKODE). We expect that there is some γ_c for which (YKODE) has a nonsmooth traveling wave solution analogous to the solution of (PMODE) for $\gamma = 1$.

4.6. Traveling wave solutions of (YK). Solutions of ODE (4.1) that correspond to traveling waves connecting L to R are given by the intersection of $W^u(L)$ with $W^s(R)$. Our study of the phase space suggests that there is at most one such intersection for any given γ. The traveling waves shown in Figure 10 were produced by finding this intersection.

In Figure 11, we provide graphs of the second derivative of traveling wave solutions. In each case, $|\phi''|$ is bounded by 1 as expected. The local extrema of ϕ'' are achieved at $\phi = \pm\gamma$, where $\phi''' = 0$. As γ increases, these extreme values approach the singular values $\phi'' = \pm 1$. Because of the ODE's symmetry, ϕ'' approaches a singular value in two places. ϕ'' approaches -1 when $\phi = \gamma$, and it approaches $+1$ when $\phi = -\gamma$.

To illustrate that the traveling waves are stable for the PDE dynamics, we implement (YK) with a fully implicit scheme. We use centered differences for all spatial derivatives, including the Burgers term, which is approximated by centered differences in flux form. We use a Newton solver and an adaptive time step. The time step was adjusted to expedite convergence of the Newton method, as was done for (PM) in

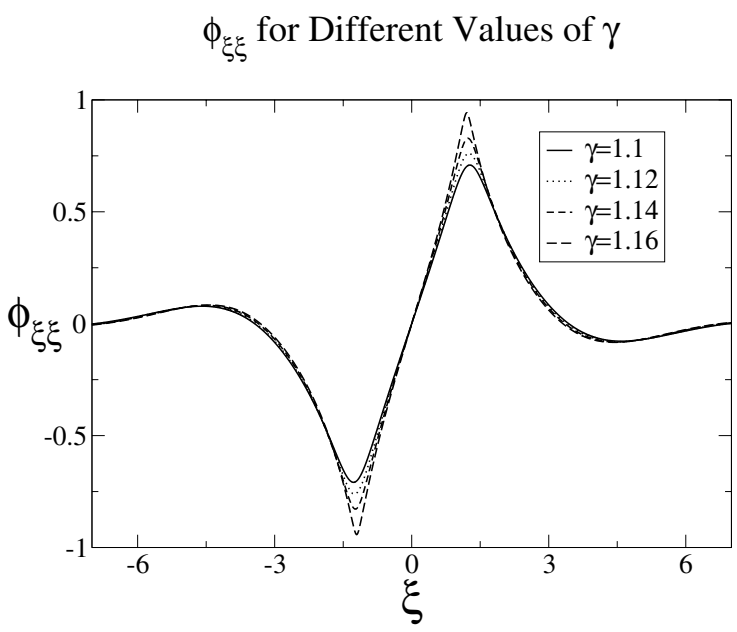

FIG. 11. *The second derivatives of traveling wave solutions of* (YK). *Traveling waves for* γ *near* 1.16 *have second derivatives near the singular value* $w = -1$.

section 3.4. The correspondence between (YK) and (YKODE) is not as clear as it is for (PM) and (PMODE). The numerics become very difficult for γ near the range of nonexistence of traveling waves. In this parameter range, the PDE numerics do not converge nicely to a traveling wave solution, even when our ODE numerics suggest one exists. It is not clear whether this difficulty results from the numerics or from the PDE. We show an example with a smaller γ in Figure 12. In this case, the PDE solution clearly converges to the solution of (YKODE).

5. Tumblin–Turk with advection. We show that (TT) is qualitatively different from both (PM) and (YK). We first use a topological argument to prove that for all $\gamma > 0$, (TTODE) has an orbit corresponding to a traveling wave solution of (TT). Our primary tool is the Conley index, as discussed in [40]. We use standard methods [8, 37], but the particular nonlinear structure of (TTODE) requires new a priori bounds and estimates. We rely on the observation that (TTODE) can be rewritten as

$$r(\phi) = -(\arctan(\phi''))' \tag{5.1}$$

when $g(s) = \frac{1}{1+s^2}$. The analysis consequently depends very much on this particular choice of g.

In section 5.3, we present phase plane illustrations that contrast solutions of (TTODE) to those of (YKODE) and (PMODE). We conclude our discussion of the Tumblin–Turk equations with numerical simulations of (TT).

5.1. Lyapunov function for (TTODE). We seek a Lyapunov function, $\mathcal{L}_2(\xi)$, for (TTODE). Let $\mathcal{R}(s)$ denote a primitive of $r(s)$. Multiplying (5.1) by ϕ' and integrating produces

$$\mathcal{R}(\phi) = -\arctan(\phi'')\phi' + \int^{\xi} \arctan(\phi''(s))\phi''(s)ds.$$

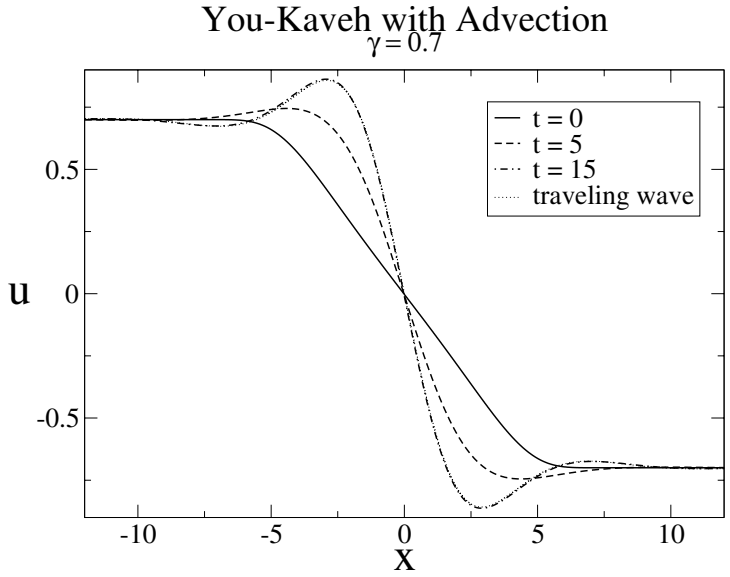

FIG. 12. *Approximate solution of* (YK). *When* $\gamma = 0.7$, u *approaches the traveling wave solution given by* (YKODE).

Since $\arctan(s)s \geq 0$ for all s, we easily check that

$$\mathcal{L}_2(\xi) = \mathcal{R}(\phi(\xi)) + \arctan(\phi''(\xi))\phi'(\xi) \tag{5.2}$$

satisfies

$$\frac{d}{d\xi}\mathcal{L}_2(\xi) = \arctan(\phi''(\xi))\phi''(\xi) \geq 0. \tag{5.3}$$

As was the case for $\mathcal{L}_1$, $\mathcal{L}_2(\xi) = \mathcal{R}(\phi(\xi))$ at zeros of ϕ' and ϕ''. This establishes the entropy condition $\gamma > 0$ and the following lemma.

LEMMA 5.1. *Let* $\overline{\phi}$ *and* $\underline{\phi}$ *be defined by* (4.5). *Any bounded smooth solution* ϕ *of* (TTODE) *that is defined on the real line must satisfy*

$$\underline{\phi} \;\leq\; \phi(\xi) \;\leq\; \overline{\phi} \tag{5.4}$$

for all $\xi \in \mathbb{R}$.

Proof. The proof follows the same argument as that of Lemma 4.2. ∎

5.2. System of ODEs for (TTODE). We rewrite (TTODE) as a system of three ODEs:

$$\phi' = v, \quad v' = \tan(w), \quad w' = -r(\phi). \tag{5.5}$$

System (5.5) has two equilibrium points, $L = (\gamma, 0, 0)$ and $R = (-\gamma, 0, 0)$. We use Conley index theory to prove the existence of a heteroclinic orbit connecting L to R. To do this, we first find uniform bounds for all bounded solutions (ϕ, v, w) of (5.5). Lemma 5.1 provides such a bound for ϕ. It is particularly important to find a bound C such that $|w| \leq C < \frac{\pi}{2}$. To do so, we first examine $v' = \phi''$.

LEMMA 5.2. *Any bounded smooth solution ϕ of* (TTODE) *satisfies*

$$\int_{-\infty}^{\infty} \arctan(\phi''(s))\phi''(s)ds \;\leq\; \mathcal{R}(-\gamma) - \mathcal{R}(\gamma) \;=\; \frac{2}{3}\gamma^3. \tag{5.6}$$

Proof. We follow an argument used in the proof of Theorem 4.8 in [10]. Let ϕ be a bounded solution of (TTODE). Bound (5.6) is obvious if either $\phi(\xi) = \gamma$ or $\phi(\xi) = -\gamma$ for all ξ. Since $L = (\gamma, 0, 0)$ and $R = (-\gamma, 0, 0)$ are the only equilibrium points of (TTODE), we now assume that ϕ is nonconstant. We first examine the behavior of $\phi(\xi)$ as $\xi \to \infty$. There are two cases to consider, depending on the set of extrema of ϕ.

Case 1. Suppose there exists a ξ_M such that ϕ has no extrema for $\xi > \xi_M$. Then ϕ approaches an equilibrium point as $\xi \to \infty$. Since L is increasing, $\phi \to -\gamma$ as $\xi \to \infty$; otherwise all extrema of ϕ would be less than $\underline{\phi}$, and ϕ would grow without bound as $\xi \to -\infty$. We therefore have

$$\int_{0}^{\infty} \arctan(\phi''(s))\phi''(s)ds \;=\; \mathcal{R}(-\gamma) - \mathcal{R}(\phi(0)).$$

Case 2. Now assume that there is no such ξ_M. Since ϕ solves (TTODE), it is analytic (see, e.g., [41]) and must have a countable set of extrema with no limit point. Suppose the extrema occur at ξ_i with $\xi_i > 0$ and $\xi_i < \xi_{i+1}$. The Lyapunov function implies that $\mathcal{R}(\xi_i)$ is a bounded increasing sequence, and we therefore have $\mathcal{R}(\xi) \to \mathcal{R}_+$ for some $\mathcal{R}_+ \leq \mathcal{R}(-\gamma)$. For each ξ_i,

$$\int_{0}^{\xi_i} \arctan\left(\phi''(s)\right)\phi''(s)ds \;=\; \mathcal{R}\left(\phi(\xi_i) - \mathcal{R}\left(\phi(0)\right) \;\leq\; \mathcal{R}(-\gamma) - \mathcal{R}(\phi(0)).\right.$$

The monotone convergence theorem gives us

$$\int_{0}^{\infty} \arctan(\phi''(s))\phi''(s)ds = \mathcal{R}_+ - \mathcal{R}(\phi(0)) \leq \mathcal{R}(-\gamma) - \mathcal{R}(\phi(0)). \tag{5.7}$$

Similar arguments show

$$\int_{-\infty}^{0} \arctan(\phi''(s))\phi''(s)ds \leq \mathcal{R}(\phi(0)) - \mathcal{R}(\gamma). \tag{5.8}$$

Combining (5.7) and (5.8) completes the proof. ∎

We interpret Lemma 5.2 to mean that $\phi'' = v'$ is *almost* L^1, since $\arctan(s)s$ is linear in s for large s. Specifically, for any $\epsilon > 0$, we define $S = \{s : |\phi''(s)| > \epsilon\}$ and discover

$$\begin{aligned} \int_S |\phi''(s)|ds &\leq \frac{1}{\arctan\epsilon} \int_S |\arctan(\phi''(s))\phi''(s)|ds \\ &\leq \frac{1}{\arctan\epsilon} \int_{-\infty}^{\infty} \arctan(\phi''(s))\phi''(s)ds \leq \frac{2}{3\arctan\epsilon}\gamma^3. \end{aligned} \tag{5.9}$$

We now show that w is bounded away from $\pm\frac{\pi}{2}$, the asymptotes of $\tan w$.

LEMMA 5.3. *There exists a positive $C_1 < \frac{\pi}{2}$ such that for any bounded solution (ϕ, v, w) of system* (5.5)*, $|w| \leq C_1$ for all $\xi \in \mathbb{R}$.*

Proof. Since $\underline{\phi} \leq \phi(\xi) \leq \overline{\phi}$ for all ξ, $|-r(\phi)| \leq \gamma^2$, which by (5.5) implies a uniform Lipschitz bound for w,

$$w(\xi_0 - h) \geq w(\xi_0) - \gamma^2 h \text{ for all } \xi_0 \text{ and all } h > 0. \tag{5.10}$$

We use (5.9) with the uniform Lipschitz continuity of w to derive a pointwise bound on w. We focus on bounding w away from $w = +\frac{\pi}{2}$. To make use of (5.9), we must find an interval on which w is bounded away from zero. Pick ξ_0 with $\frac{\pi}{4} < w(\xi_0) < \frac{\pi}{2}$. If no such ξ_0 exists, then $w(\xi) \leq \frac{\pi}{4}$. Choose $\delta > 0$ so

$$\frac{\pi}{4} > w(\xi_0) - \gamma^2\delta \geq \frac{\pi}{6}, \tag{5.11}$$

also implying by (5.10) that $w(\xi) \geq \frac{\pi}{6}$ for all $\xi \in [\xi_0 - \delta, \delta]$. Let

$$S = \left\{ \xi : \phi''(\xi) \geq \frac{\sqrt{3}}{3} = \tan\frac{\pi}{6} \right\}.$$

Then Lemma 5.2 ensures

$$\int_S |v'| = \int_S |\phi''| \leq \frac{6}{\pi}\left(\mathcal{R}(-\gamma) - \mathcal{R}(\gamma)\right) = \frac{4}{\pi}\gamma^3. \tag{5.12}$$

Now using (5.10) and (5.11), we calculate

$$\begin{aligned}
\int_S |v'(s)|ds &\geq \int_{\xi_0-\delta}^{\xi_0} |v'(s)|ds \\
&= \int_{\xi_0-\delta}^{\xi_0} |\tan(w(s))|ds \\
&\geq \int_{\xi_0-\delta}^{\xi_0} \tan(w(\xi_0) - \gamma^2(\xi_0 - s))ds \\
&= \frac{1}{\gamma^2}\log\left|\frac{\cos(w(\xi_0) - \gamma^2\delta)}{\cos(w(\xi_0))}\right| \\
&\geq \frac{1}{\gamma^2}\log\left|\frac{\cos\frac{\pi}{4}}{\cos(w(\xi_0))}\right|.
\end{aligned}$$

Combining this with (5.12), we see

$$\frac{1}{\gamma^2}\log\left|\frac{\sqrt{2}}{2\cos(w(\xi_0))}\right| \leq \frac{4}{\pi}\gamma^3, \tag{5.13}$$

so

$$\cos(w(\xi_0)) \geq \frac{\sqrt{2}}{2}e^{-\gamma^2(\frac{4}{\pi}\gamma^3)} > 0, \tag{5.14}$$

and

$$w(\xi_0) \leq C_1 := \arccos\left(\frac{\sqrt{2}}{2}e^{-\frac{4}{\pi}\gamma^5}\right) < \frac{\pi}{2}. \tag{5.15}$$

The same argument with slight adjustments shows that

$$w(\xi_0) \geq -C_1 = -\arccos\left(\frac{\sqrt{2}}{2}e^{-\frac{4}{\pi}\gamma^5}\right) > -\frac{\pi}{2}. \qquad \square$$

COROLLARY 5.4. *There exists a $C_2 > 0$ satisfying $|v| \leq C_2$.*

Proof. Since w is bounded in an interval strictly contained within $(-\frac{\pi}{2}, \frac{\pi}{2})$, we have a bound on $\phi'' = \tan w$. We use Lemma 4.5 to bound $v = \phi'$. $\square$

THEOREM 5.5. *Given any $\gamma > 0$, there exists a solution ϕ of* (TTODE) *such that $\phi(\xi) \to \gamma$ as $\xi \to -\infty$, and $\phi(\xi) \to -\gamma$ as $\xi \to \infty$.*

Proof. Our proof centers on the Conley index. We refer the reader to [40], which contains an excellent description of Conley index theory. Let C_1 and C_2 be given by Lemma 5.3 and Corollary 5.4. Define the set

$$N = \left\{ (\phi, v, w) : \begin{array}{r} \underline{\phi} \leq \phi \leq \overline{\phi} \\ |v| \leq C_2 \\ |w| \leq C_1 \end{array} \right\}. \tag{5.16}$$

N is an isolating neighborhood, as all bounded trajectories are strictly contained within the interior of N. As explained in Theorem 22.18 of [40], N contains an isolating block, B. Isolating blocks of (TTODE) are special isolating sets whose boundary points immediately leave the set in positive or negative time under the flow defined by (TTODE). The Conley index is the homotopic equivalence class of the quotient space B/b^+, where b^+ is the set of all points on ∂B that leave B in positive time.

Let $\beta \in \mathbb{R}$, and define the continuous deformation of (TTODE),

$$r(\phi) + \beta = -g(\phi'')\phi'''. \tag{5.17}$$

Let $\beta_0 = \frac{\gamma^2}{2}$. Consider $\beta \in [0, \beta_0)$. The new system has a new function

$$\mathcal{R}_\beta(\phi) = \frac{1}{6}\phi^3 + \left(\beta - \frac{1}{2}\gamma^2\right)\phi.$$

The new Lyapunov function is found by replacing $\mathcal{R}$ with $\mathcal{R}_\beta$. The system has two equilibrium points: $L_\beta = (0, 0, \sqrt{\gamma^2 - 2\beta})$ and $R_\beta = (0, 0, -\sqrt{\gamma^2 - 2\beta})$. The system has new upper and lower bounds for all bounded solutions:

$$\overline{\phi}_\beta = 2\sqrt{\gamma^2 - 2\beta} \leq \overline{\phi}$$

and

$$\underline{\phi}_\beta = -2\sqrt{\gamma^2 - 2\beta} \geq \underline{\phi}.$$

It is easy to check that B is an isolating block for the adjusted system with $0 \leq \beta \leq \beta_0$.

When $\beta = \beta_0$, the only bounded trajectory of (5.17) is the constant function $\phi = 0$, so B remains an isolating block. Choosing $\beta > \beta_0$ produces a differential equation with no equilibrium points. B remains an isolating block of the flow and contains no isolated invariant set (other than the null set). It follows that the homotopic equivalence class of $B/b+$ is that of the null set, implying the existence of an orbit of (TTODE) connecting L and R (see Theorem 22.33 in [40]). The Lyapunov function ensures that the trajectory flows from L to R. $\square$

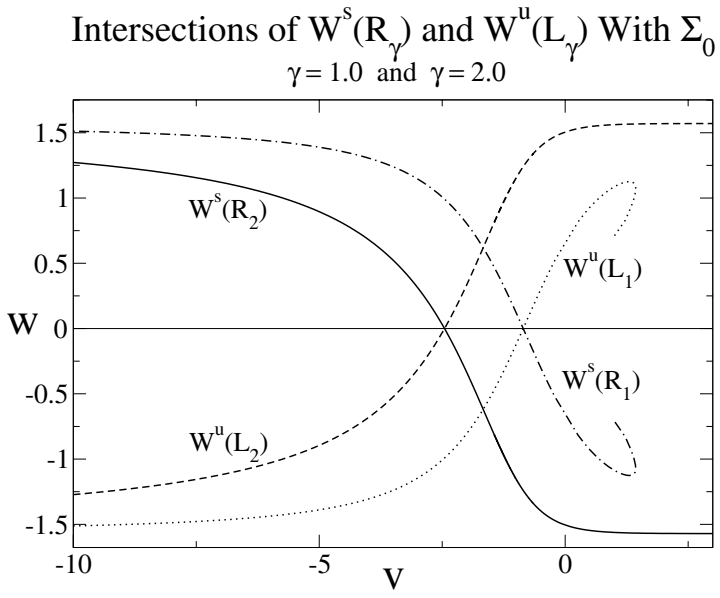

FIG. 13. *Changes of the manifolds for* (TTODE) *with increasing* γ. *The intersections of* $W^u(L)$ *and* $W^s(R)$ *with* Σ_0 *are shown for* $\gamma = 1.0$ *and* $\gamma = 2.0$.

5.3. The (TTODE) phase space. As suggested by our analysis of both equations, the phase plane geometry of (TTODE) is remarkably different from that of (YKODE). Using the method discussed in section 4.4, we visualize the phase space by considering the cross-section $u = 0$, denoted by Σ_0. Any intersection of $W^u(L)$ with $W^s(R)$ is visible on Σ_0, where it must occur on the line $w = 0$. We draw $W^u(L)$ by computing trajectories with initial conditions near L and marking their intersections with Σ_0. $W^s(R)$ is drawn similarly but by numerically integrating (TTODE) backward in time.

Smooth curves in the phase space must lie between the two planes $w = \pm\frac{\pi}{2}$, since $v = \tan w$. Figures 13 and 14 show the intersections of $W^s(R)$ and $W^u(L)$ with Σ_0 for various values of γ. Since $W^s(R)$ and $W^u(L)$ do not have boundaries caused by singularities of (TTODE), both manifolds stretch from $w = -\frac{\pi}{2}$ to $w = \frac{\pi}{2}$, even for large γ. This allows an intersection at $w = 0$ for all $\gamma > 0$; increasing γ shifts only the manifolds in the $-v = -\phi'$ direction. This is remarkably different from the You–Kaveh ODE (YKODE), for which $W^s(R)$ and $W^u(L)$ have boundaries that allow the manifolds to shift away from each other when γ is increased.

5.4. Traveling wave solutions of (TT). Figure 15 shows traveling wave solutions of (TT) for a series of γ-values. Each traveling wave was produced by finding the intersection of $W^u(L)$ with $W^s(R)$ in the phase space of (TTODE). As the jump height from u_L to u_R increases, so does the traveling wave's slope near the jump. Although the ODE solutions are smooth, the jump transition can be so severe that when viewed at large length scales the solution appears to have a shock. This is demonstrated when $\gamma = 7$, as shown in Figure 15.

Numerical examples suggest that the heteroclinic orbits of (TTODE) are stable traveling wave solutions of (TT). To numerically integrate (TT), we use the change

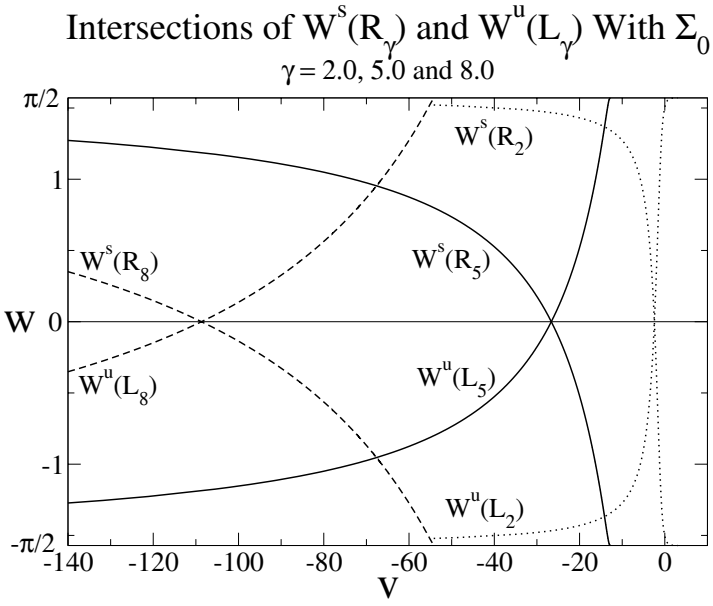

FIG. 14. *Changes of the manifolds for* (TTODE) *with increasing* γ. *The intersections of* $W^u(L)$ *and* $W^s(R)$ *with* Σ_0 *are shown for* $\gamma = 2.0, 5.0,$ *and* 8.0. *Each manifold's structure persists while increasing* γ.

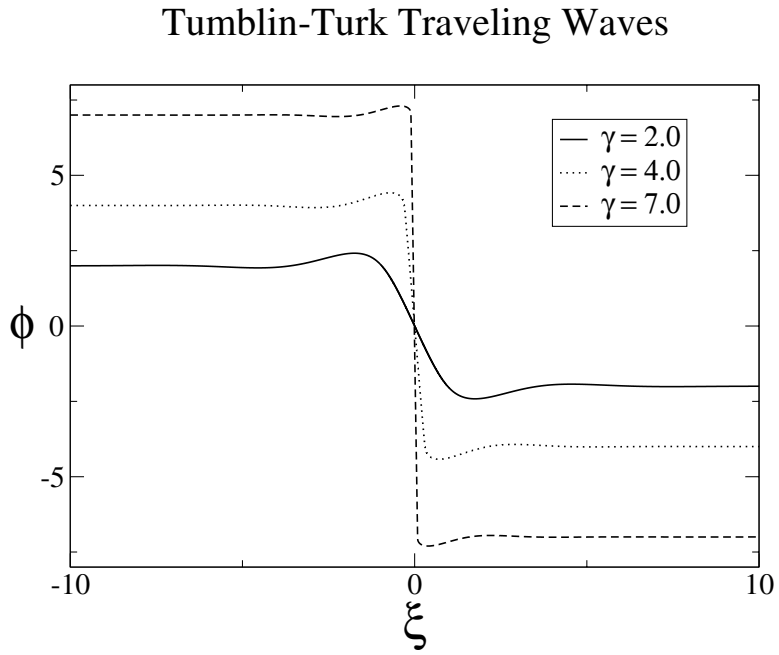

FIG. 15. *Heteroclinic orbits of* (TTODE). *At this length scale, the traveling wave solution for* $\gamma = 7$ *appears to have a shock.*

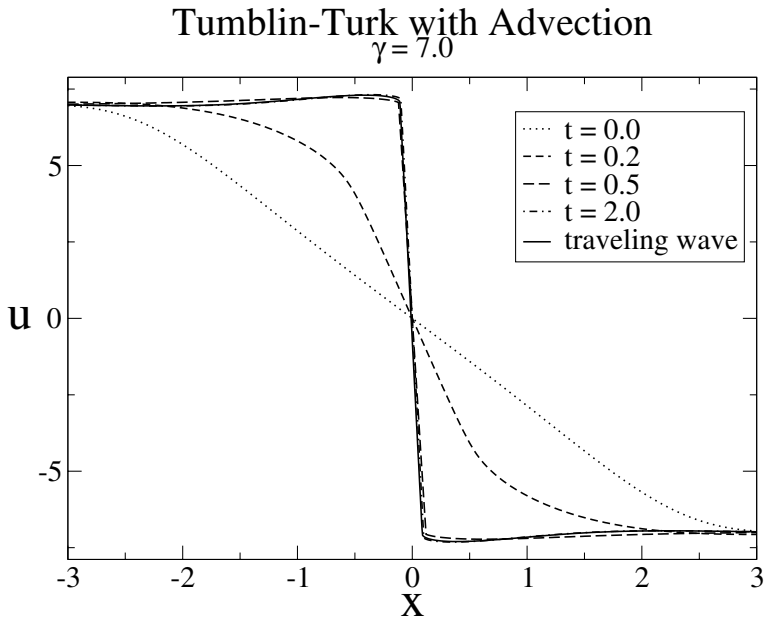

FIG. 16. *Numerically integrated solution of* (TT) *for* $\gamma = 7.0$.

of variables $w = \arctan u_{xx}$ and solve the nonlinear system

$$u_t + uu_x = w_{xx}, \qquad \tan w = -u_{xx} \tag{5.18}$$

using a fully implicit scheme with centered differences in space. We use Newton's method and an adaptive time step, as we did for (PM) in section 3.4. The change of variables $w = \arctan u_{xx}$ is used to ensure that u_{xx} remains bounded. See [7] for a discussion on numerically implementing the fourth order diffusion.

Figure 16 shows the behavior of u, given an initial condition near the traveling wave profile. The computations suggest that the traveling wave is a stable solution of the PDE.

6. Conclusions. We have considered traveling wave solutions of the advection-diffusion equations

$$u_t + \left(\frac{1}{2}u^2\right)_x = -(g(u_{xx})u_{xx})_{xx} \tag{YK}$$

and

$$u_t + \left(\frac{1}{2}u^2\right)_x = -(g(u_{xx})u_{xxx})_x, \tag{TT}$$

with $g(s) = \frac{1}{1+s^2}$, in order to clearly illustrate the features of higher order nonlinear diffusion equations recently proposed for use in image processing.

The advection term $(\frac{1}{2}u^2)_x$ in (YK) and (TT) serves two roles. First, it allows for traveling wave solutions that approximate shocks, which in images correspond to edges. By converting the problem to one of traveling waves, we reduce a fourth order PDE to a third order ODE for which we are able to prove rigorous results and perform clear phase space computations. Second, advective PDEs combining similar diffusion terms are being used for such processes as image inpainting [2, 3]. Thus these kinds of equations are interesting for image processing in their own right.

We discover a fundamental difference between solutions of (YK) and (TT). Smooth traveling waves solutions of (YK) do not exist for sufficiently large jump height, whereas solutions of (TT) exist for all jumps. This suggests that the dynamics of the full PDE (YK) is quite different from that of (TT). In a separate paper, we prove that in one dimension the PDE (TT) without advection has globally smooth solutions, given smooth initial data. The study in this paper would lead us to conjecture that (YK) without advection does have finite time singularities in u_{xx}, just as the classical Perona–Malik equation has finite time singularities in the slope.

Although the PDE numerics suggest that the smooth traveling waves are stable, a rigorous proof of this is still forthcoming. Rigorous stability results for traveling wave solutions of second order convection-diffusion equations include Goodman's proof of multidimensional stability of viscous scalar shock fronts [22] and Osher and Ralston's proof of stability of traveling wave solutions of the convective porous media equation [35]. Fourth order traveling waves are more difficult to analyze due to the lack of a maximum principle and the fact that the traveling waves themselves often do not have a closed form expression. In [9], Evans function techniques are used to prove instability of fourth order thin film traveling waves, although they establish only a consistent condition for stability.

Our work is done entirely in one dimension, but there is at least one obvious extension to two dimensions. Traveling wave solutions of the model equations correspond to plane wave solutions of the equations with diffusions in two dimensions, while the advection term remains only in the x-direction. These plane waves move in the x-direction and do not depend on y. In physical applications, the existence and stability of plane waves is relevant for pattern formation [5, 22, 42]. Analagous questions in imaging are interesting and have not been explored to our knowledge.

REFERENCES

[1] L. Alvarez, F. Guichard, P.-L. Lions, and J.-M. Morel, *Axioms and fundamental equations of image processing*, Arch. Rational Mech. Anal., 123 (1993), pp. 199–257.

[2] M. Bertalmio, A. Bertozzi, and G. Sapiro, *Navier-Stokes, fluid dynamics and image and video inpainting*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Kayai, HI, 2001, pp. 355–362.

[3] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, *Image inpainting*, in Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, K. Akeley, ed., ACM Press/Addison-Wesley, New York, 2000, pp. 417–424.

[4] A. L. Bertozzi, *Symmetric singularity formation in lubrication-type equations for interface motion*, SIAM J. Appl. Math., 56 (1996), pp. 681–714.

[5] A. L. Bertozzi and M. P. Brenner, *Linear stability and transient growth in driven contact lines*, Phys. Fluids, 9 (1997), pp. 530–539.

[6] A. L. Bertozzi, M. P. Brenner, T. F. Dupont, and L. P. Kadanoff, *Singularities and similarities in interface flow*, in Trends and Perspectives in Applied Mathematics, L. Sirovich, ed., Appl. Math. Sci. 100, Springer-Verlag, New York, 1994, pp. 155–208.

[7] A. L. Bertozzi and J. B. Greer, *Low-curvature image simplifiers: Global regularity of smooth solutions and Laplacian limiting schemes*, Comm. Pure Appl. Math., 57 (2004), pp. 764–790.

[8] A. L. Bertozzi, A. Münch, and M. Shearer, *Undercompressive shocks in thin film flows*, Phys. D, 134 (1999), pp. 431–464.
[9] A. L. Bertozzi, A. Münch, M. Shearer, and K. Zumbrun, *Stability of compressive and undercompressive thin film travelling waves*, European J. Appl. Math., 12 (2001), pp. 253–291.
[10] A. L. Bertozzi and M. Shearer, *Existence of undercompressive traveling waves in thin film equations*, SIAM J. Math. Anal., 32 (2000), pp. 194–213.
[11] S. Boatto, L. Kadanoff, and P. Olla, *Travelling wave solutions to thin film equations*, Phys. Rev. E(3), 48 (1993), pp. 4423–4431.
[12] A. Bourlioux, A. J. Majda, and V. Roytburd, *Theoretical and numerical structure for unstable one-dimensional detonations*, SIAM J. Appl. Math., 51 (1991), pp. 303–343.
[13] M. Brenner, P. Constantin, L. P. Kadanoff, and S. C. Venkataramani, *Diffusion, attraction, and collapse*, Nonlinearity, 12 (1999), pp. 1071–1098.
[14] R. Buckingham, M. Shearer, and A. Bertozzi, *Thin film traveling waves and the Navier slip condition*, SIAM J. Appl. Math., 63 (2003), pp. 722–744.
[15] J. Canny, *A computational approach to edge detection*, IEEE Trans. Pattern Anal. Machine Intell., PAMI-8 (1986), pp. 679–698.
[16] F. Catté, P.-L. Lions, J.-M. Morel, and T. Coll, *Image selective smoothing and edge detection by nonlinear diffusion*, SIAM. J. Numer. Anal., 29 (1992), pp. 182–193.
[17] A. Chambolle and P.-L. Lions, *Image recovery via total variation minimization and related problems*, Numer. Math., 76 (1997), pp. 167–188.
[18] T. Chan, A. Marquina, and P. Mulet, *High-order total variation-based image restoration*, SIAM J. Sci. Comput., 22 (2000), pp. 503–516.
[19] P. Colella, A. Majda, and V. Roytburd, *Theoretical and numerical structure for reacting shock waves*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 1059–1080.
[20] P. Constantin, P. D. Lax, and A. Majda, *A simple one dimensional model for the three dimensional vorticity equation*, Comm. Pure Appl. Math., 38 (1985), pp. 715–724.
[21] S. Esedoglu, *An analysis of the Perona-Malik scheme*, Comm. Pure Appl. Math., 54 (2001), pp. 1442–1487.
[22] J. Goodman, *Stability of viscous scalar shock fronts in several dimensions*, Trans. Amer. Math. Soc., 311 (1989), pp. 683–695.
[23] J. Goodman, A. Kurganov, and P. Rosenau, *Breakdown in Burgers-type equations with saturating dissipation fluxes*, Nonlinearity, 12 (1999), pp. 247–268.
[24] J. B. Greer and A. L. Bertozzi, *H^1 solutions of a class of fourth order nonlinear equations for image processing*, Discrete Contin. Dynam. Systems, 10 (2004), pp. 349–366.
[25] L. Kadanoff, *Singularities and blowups*, Phys. Today, 50 (1997), pp. 11–12.
[26] B. Kawohl and N. Kutev, *Maximum and comparison principle for one-dimensional anisotropic diffusion*, Math. Ann., 311 (1998), pp. 107–123.
[27] S. Kichenassamy, *The Perona–Malik paradox*, SIAM J. Appl. Math., 57 (1997), pp. 1328–1342.
[28] J. J. Koenderink, *The structure of images*, Biol. Cybernet., 50 (1984), pp. 363–370.
[29] A. Kurganov, D. Levy, and P. Rosenau, *On Burgers-type equations with nonmonotonic dissipative fluxes*, Comm. Pure Appl. Math., 51 (1998), pp. 443–473.
[30] A. Kurganov and P. Rosenau, *Effects of a saturating dissipation in Burgers-type equations*, Comm. Pure Appl. Math., 50 (1997), pp. 753–771.
[31] P. D. Lax, *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 11, SIAM, Philadelphia, 1973.
[32] M. Lysaker, A. Lundervold, and X.-C. Tai, *Noise removal using fourth-order partial differential equations with applications to medical magnetic resonance images in space and time*, IEEE Trans. Image Process., 12 (2003), pp. 1579–1590.
[33] M. Lysaker, S. Osher, and X.-C. Tai, *Noise removal using smoothed normals and surface fitting*, IEEE Trans. Image Process., to appear.
[34] D. Marr and E. Hildreth, *Theory of edge detection*, Pro. Roy. Soc. London Ser. B, 207 (1980), pp. 187–217.
[35] S. Osher and J. Ralston, *L^1 stability of travelling waves with applications to convective porous media flow*, Comm. Pure Appl. Math., 35 (1982), pp. 737–749.
[36] P. Perona and J. Malik, *Scale-space and edge detection using anisotropic diffusion*, IEEE Trans. Pattern Anal. Machine Intell., 12 (1990), pp. 629–639.
[37] M. Renardy, *A singularly perturbed problem related to surfactant spreading on thin films*, Nonlinear Anal., 27 (1996), pp. 287–296.
[38] L. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D, 60 (1992), pp. 259–268.
[39] D. C. Sarocka and A. J. Bernoff, *An intrinsic equation of interfacial motion for the solid-*

ification of a pure hypercooled melt, Phys. D, 85 (1995), pp. 348–374.
[40] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, 2nd ed., Grundlehren Math. Wiss. 258 Springer-Verlag, New York, 1994.
[41] M. E. Taylor, *Partial Differential Equations.* I. Basic Theory, Appl. Math. Sci. 115, Springer-Verlag, New York, 1996.
[42] S. M. Troian, E. Herbolzheimer, S. A. Safran, and J. F. Joanny, *Fingering instabilities of driven spreading films*, Europhys. Lett., 10 (1989), pp. 25–30.
[43] J. Tumblin and G. Turk, *LCIS: A boundary hierarchy for detail-preserving contrast reduction*, in Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, ACM Press/Addison-Wesley, New York, 1999, pp. 83–90.
[44] G. W. Wei, *Generalized Perona-Malik equation for image processing*, IEEE Signal Processing Letters, 6 (1999), pp. 165–167.
[45] J. Weickert, B. M. ter Haar Romeny, and M. A. Viergever, *Efficient and reliable schemes for nonlinear diffusion filtering*, IEEE Trans. Image Process., 7 (1998), pp. 398–410.
[46] R. Whitaker and S. Pizer, *A multi-scale approach to nonuniform diffusion*, CVGIP: Image Understanding, 57 (1993), pp. 99–110.
[47] T. P. Witelski, *Shocks in nonlinear diffusion*, Appl. Math. Lett., 8 (1995), pp. 27–32.
[48] T. P. Witelski, *The structure of internal layers for unstable nonlinear diffusion equations*, Stud. Appl. Math., 97 (1996), pp. 277–300.
[49] T. P. Witelski, D. G. Schaeffer, and M. Shearer, *A discrete model for an ill-posed nonlinear parabolic PDE*, Phys. D, 160 (2001), pp. 189–221.
[50] A. Witkin, *Scale-space filtering*, in Proceedings of the International Joint Conference on Artificial Intelligence, Karlsruhe, Germany, 1983, pp. 1019–1021.
[51] Y.-L. You and M. Kaveh, *Fourth-order partial differential equations for noise removal*, IEEE Trans. Image Process., 9 (2000), pp. 1723–1730.

SIAM J. MATH. ANAL.
Vol. 36, No. 1, pp. 69–93

THE SUBCRITICAL MOTION OF A SEMISUBMERGED BODY: SOLVABILITY OF THE FREE BOUNDARY PROBLEM*

CARLO D. PAGANI† AND DARIO PIEROTTI†

Abstract. We discuss existence and regularity of the solutions of the wave-resistance problem for a thin semisubmerged body moving at uniform subcritical velocity in a heavy fluid (e.g., water) of constant depth. The main assumption (on the geometry of the body) is that the flow is two-dimensional; i.e., it can be completely described in the vertical plane containing the direction of the motion. Then the problem can be formulated in terms of a boundary value problem for a holomorphic function (the complex velocity field) satisfying a nonlinear condition (the Bernoulli condition) on a free boundary (the free surface of the fluid). By a hodograph transformation and choosing an appropriate functional setting, we first reduce the problem to the resolution of a nonlinear functional equation depending on two unknown parameters, which are related to the positions in the hodograph plane of the points of contact between the free surface and the body. The main result of this paper is the proof of the existence, under mild assumptions on the body's profile, of an exact solution of the nonlinear problem: the resulting free surface is asymptotically flat at infinity upstream and is oscillating downstream; moreover, it is tangent to the body's profile at the contact points.

Key words. free boundary, nonlinear boundary condition, hodograph transformation

AMS subject classifications. 35J65, 35R35, 76B10

DOI. 10.1137/S0036141003425982

1. Introduction and statement of the problem.

Let us consider an infinitely long, semisubmerged horizontal cylinder, moving at a uniform speed on the free surface of a heavy fluid, in the direction orthogonal to its generators. The unperturbed fluid, which is at rest, has finite constant depth H. Compressibility and viscosity are neglected as well as surface tension; moreover, the fluid motion is assumed to be irrotational.

We want to find the steady flow generated by the cylinder's motion. Because of the geometry of the problem, the flow can be completely described in the vertical plane containing the direction of the motion. Then the problem can be formulated in terms of a boundary value problem for a holomorphic function (the complex velocity field) satisfying a nonlinear condition (the Bernoulli condition) on a free boundary (the free surface of the fluid); moreover, the free boundary is the union of two disconnected curves ending on the cylinder's profile at unknown points (see Figure 1).

The solvability of this problem was established in [1] (for a cylinder with symmetric cross section) and in [2] (for a generic cylinder) in the case of *supercritical velocity* (see below). The proof relies on the assumption that the piercing part of the cylinder is small compared to its length (and to the fluid's depth) and essentially consists in the application of the implicit function theorem to a functional equation in the hodograph plane. In this approach, a crucial step is the proof of the unique solvability of a *linear problem*, which is obtained by considering the limit when the cylinder's section becomes a beam and the flow (in a reference system connected with the cylinder) approaches the constant, parallel flow [3].

*Received by the editors April 15, 2003; accepted for publication (in revised form) December 5, 2003; published electronically June 22, 2004.
http://www.siam.org/journals/sima/36-1/42598.html
†Dipartimento di Matematica del Politecnico, Piazza L. da Vinci 32, 20133 Milano, Italy (carpag@mate.polimi.it, darpie@mate.polimi.it).

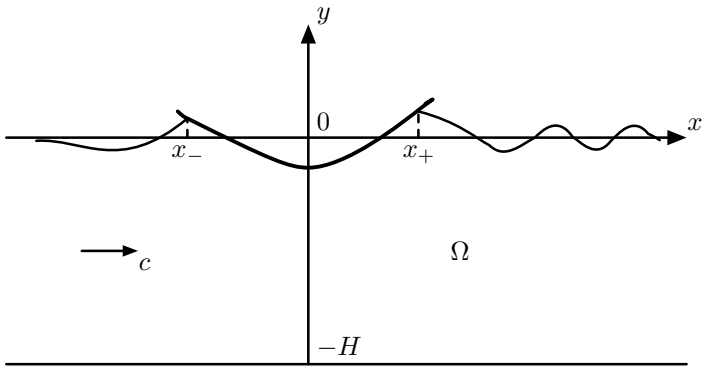

FIG. 1.

In the present work, we prove the solvability of the free boundary problem in the case of *subcritical velocities*. As it is already clear at the level of the linearized problem [4] and from numerical experiments [5], the properties of the solutions are quite different for subcritical and supercritical flows. For example, in the former case the flow may have nontrivial oscillations at infinity downstream, while in the latter case it is asymptotically parallel. This behavior is of course related to the situation that arises in the free-surface water waves: there, a supercritical flow is associated with solitary waves (that exponentially decay at infinity in both directions; see, e.g., [6], [7]), while subcritical flows develop a periodic wave train [7], [8, Chap. 71]. As a consequence, the former proof of the solvability will not extend in a trivial way to the subcritical flow. Nevertheless, we can still formulate the problem in terms of a functional equation in the hodograph plane with the same assumptions on the geometry of the cylinder. As in the case of supercritical velocities, we seek a "local" result of existence for a solution which, for some small parameter ϵ tending to zero, approaches the constant parallel flow. To reach this goal, in contrast with the supercritical case, *we will not fix a priori the asymptotic velocity of the perturbed field at infinity upstream*; the solution that we obtain will be a perturbation of the constant flow with prescribed subcritical velocity

$$c_0 < \sqrt{gH} \tag{1.1}$$

(here g is the acceleration of gravity); the perturbed flow will be parallel at upstream infinity, but its velocity c will depend on the parameter ϵ and will approach the unperturbed velocity c_0 as $\epsilon \to 0$. Similarly, the origin in the hodograph plane will not be completely fixed a priori, but we let it depend on a parameter changing with ϵ; we prescribe only its value in the limit $\epsilon = 0$ when the hodograph map is linear (see section 2). Both these quantities will be determined, as functions of ϵ, from the resolution of the problem, together with the free surface and the velocity field. The necessity of considering additional unknowns comes from the requirement to satisfy two *nonresonance conditions* at infinity downstream, where the velocity field is oscillating. More precisely, one finds that in the linearized problem (see below) the *wave number* and the *phase* of the oscillating far field are directly related to the above parameters; hence, by suitably setting their values, we can search the perturbed solutions in a common space of functions oscillating with the same wave number at downstream infinity.

In order to state the various equations of the problem, we choose a coordinate system connected with the cylinder and such that the xy-plane is orthogonal to the

horizontal generators of the cylinder; the x-axis is directed as the unperturbed flow, the undisturbed free surface is at $y = 0$, and the bottom of the region occupied by the fluid is at $y = -H$. The cross section of the "hull" is described by the equation

$$y = \epsilon f(x), \tag{1.2}$$

where $\epsilon > 0$ is a small parameter and f is a $\mathcal{C}^1$ function defined in some neighborhood of the origin, say J, and such that, for some other neighborhood of the origin $J' = (a, b) \subset J$, we have

$$\begin{aligned} &f(x) < 0 \qquad \text{for} \quad x \in J', \\ &f(x) = 0 \qquad \text{for} \quad x = a \quad \text{and} \quad x = b, \\ &f(x) > 0 \qquad \text{for} \quad x \in J \backslash \bar{J}', \\ &xf'(x) > 0 \qquad \text{for} \quad x \in J \backslash \{0\}, \\ &f'(0) = 0. \end{aligned} \tag{1.3}$$

The fluid surface is described by the equation $y = h(x)$, where h is an unknown smooth function defined in $\mathbf{R}\backslash[x_-, x_+]$, with $x_\pm \in J$. The two numbers $x_\pm$ are the abscissae of the points where the free surface meets the hull so that $h(x_\pm) = \epsilon f(x_\pm)$. Note that the values $x_\pm$ are unknown, and their determination is part of the problem. It is natural to assume that x_- and x_+ lie in small neighborhoods of the points a and b, respectively, which are bounded away from the origin.

We set

$$h^*(x) = \begin{cases} h(x) & \text{for } x \le x_-,\ x \ge x_+ \\ \epsilon f(x) & \text{for } x_- \le x \le x_+. \end{cases} \tag{1.4}$$

Then

$$S^* = \{(x, y) \in \mathbf{R}^2 \,:\, -H < y < h^*(x)\} \tag{1.5}$$

will denote the region filled with the fluid. We assume (as usual) that the curve $y = h^*(x)$ is a streamline; i.e., the free surface and the wetted part of the cylinder form a single streamline; the bottom $\{y = -H\}$ is also assumed to be a streamline. Let us introduce the complex variable $z = x + iy$ and the complex velocity function $\omega(z) = u(x, y) - iv(x, y)$, holomorphic in S^*, with u and v components of the velocity vector. We can now state our problem in the following form: find three scalars, the asymptotic velocity c, the abscissae $x_+ > 0$ and $x_- < 0$, and a real function $h \in \mathcal{C}^1(\mathbf{R}\backslash[x_-, x_+])$ and a complex function $\omega = u - iv$ holomorphic in S^* and bounded in $\overline{S^*}$, such that the following boundary conditions hold:

$$\frac{1}{2}|\omega(x, h(x))|^2 + gh(x) = \text{constant}, \qquad x < x_- \text{ or } x > x_+, \tag{1.6}$$

$$v(x, h(x)) = h'(x)u(x, h(x)), \qquad x < x_- \text{ or } x > x_+, \tag{1.7}$$

$$v(x, \epsilon f(x)) = \epsilon f'(x)u(x, \epsilon f(x)), \qquad x_- \le x \le x_+, \tag{1.8}$$

$$v(x, -H) = 0, \qquad x \in \mathbf{R}, \tag{1.9}$$

$$\lim_{x \to -\infty} \omega(z) = c, \tag{1.10}$$

$$\lim_{x \to -\infty} h(x) = 0. \tag{1.11}$$

Equations (1.7), (1.8) indicate that the free surface and the wetted hull are arcs of a streamline; (1.9) expresses the same property for the bottom, while (1.6) is the

Bernoulli condition on the free surface. The asymptotic conditions (1.10), (1.11) state that at infinity upstream the flow approaches a constant parallel flow, and the free boundary approaches the straight line $h(x) = 0$. We stress that the asymptotic velocity c is an unknown function of ϵ which tends to c_0 as $\epsilon \to 0$. As discussed above, the perturbation due to the presence of the cylinder does not in general vanish (in two dimensions) at infinity downstream if (1.1) holds. The statement of the problem is completed by the *continuity conditions*

$$h(x_\pm) = \epsilon f(x_\pm). \tag{1.12}$$

Rigorous mathematical results about nonlinear ship waves are quite rare in the literature; the problem appears in a linearized version (the Neumann–Kelvin problem; see, e.g., [4, Part 2] and references cited therein), or it has been treated by numerical methods [5]. Some authors [9], [10] (see also [7]) consider the water waves problem by assuming a variable pressure of the form: p_0 (atmospheric pressure) $+\epsilon p(x)$ acting on the free surface; if $p(x)$ is compactly supported, this extra pressure may simulate the action of a ship.

The aim of this paper is to prove the existence, for small values of the parameter ϵ, of an exact solution of the nonlinear problem, which, for $\epsilon \to 0$, reduces to the trivial parallel flow $\omega = c_0$, $h = 0$. The main steps in implementing this program are the following: in the next section, we use a hodograph transformation which (partially) overcomes the difficulties due to the free boundary; the transformed problem proves to be convenient for a functional reformulation. The proof of solvability is achieved in two steps: see sections 3 and 4; in particular, in section 3 we exploit the results obtained in [3] for the linearized problem. Some technical results and side properties of the solution are described in the appendix.

The main result of the paper (the precise statement is Theorem 5.6) is that, for a given profile $\epsilon f(x)$ with $\epsilon > 0$ small, f satisfying (1.3), and some additional technical conditions (also involving the data c_0 and H) there is a solution of the system (1.6)–(1.12). More specifically, there is a flow $\omega_\epsilon(z)$ which is asymptotically parallel when $x \to -\infty$; the asymptotic velocity c is a known quantity depending on ϵ and tending to c_0 as $\epsilon \to 0$. The free surface and the cylinder profile form a single $\mathcal{C}^1$ streamline: they match at known points x_- and x_+ (depending on ϵ). Moreover, the free surface is exponentially vanishing for $x \to -\infty$ and is bounded and asymptotically periodic when $x \to +\infty$; the period is also a known function of ϵ. This result qualitatively agrees with the numerical experiments presented in [5, Par. 3] for Froude numbers approximately ranging from 0.35 to 0.6 and for a parabolic profile. Also, the analysis developed in [9] (where we still have a localized obstacle on the bottom and a localized extra pressure on the free surface) shows that, for Froude numbers strictly less than 1, all bounded solutions are asymptotically periodic at infinity downstream.

2. The hodograph transformation. By means of the hodograph transformation (see [2] for details) we can reformulate the problem by taking the complex potential

$$w = \varphi + i\psi \tag{2.1}$$

(where $\varphi(x, y)$ is the velocity potential and $\psi(x, y)$ the stream function) as the independent variable and the reciprocal of the velocity field

$$\frac{1}{\omega(z)} = \Omega(w), \qquad \Omega = U - iV, \tag{2.2}$$

as the unknown. Given the function Ω, the (inverse) hodograph map $w \mapsto z = x + iy$ is defined by the relation $dz/dw = \Omega$ modulo an additive complex constant; the imaginary part of this constant can be fixed in such a way that the streamline consisting of the free surface and the wetted hull corresponds to $\psi = 0$ (see (2.5)). Then the domain S^* of the physical plane is mapped onto the strip

$$A_H \equiv \{(\varphi, \psi) \in \mathbf{R}^2 : \quad -cH < \psi < 0\}. \tag{2.3}$$

The real part of the additive constant is left undetermined for the moment and is assumed to change with ϵ. For $\epsilon = 0$ we assume that the image of the point $w = 0$ coincides with the origin in the physical plane, which can be placed at a minimum point of the function f on the x-axis; for $\epsilon > 0$, the origin of the hodograph plane will be mapped (for small enough ϵ) to an unknown point $(\bar{x}, \epsilon f(\bar{x}))$ on the cylinder's profile according to the discussion of the introduction. Taking account of the above conditions, the relation between the physical plane variables and the hodograph plane ones can be written

$$x(\varphi, \psi) = \bar{x} + \int_0^{\varphi} U(s, \psi) ds + \int_0^{\psi} V(0, t) dt, \tag{2.4}$$

$$y(\varphi, \psi) = \int_{-cH}^{\psi} U(\varphi, t) dt - H = \frac{1}{c}\psi - \int_{-\infty}^{\varphi} V(s, \psi) ds. \tag{2.5}$$

We stress that the functions U, V and the parameters $\bar{x}$, c in (2.4), (2.5) depend on ϵ; for $\epsilon = 0$, we have $U = 1/c_0$, $V = 0$, $\bar{x} = 0$, and $c = c_0$, and the map is simply the multiplication by $1/c_0$.

We call φ_- and φ_+ the values of φ at the separating points $P_- = (x_-, \epsilon f(x_-))$ and $P_+ = (x_+, \epsilon f(x_+))$, respectively. Then the upper boundary of the strip consists of the segment

$$I = \{(\varphi, \psi) : \quad \psi = 0, \quad \varphi_- < \varphi < \varphi_+\}, \tag{2.6}$$

which is the image of the cylinder's hull, and the two half-lines

$$F = \{(\varphi, \psi) : \quad \psi = 0, \quad \varphi < \varphi_-\} \cup \{(\varphi, \psi) : \quad \psi = 0, \quad \varphi > \varphi_+\}, \tag{2.7}$$

which are the image of the free surface; we stress that the separating abscissae $\varphi_- < 0$ and $\varphi_+ > 0$ are also unknown. The bottom is mapped onto the line

$$B = \{(\varphi, \psi) : \quad \psi = -cH, \quad \varphi \in \mathbf{R}\}. \tag{2.8}$$

Then the function Ω must be holomorphic in A_H and satisfy the boundary conditions

$$-\frac{1}{2}\frac{\partial |\Omega|^{-2}}{\partial \varphi} + gV = 0 \qquad \text{on } F, \tag{2.9}$$

$$V + \epsilon f'(x) U = 0 \qquad \text{on } I, \tag{2.10}$$

$$V = 0 \qquad \text{on } B. \tag{2.11}$$

Moreover, we require the condition at infinity upstream

$$\lim_{\varphi \to -\infty} \Omega = \frac{1}{c}. \tag{2.12}$$

The continuity conditions (1.12) are written

$$-\int_{-\infty}^{\varphi_\pm} V(s,0)ds = \epsilon f\left(\bar{x} + \int_0^{\varphi_\pm} U(s,0)ds\right). \tag{2.13}$$

As already noticed in [2] these two conditions, now written in the hodograph plane, are *not* independent if (2.10) holds. We now show, however, that there is another independent condition at the point $(\varphi_+, 0)$, related to the Bernoulli equation (1.6). In fact, for physical reasons and recalling the asymptotic conditions (1.10), (1.11), the constant appearing on the right-hand side of (1.6) must have the same value $c^2/2$ on both components of the free surface; this holds in particular at the two points $P_\pm$. In terms of the hodograph variables, we get by (2.5), (2.9), and the limit condition (2.12)

$$\frac{1}{2}\,|\,\Omega(\varphi,0)\,|^{-2} + g\,y(\varphi,0) = \frac{c^2}{2}$$

for $\varphi \le \varphi_-$ (and $\psi = 0$); by (2.2), this is equivalent to (1.6). On the other hand, there is no prescribed limit at infinity downstream (only boundedness of the flow field is required). This means that we have the *additional condition*

$$|c\,\Omega(\varphi_+,0)|^{-2} + \frac{2g}{c^2}y(\varphi_+,0) = 1,$$

which, taking account of (2.13), becomes

$$|c\,\Omega(\varphi_+,0)|^{-2} + \frac{2g}{c^2}\epsilon f\left(\bar{x} + \int_0^{\varphi_+} U(s,0)ds\right) = 1. \tag{2.13$'$}$$

Equations (2.9)–(2.13$'$) formulate the problem in the hodograph plane; by the previous discussion, we could replace (2.13) with the analogous of (2.13$'$) at φ_-.

We stress that in the above problem the size and the position of the segment I defined by (2.6) are unknown (the same is true for the depth of the bottom B of the strip in the hodograph plane; see (2.8)). Therefore, a further change of the independent variables and unknowns will prove convenient in the following. Let us first introduce the new parameters

$$\varphi^* = \frac{\varphi_+ - \varphi_-}{2c}, \qquad \varphi^m = \frac{\varphi_+ + \varphi_-}{2c}. \tag{2.14}$$

Then, by setting

$$\rho = \frac{\varphi - c\varphi^m}{c\varphi^*}, \qquad \sigma = \frac{\psi}{c\varphi^*}, \tag{2.15}$$

the beam I is mapped onto the interval $(-1,1)$ of the ρ-axis, and the strip A_H becomes

$$A^* = \{(\rho,\sigma) \in \mathbf{R}^2 \,:\quad -H^* < \sigma < 0\}, \tag{2.16}$$

where

$$H^* = \frac{H}{\varphi^*}. \tag{2.17}$$

Then we define the new unknown

$$\chi = \xi - i\eta$$

(as a function of the new variables (2.15)) by subtracting the asymptotic field $1/c$ from Ω and dividing by ϵ; namely, we set

$$U(\varphi,\psi) = \frac{1}{c}\Big(1+\epsilon\xi(\rho,\sigma)\Big), \qquad V(\varphi,\psi) = \frac{\epsilon}{c}\eta(\rho,\sigma). \tag{2.18}$$

We now want to write the nonlinear boundary conditions (2.9), (2.10) as formal operator equations in the new variables. We first note that on the line $\sigma = 0$ (that is, $\psi = 0$) the right-hand side of (2.4) takes the form

$$\bar{x} + \varphi^* \int_{\rho^m}^{\rho} (1+\epsilon\xi(s,0))ds \equiv x(\rho), \tag{2.19}$$

where we set $\rho^m = -\varphi^m/\varphi^*$; note that (2.19) is independent of c.

We now define the functions

$$G(\rho) = f'(x(\rho)) \tag{2.20}$$

and

$$B^I(\chi,\bar{x};\varphi^m,\varphi^*,\epsilon) = \Big\{\eta + G(\,\cdot\,)(1+\epsilon\xi)\Big\}\Big|_{|\rho|<1,\,\sigma=0}. \tag{2.21}$$

Furthermore, by introducing the parameter

$$\nu^* = \varphi^*\nu = \varphi^*\frac{g}{c^2}, \tag{2.22}$$

we define

$$B^F(\chi,\nu^*;\epsilon) = \Big\{ -\frac{1}{2\epsilon}\frac{\partial}{\partial\rho}|1+\epsilon\chi|^{-2} + \nu^*\eta\Big\}\Big|_{|\rho|>1,\,\sigma=0} \tag{2.23}$$

and

$$\mathbf{B}(\chi,\bar{x},\nu^*;\varphi^m,\varphi^*,\epsilon) = (B^I(\chi,\bar{x};\varphi^m,\varphi^*,\epsilon), B^F(\chi,\nu^*;\epsilon)). \tag{2.24}$$

Then, for every $\epsilon > 0$, the equation

$$\mathbf{B}(\chi,\bar{x},\nu^*;\varphi^m,\varphi^*,\epsilon) = 0 \tag{2.25}$$

is equivalent to the conditions (2.9), (2.10). Moreover, the function χ must be holomorphic in A^*, vanishing for $\rho \to -\infty$, and satisfying the *linear* condition $\eta(\rho,-H^*) = 0$. The notation used stresses the dependence of the differential system on the various parameters of the problem: $\bar{x}$, ν^* (and then c), φ^m, φ^*; such quantities are unknown functions of ϵ as well as the field χ. Our strategy for solving the problem in the hodograph plane will consists of two steps: first, *we fix φ^* and φ^m independent of* ϵ and solve (via the implicit function theorem) (2.25) with respect to χ, ν^*, and $\bar{x}$ for small ϵ, starting with the solution of a linear problem at $\epsilon = 0$ (see section 3.1 below). The values of ν^* and $\bar{x}$ for $\epsilon > 0$ will be determined by requiring the solution χ to belong to an appropriate Banach space (see section 3.2). Thus, in this way, we determine a family of hodograph maps depending on the parameters φ^*, φ^m, and ϵ

(note that also the strip A^* where the function χ is defined depends on φ^*; see (2.16), (2.17)). Then, in the second step, we select a pair $\varphi^* = \varphi^*(\epsilon)$, $\varphi^m = \varphi^m(\epsilon)$ by solving (2.13), (2.13′): the selected map, corresponding to those values of the parameters, will finally determine the solution of the problem in the physical plane. The reason for this two-step procedure is that we do not know a priori the limit positions for $\epsilon \to 0$ of the points $P_\pm$ so that we cannot linearize the whole problem around a known solution at $\epsilon = 0$.

In the next section, we will formulate (2.25) as an operator equation between suitable Banach spaces; this equation will be solved, for fixed φ^m, φ^*, in section 4, while in section 5 we discuss (2.13), (2.13′) determining φ^m, φ^*.

3. The functional setting of the problem.

3.1. The problem at $\epsilon = 0$ in the hodograph plane. According to the previous discussion, we fix the two parameters $\varphi^* > 0$ and φ^m and discuss the linear problem obtained from (2.25) by letting $\epsilon \to 0$ (formally) in the expressions (2.21), (2.23); the results obtained will suggest the correct functional setting of the nonlinear problem.

We first recall that $c \to c_0$ and $\bar{x} \to 0$ as $\epsilon \to 0$; hence, by recalling (2.22) we also have $\nu^* \to \nu_0^*$, where

$$\nu_0^* = \varphi^* g / c_0^2. \tag{3.1}$$

Then, for $\epsilon \to 0$ the system (2.25), together with the condition on the bottom and the asymptotic condition, leads to the following problem for a holomorphic function $\chi_0 = \xi_0 - i\eta_0$ in the domain A^* (see [2]):

$$\partial_\rho \xi_0 + \nu_0^* \eta_0 = 0 \qquad \text{for} \quad \sigma = 0, \ \ |\rho| > 1,$$

$$\eta_0(\rho, 0) = -f'(\varphi^* \rho + \varphi^m) \qquad \text{for} \quad |\rho| < 1,$$

$$\eta_0 = 0 \qquad \text{for} \quad \sigma = -H^*, \ \ \rho \in \mathbf{R},$$

$$\lim_{\rho \to -\infty} \chi_0 = 0.$$

By substituting, in the first equation, $\partial_\rho \xi_0$ with $-\partial_\sigma \eta_0$, we obtain a boundary value problem for the harmonic function η_0 (the harmonic conjugate ξ_0 is then determined by the requirement of vanishing at infinity upstream).

Problem $\mathbf{L_0}$. Find η_0 harmonic in A^* such that

$$\partial_\sigma \eta_0(\rho, 0) - \nu_0^* \, \eta_0(\rho, 0) = 0 \qquad \text{for} \quad |\rho| > 1, \tag{3.2}$$

$$\eta_0(\rho, 0) = -f'(\varphi^* \rho + \varphi^m) \qquad \text{for} \quad |\rho| < 1, \tag{3.3}$$

$$\eta_0(\rho, -H^*) = 0 \qquad \text{for} \quad \rho \in \mathbf{R}, \tag{3.4}$$

$$\lim_{\rho \to -\infty} \eta_0(\rho, \cdot) = 0. \tag{3.5}$$

By adding to (3.2)–(3.5) the natural requirement that the solution is H^1_{loc} and bounded (more generally, polynomially bounded) in the strip outside any neighborhood of the interval $[-1, 1] \times \{0\}$, Problem $\mathbf{L_0}$ coincides with the problem obtained by formal linearization of the original nonlinear problem in the physical plane; see [3]. Then, by the results of [3], we have the following.

THEOREM 3.1. *Given $f' \in H^{1/2}(I)$ (with I defined by (2.6)), Problem* $\mathbf{L_0}$ *is uniquely solvable for $\nu_0^* H^* = gH/c_0^2 > 1$, provided the positive solution μ of*

$$\tanh(\mu H^*) = \frac{\mu}{\nu_0^*} \tag{3.6}$$

is different from $n\pi/2$, $n = 1, 2, \ldots$. Furthermore, if $f' \in H^{3/2}(I)$, the solution is continuous and bounded in the closed strip $\overline{A^*}$, and there are real constants A_0, B_0 such that

$$\sup_{(\rho,\sigma)\in A^*} e^{\lambda_0|\rho|}\Big|\eta_0(\rho,\sigma) - \theta(\rho)[A_0\sin(\mu\rho) + B_0\cos(\mu\rho)]\sinh(\mu(\sigma + H^*))\Big| < \infty, \tag{3.7}$$

where λ_0 is the first positive solution of the equation

$$\tan(\lambda H^*) = \frac{\lambda}{\nu_0^*} \tag{3.8}$$

and θ is the characteristic function of the interval $(0, +\infty)$.

From Theorem 3.1 we get no information on the solvability of Problem $\mathbf{L_0}$ at the "singular values" $\mu = n\pi/2$, $n = 1, 2, \ldots$. However, by a careful reconsideration of some arguments of [3], it can be shown that the solutions defined by Theorem 3.1 have well defined limits for $\mu \to n\pi/2$ and that these limits are still solutions to Problem $\mathbf{L_0}$ with the same regularity and asymptotic properties (3.7). In fact, we can now state the following.

THEOREM 3.2. *Let f be given as in Theorem* 3.1*, and suppose that the positive solution of* (3.6) *satisfies $\mu = n\pi/2$, $n = 1, 2, \ldots$. Then there is a unique solution η_0 of Problem* $\mathbf{L_0}$*, which is defined as the limit for $\mu \to n\pi/2$ of the solutions given by Theorem* 3.1*.*

The proof is given in the appendix.

Remark 3.3. It is worthwhile to point out further properties of the solution of Problem $\mathbf{L_0}$ which will be useful for the definition of an appropriate functional setting for (2.25). We stress that these properties, as well as the asymptotic representation (3.7), *hold for every positive value of μ.*

(i) It can be shown (see also [1]) that if the datum in (3.2) belongs to the Sobolev space $W_p^{2-\frac{1}{p}}(-1, 1)$, with $p \in (1, 4/3)$, then η_0 belongs to $W_p^2(B)$ for every bounded, measurable $B \subset A^*$. We recall the inclusion $W_p^2(B) \subset \mathcal{C}^{0,\alpha}(\bar{B})$, with $\alpha = 2 - 2/p$; also notice that the space W_p^2 is *an algebra* for $p > 1$ and that the product between functions of W_p^2 is continuous. Moreover, since the gradient of η_0 is locally integrable along any curve contained in the closed strip $\overline{A^*}$, the harmonic conjugate ξ_0 is continuous on $\overline{A^*}$ (and also in W_p^2 of any bounded subset).

(ii) The holomorphic function $\chi_0 = \xi_0 - \eta_0$ is everywhere bounded in $\overline{A^*}$ and smooth up to the boundary outside any neighborhood of the interval $[-1, 1] \times \{0\}$. Moreover, a bound similar to (3.7) also holds for the function $\partial_\rho \xi_0$.

(iii) For coefficients A_0 and B_0, in the representation (3.7), the following formulas hold (see Proposition A.4 of the appendix):

$$A_0 = \int_{-1}^{1} f'(\varphi^*\rho + \varphi^m)\alpha(\rho)d\rho, \tag{3.9}$$

$$B_0 = \int_{-1}^{1} f'(\varphi^*\rho + \varphi^m)\beta(\rho)d\rho, \tag{3.10}$$

where α, β are real continuous functions on $[-1,1]$ depending only on ν_0^* and H^*.

Remark 3.4. By Theorems 3.1 and 3.2, we have that *the linearized problem of the flow past a surface-piercing obstacle is uniquely solvable for every subcritical value of the velocity* c_0. This property was never proved before in the linear theory of wave-body interaction, [4] and will be discussed in a more general framework in a forthcoming paper.

Let us now go back to the representation (3.7); we note that the parameter ν_0^* determines the wave number of the perturbed flow at downstream infinity. Moreover, the choice $\bar{x} = 0$ at $\epsilon = 0$ affects the asymptotic phase of η_0. Actually, for $\rho \to +\infty$ we can write

$$\eta_0(\rho,\sigma) \approx C_0 \sin(\mu\rho + \delta_0)\sinh(\mu(\sigma + H^*)), \tag{3.11}$$

where $\delta_0 = \arctan(B_0/A_0)$; since A_0 and B_0 are linear functionals of the boundary datum of Problem $\mathbf{L_0}$, i.e., of $-f'(\varphi^*\rho + \varphi^m)$ for $|\rho| < 1$, a different choice of the limit value of $\bar{x}$ corresponds to a shift of the argument of f' and therefore to a change of δ_0. Now, it is clear that, given φ^*, φ^m, ν_0^*, and H^*, the constant δ_0 in (3.11) is fixed by the integrals on the right-hand sides of (3.9), (3.10) (we suppose that at least one of these integrals is not vanishing; if $A_0 = 0$ we take $\delta_0 = \pi/2$). In the following, for the sake of simplicity, we shall assume that the function $\rho \mapsto f'(\varphi^*\rho + \varphi^m)$ is orthogonal to β so that $B_0 = 0$ and $\delta_0 = 0$. In this case, the solution η_0 is *asymptotically odd* with respect to ρ for $\rho \to +\infty$. Then we will look for solutions of the nonlinear problem (2.25) with the same symmetry at infinity. We point out that the restriction to invariant subspaces of functions with definite symmetry is also a crucial step in the proof of the existence of periodic water waves by bifurcation methods [8], [11]. Later, we will show how to get rid of the previous orthogonality assumption; we remark only here that, in the general case, the function $\rho \mapsto \eta_0(\rho - \delta_0^*, \sigma)$ is asymptotically odd, where

$$\delta_0^* = \delta_0/\mu. \tag{3.12}$$

Remark 3.5. When A_0, B_0 are both vanishing, we get $C_0 = 0$ in (3.11), and the phase δ_0 is undetermined. In this case, we have a waveless solution of the linear Problem $\mathbf{L_0}$ (see [4], [14]), which is also uniquely determined.

3.2. The functional equation. It is now convenient to outline our strategy for solving (2.25); we want to solve such an equation for every pair φ^*, φ^m in a neighborhood of the previously discussed solution at $\epsilon = 0$. We remark that this solution is a function $\chi_0 = \xi_0 - i\eta_0$ holomorphic in the strip (2.16); moreover, χ_0 vanishes for $\rho \to -\infty$ and approaches, for $\rho \to +\infty$, a holomorphic function $\chi_0^\#$ which is $2\pi/\mu$ periodic with respect to ρ, where μ is the positive solution of (3.6). Finally, by the discussion at the end of the previous section, $\chi_0^\#$ satisfies the *symmetry condition* $\chi_0^\#(-\rho,\sigma) = \overline{\chi_0^\#(\rho,\sigma)}$.

Thus, it is natural to solve the functional equation (2.25) in a space of functions defined in the *fixed* strip A^* and with the above asymptotic properties. We note in particular that we look for solutions having *the same wave number and symmetry*, in the limit $\rho \to +\infty$, for every positive (small enough) ϵ; as we will see, this can be accomplished by letting the parameters ν^*, $\bar{x}$ vary from the initial values ν_0^*, 0. This means that we will solve the functional equation with respect to the unknowns $(\chi, \bar{x}, \nu^*)$ in a neighborhood of the solution $(\chi_0, 0, \nu_0^*)$. We first define suitable Banach spaces for

the discussion of the functional equation (2.25); then we will show that the operator $\mathbf{B}$ is a continuously differentiable map between these spaces. In the next section, we will prove the invertibility of the Frechet derivative $\mathbf{B}'$ at $\epsilon = 0$ and solve (2.25) by the implicit function theorem. Taking account of Theorem 3.1 and Remark 3.3, we now introduce a Banach space X of holomorphic functions defined in A^* and continuous up to the boundary. Let us fix $\rho_0 > 1$ and set $Q_{\rho_0} = [-H^*, 0] \times \mathbf{R}\backslash(-\rho_0, \rho_0)$; moreover, given $\bar\rho > \rho_0$, let $B_{\bar\rho} \subset A^*$ be the bounded rectangle $(-H^*, 0) \times (-\bar\rho, \bar\rho)$. Finally, take $1 < p < 4/3$, $\alpha = 2 - 2/p$, and define

$$X = \Big\{\chi = \xi - i\eta \in \mathrm{Hol}(A^*), \quad \chi|_{B_{\bar\rho}} \in W^2_p(B_{\bar\rho}),\ \chi|_{Q_{\rho_0}} \in \mathcal{C}^{1,\alpha}(Q_{\rho_0}),$$
$$\eta(\cdot, -H^*) = 0, \quad \lim_{\rho\to-\infty} \chi = 0, \quad \lim_{\rho\to+\infty} |\chi - \chi^\#| = 0, \Big\}, \tag{3.13}$$

where $\chi^\# = \xi^\# - i\eta^\#$ is holomorphic in the strip and $2\pi/\mu$-periodic with respect to ρ and such that $\chi^\#(-\rho, \sigma) = \overline{\chi^\#(\rho, \sigma)}$. The limits in (3.13) are uniform with respect to σ. The space X is endowed with the following norm:

$$||\chi||_X = ||\chi||_{\mathcal{C}^{1,\alpha}(Q_{\rho_0})} + ||\chi||_{W^2_p(B_{\bar\rho})}$$
$$+ \sup_{Q_{\rho_0}} e^{\lambda^*|\rho|}\{|\eta(\rho,\sigma) - \theta(\rho)\eta^\#(\rho,\sigma)| + |\partial_\rho\xi(\rho,\sigma) - \theta(\rho)\partial_\rho\xi^\#(\rho,\sigma)|\}, \tag{3.14}$$

where $0 < \lambda^* < \lambda_0$ and λ_0 is the lowest positive solution of (3.8). We note that X is a linear space of bounded, continuous functions up to the boundary of A^*; furthermore, X is complete with respect to the norm (3.14). In fact, if χ_n is a Cauchy sequence in X, we have in particular that χ_n converges uniformly on the closure of the strip A^* to a continuous function χ which is holomorphic in A^*; moreover, it can be shown that χ is the limit in X of the sequence. For, by (3.14), if $\chi_n \in X$ is a Cauchy sequence of functions asymptotic to the periodic functions $\chi^\#_n$, then the $\chi^\#_n$ form is also a Cauchy sequence in $\mathcal{C}^0([\varrho, \varrho + 2\pi/\mu] \times [-H^*, 0])$, with $\varrho \geq \rho_0$; hence, $\chi^\#_n \to \chi^\#$ uniformly, with $\chi^\#$ holomorphic and satisfying the properties described below (3.13). Now, by writing explicitly the Cauchy condition for $\chi_n - \chi_m$ and taking the limit for $m \to \infty$ at every point of the closed strip, we find that $\chi \in X$ with the above limit $\chi^\#$ in the definition (3.13); moreover, $\lim_{n\to\infty} \chi_n = \chi$ in X. Let us now define the space

$$Y = W_p^{2-\frac{1}{p}}(-1, 1) \times Y_{\rho_0,\bar\rho}\,, \tag{3.15}$$

where $Y_{\rho_0,\bar\rho}$ is the set of the real functions l defined (a.e.) in $\mathbf{R}\backslash[-1, 1]$ and with the following properties:

$$l|_{(-\bar\rho,-1)\cup(1,\bar\rho)} \in W_p^{1-\frac{1}{p}}((-\bar\rho, -1) \cup (1, \bar\rho)),$$
$$l|_{\mathbf{R}\backslash(-\rho_0,\rho_0)} \in \mathcal{C}^{0,\alpha}(\mathbf{R}\backslash(-\rho_0, \rho_0)),$$
$$\sup_{|\rho|\geq\rho_0} e^{\lambda^*|\rho|}|l(\rho) - \theta(\rho)l_\#(\rho)| < \infty, \tag{3.16}$$

where $l_\#$ is continuous, $2\pi/\mu$-periodic, and odd. The linear space $Y_{\rho_0,\bar\rho}$, equipped with the norm

$$||l||_Y = ||l||_{W_p^{1-\frac{1}{p}}((-\bar\rho,-1)\cup(1,\bar\rho))} + ||l||_{\mathcal{C}^{0,\alpha}(\mathbf{R}\backslash(-\rho_0,\rho_0))}$$
$$+ \sup_{|\rho|\geq\rho_0} e^{\lambda^*|\rho|}|l(\rho) - \theta(\rho)l_\#(\rho)|, \tag{3.17}$$

is a Banach space. The crucial result of this section is the following.

THEOREM 3.6. *Let f be a $\mathcal{C}^{3,1}$ function defined in an interval J containing the interval* (2.6)*. Then, for every bounded domain $\Phi \subset \mathbf{R}^2$ of the form $\varphi^* \in [a,b] \subset \mathbf{R}_+$, $\varphi^m \in [-\varphi^*, \varphi^*]$, there exists $\epsilon_0 > 0$ and a bounded open set $\mathcal{U} \subset X \times \mathbf{R}^2$ such that the operator*

$$\mathbf{B} : \mathcal{U} \times \Phi \times [0, \epsilon_0) \to Y,$$

defined by (2.24)*, is continuously differentiable with respect to $(\chi, \bar{x}, \nu^*)$; furthermore, the solution $(\chi_0, 0, \nu_0^*)$ of the linear problem*

$$\mathbf{B}(\chi, \bar{x}, \nu^*; \varphi^m, \varphi^*, 0) = 0$$

belongs to $\mathcal{U}$.

Proof. This proof follows along the same lines as the proof of Theorem 3.5 in [1] and of Theorem 3.2 in [2] and is only sketched here. We recall that $\mathbf{B}$ is a *family of operators*, acting on the variables $(\chi, \bar{x}, \nu^*) \in X \times \mathbf{R}^2$, depending on three parameters: φ^*, φ^m, and ϵ. We assume that ϵ belongs to some interval $[0, \epsilon_0)$ and the pair φ^*, φ^m to some bounded domain Φ as defined above. We prove the assertions of the theorem separately for the two components of the operator $\mathbf{B}$. We stress that the assumption $f \in \mathcal{C}^{3,1}(J)$ guarantees the continuity of the Nemitski operator associated with f'' from $W_p^{3-\frac{1}{p}}(-1,1)$ to $W_p^{2-\frac{1}{p}}(-1,1)$; furthermore, by Remark 3.3(i) it follows that $W_p^{2-\frac{1}{p}}(-1,1)$ is an algebra and that the product between functions in this space is continuous (see also [12, Theorem 1.4.4.2]). These properties allow us to conclude that, for suitably chosen $\mathcal{U}_I \subset X \times \mathbf{R}$ (containing the point $(\chi_0, 0)$) and $\epsilon_0 > 0$, the operator $\mathbf{B}^I$ given by (2.21) is continuous from $\mathcal{U}_I \times \Phi \times [0, \epsilon_0)$ into $W_p^{2-\frac{1}{p}}(-1,1)$ and continuously differentiable with respect to $(\chi, \bar{x})$; its G-differential at the point $(\chi_0, 0) \in \mathcal{U}_I$ is given by

$$\begin{aligned} &d_G\mathbf{B}^I(\chi_0, 0; \epsilon)[\chi, \bar{x}] \\ &= \eta(\rho, 0) + \epsilon f' \left(\varphi^* \int_{\rho^m}^{\rho} \big(1 + \epsilon\xi_0(s,0)\big) ds \right) \xi(\rho, 0) \\ &+ f'' \left(\varphi^* \int_{\rho^m}^{\rho} \big(1 + \epsilon\xi_0(s,0)\big) ds \right) \big(1 + \epsilon\xi_0(\rho, 0)\big) \left(\bar{x} + \epsilon\varphi^* \int_{\rho^m}^{\rho} \xi(s,0) ds \right) \\ &= \eta(\rho, 0) + \bar{x} f''(\varphi^* \rho + \varphi^m) + \mathcal{O}(\epsilon), \qquad |\rho| < 1. \end{aligned} \tag{3.18}$$

It is easy to check that the map $(\chi, \bar{x}; \varphi^m, \varphi^*, \epsilon) \mapsto d_G\mathbf{B}^I(\chi, \bar{x}; \varphi^m, \varphi^*, \epsilon)$ is continuous; then $\mathbf{B}^I$ is Frechet differentiable with the continuous derivative in $\mathcal{U}_I \times \Phi \times [0, \epsilon_0)$. Similarly, we can check differentiability of $\mathbf{B}^I$ with respect to the parameters φ^m, φ^*, and ϵ.

In order to exploit similar arguments for the second component of $\mathbf{B}$, i.e., $\mathbf{B}^F$ given by (2.23), it is convenient to write it in the form

$$\mathbf{B}^F(\chi, \nu^*; \epsilon) = \left\{ \nu^* \eta + \xi_\rho + \epsilon \frac{\partial}{\partial \rho} \frac{\frac{1}{2}(\eta^2 - 3\xi^2) - \epsilon\xi(\xi^2 + \eta^2)}{(1 + \epsilon\xi)^2 + \epsilon^2\eta^2} \right\} \Big|_{|\rho|>1,\, \sigma=0}. \tag{3.19}$$

Again by the properties of the spaces $W_p^{2-\frac{1}{p}}$ with $p > 1$ and recalling the definitions (3.16), (3.17), one can verify that, for suitably chosen $\mathcal{U}_F \subset X \times \mathbf{R}_+$ (containing the

point (χ_0, ν_0^*)) and $\epsilon_0 > 0$, the operator $\mathbf{B}^F$ acts continuously from $\mathcal{U}_F \times [0, \epsilon_0)$ into $Y_{\rho_0,\bar\rho}$; we particularly emphasize that if χ satisfies the asymptotic symmetry condition specified in the definition of the space X, then the right-hand side of (3.19) satisfies the analogous condition for the space $Y_{\rho_0,\bar\rho}$ (see (3.16)). The operator $\mathbf{B}^F$ is also differentiable with respect to (χ, ν^*), and its G-differential at the point (χ_0, ν_0^*) can be written

$$d_G\mathbf{B}^F(\chi_0, \nu_0^*; \epsilon)[\chi, \nu^*] = \nu^*\eta_0(\rho, 0) + \nu_0^*\eta(\rho, 0) + \xi_\rho(\rho, 0) + \mathcal{O}(\epsilon), \qquad |\rho| > 1. \tag{3.20}$$

The map $(\chi, \nu^*; \epsilon) \mapsto d_G\mathbf{B}^F(\chi, \nu^*; \epsilon)$ is continuous; then $\mathbf{B}^F$ is Frechet differentiable with the continuous derivative in $\mathcal{U}_F \times [0, \epsilon_0)$. We can also readily check the differentiability of $\mathbf{B}^F$ with respect to ϵ. By collecting all these facts, we get the proof of the theorem.

4. Solvability of the functional equation. In this section we solve (2.25) in a neighborhood of the solution at $\epsilon = 0$. To this aim, we prove the invertibility of the Frechet derivative $\mathbf{B}'(\chi_0, 0, \nu_0^*; \varphi^m, \varphi^*, 0)$; by Theorem 3.6 and evaluating (3.18), (3.20) at $\epsilon = 0$, we are led to consider the following boundary value problem.

Problem **L**. Find $\chi = \xi - i\eta \in X$ such that

$$\eta_\sigma(\rho, 0) - \nu_0^*\,\eta - \nu^*\,\eta_0(\rho, 0) = l(\rho) \qquad \text{for} \quad |\rho| > 1, \tag{4.1}$$

$$\eta(\rho, 0) + \bar x f''(\varphi^*\rho + \varphi^m) = k(\rho) \qquad \text{for} \quad |\rho| < 1, \tag{4.2}$$

$$\eta(\rho, -H^*) = 0, \tag{4.3}$$

where the pair (k, l) belongs to the space Y defined by (3.15), (3.16). We will show that problem (4.1)–(4.3) is uniquely solvable in the space X (see (3.13)) for a unique choice of the pair $(\bar x, \nu^*)$ in a neighborhood of $(0, \nu_0^*)$. We search a solution of the problem in the form

$$\eta = \eta_1 + \eta_2,$$

where η_1, η_2 are harmonic in the strip and satisfy, respectively, the conditions

$$\partial_\sigma\eta_1(\rho, 0) - \nu_0^*\eta_1 = 0 \qquad \text{for} \quad |\rho| > 1, \tag{4.4}$$

$$\eta_1(\rho, 0) = k(\rho) - \bar x f''(\varphi^*\rho + \varphi^m) - \eta_2(\rho, 0) \qquad \text{for} \quad |\rho| < 1, \tag{4.5}$$

$$\eta_1(\rho, -H^*) = 0 \qquad \text{for} \quad \rho \in \mathbf{R}, \tag{4.6}$$

$$\partial_\sigma\eta_2(\rho, 0) - \nu_0^*\,\eta_2 = l(\rho) + \nu^*\,\eta_0(\rho, 0) \qquad \text{for} \quad |\rho| > 1, \tag{4.7}$$

$$\eta_2(\rho, -H^*) = 0 \qquad \text{for} \quad \rho \in \mathbf{R}. \tag{4.8}$$

It is readily verified that if η_1, η_2 solve the system (4.4)–(4.8), their sum η solves Problem **L**; on the other hand, we know that (4.4)–(4.6) is solvable by Theorem 3.1. Now we will consider problem (4.7)–(4.8). By the definition of the space $Y_{\rho_0,\bar\rho}$ and by the continuation properties of Sobolev space functions [12, Par. 1.4.3], we can assume that the datum l on the right-hand side of (4.7) is defined on $\mathbf{R}$ and satisfies $l|_{(-\bar\rho,\bar\rho)} \in W_p^{1-\frac{1}{p}}(-\bar\rho, \bar\rho)$. Furthermore, let us define

$$l^*(\rho) = l(\rho) + \nu^*\,\eta_0(\rho, 0), \quad \rho \in \mathbf{R}. \tag{4.9}$$

We denote by Z the set of functions l^* defined in $\mathbf{R}$ such that

$$l^*|_{(-\bar\rho,\bar\rho)} \in W_p^{1-\frac{1}{p}}(-\bar\rho,\bar\rho) \quad \text{and have property (3.16) in } \mathbf{R}\backslash(-1,1).$$

We stress that a function $l^* \in Z$ may be decomposed as

$$l^*(\rho) = l_0^*(\rho) + \theta(\rho) l^\#(\rho), \tag{4.10}$$

where l_0^* is integrable and exponentially decaying for $|\rho| \to \infty$, while $l^\#$ is $2\pi/\mu$-periodic, continuously differentiable, and odd. We now discuss an auxiliary problem in the strip whose solution proves the existence of the required function η_2.

4.1. An auxiliary problem. Let us consider the following problem:

$$\Delta\Psi = 0 \qquad \text{in} \quad A^*, \tag{4.11}$$

$$\partial_\sigma \Psi(\rho,0) - \nu_0^*\, \Psi(\rho,0) = l^*(\rho) \qquad \text{for} \quad \rho \in \mathbf{R}, \tag{4.12}$$

$$\Psi(\rho,-H^*) = 0 \qquad \text{for} \quad \rho \in \mathbf{R}. \tag{4.13}$$

Moreover, we require that Ψ vanishes for $\rho \to -\infty$. Then, we have the following.

PROPOSITION 4.1. *For every $l^* \in Z$ there exists a function Ψ satisfying* (4.11)–(4.13) *and vanishing exponentially for $\rho \to -\infty$. Moreover, if l^* satisfies the linear condition*

$$\int_{-\pi/\mu}^{\pi/\mu} \sin(\mu\rho) l^\#(\rho) = 0, \tag{4.14}$$

then Ψ is bounded and asymptotically $2\pi/\mu$-periodic (with respect to ρ) for $\rho \to +\infty$.

Proof. Let us consider the convolution

$$(K \star l^*)(\rho,\sigma) = \int_{\mathbf{R}} K(\rho-\rho',\sigma) l^*(\rho')\, d\rho', \tag{4.15}$$

where

$$K(\rho,\sigma) = \frac{1}{2\pi} \int_{\mathbf{R}} e^{ip\rho} \frac{\sinh[p(\sigma+H^*)]}{p\cosh(pH^*) - \nu_0^* \sinh(pH^*)} dp. \tag{4.16}$$

We stress that the integrand in (4.16) has two simple poles at $p = \pm\mu$ on the real axis; therefore, the integral is understood as the Fourier transform of a tempered distribution, which can be evaluated by integrating along the path in the complex plane consisting of the intervals $(-\infty, -\mu-\epsilon)$, $(-\mu+\epsilon, \mu-\epsilon)$, $(\mu+\epsilon, +\infty)$ of the real axis and two semicircles of radius ϵ and center at $(\pm\mu, 0)$ surrounding the poles in the lower half plane. As a result, we obtain

$$K(\rho,\sigma) = \kappa(\rho,\sigma) + C\,\theta(\rho)\sin(\mu\rho)\sinh[\mu(\sigma+H^*)], \tag{4.17}$$

where C is a constant, θ is the characteristic function of the interval $(0,+\infty)$, and

$$\kappa(\rho,\sigma) = \sum_{n=0}^{\infty} c_n \sin[\lambda_n(\sigma+H^*)] e^{-\lambda_n|\rho|}.$$

The coefficients c_n are given by

$$c_n = [(1 - \nu_0^* H^*)\cos(\lambda_n H^*) - \lambda_n H^* \sin(\lambda_n H^*)]^{-1}, \quad n = 0, 1, 2 \ldots,$$

where λ_n are the positive solutions of the equation

$$\tan(\lambda H^*) = \frac{\lambda}{\nu_0^*}.$$

Note that $\lambda_n H^* \approx (n + 3/2)\pi$ for large n so that $|c_n| = \mathcal{O}(1/n)$. It follows in particular that $|\kappa(\rho, \sigma)| \leq C \log(|\rho|)$ for ρ in a neighborhood of the origin so that, for every $\sigma \in [-H^*, 0]$ and $p \geq 1$, the function $\rho \mapsto \kappa(\rho, \sigma)$ belongs to $L^p(\mathbf{R})$. Moreover, we have from (4.16) that $\Delta K = 0$ in A^* and that $K_\sigma(\rho, \sigma) - \nu_0^* K(\rho, \sigma) \to \delta(\rho)$ in $\mathcal{S}'(\mathbf{R})$ for $\sigma \to 0$; then (4.15) solves (4.11)–(4.13) at least for $l^* \in \mathcal{S}(\mathbf{R})$. We will now show that (4.15) is well defined also for $l^* \in Z$ and that the proposition holds with $\Psi = (K \star l^*)$.

Let us first consider the function $\kappa \star l^*$; by (4.10), we can write

$$(\kappa \star l^*)(\rho, \sigma) = (\kappa \star l_0^*)(\rho, \sigma) + \int_0^{+\infty} \kappa(\rho - \rho', \sigma) l^\#(\rho') d\rho'.$$

By explicit bounds using the integrability and the decay properties of the factors we get $|(\kappa \star l_0^*)(\rho, \sigma)| \leq Ce^{-\lambda^* |\rho|}$, where C depends on the norms of κ and l_0^* in $L^1(\mathbf{R})$. The second term is a bounded function defined in $\mathbf{R}$ and decreasing like $Ce^{\lambda_0 \rho}$ for large negative values of ρ (recall that $\lambda_0 > \lambda^*$); moreover, we have the identity

$$\int_0^{+\infty} \kappa(\rho - \rho', \sigma) l^\#(\rho') d\rho' = \int_{\mathbf{R}} \kappa(\rho', \sigma) l^\#(\rho - \rho') d\rho' - \int_\rho^{+\infty} \kappa(\rho', \sigma) l^\#(\rho - \rho') d\rho'. \tag{4.18}$$

Recalling that $l^\#$ is periodic and odd and observing that $\kappa(-\rho, \sigma) = \kappa(\rho, \sigma)$, we have that the first term on the right-hand side is periodic and odd with respect to ρ. The second term is also vanishing as $Ce^{-\lambda_0 \rho}$ for $\rho \to +\infty$. Let us now consider the convolution between l^* and the last term of (4.17); it is proportional to the function

$$\sinh[\mu(\sigma + H^*)] \int_{-\infty}^{\rho} \sin[\mu(\rho - \rho')] l^*(\rho') d\rho', \tag{4.19}$$

which is bounded by $Ce^{\lambda^* \rho}$ for $\rho \to -\infty$. To study the other limit, we write

$$\int_{-\infty}^{\rho} \sin[\mu(\rho - \rho')] l^*(\rho') d\rho' = [A(\rho) + P_c(\rho)] \sin(\mu\rho) + [B(\rho) + P_s(\rho)] \cos(\mu\rho), \tag{4.20}$$

where

$$A(\rho) = \int_{-\infty}^{\rho} l_0^*(\rho') \cos(\mu\rho') d\rho', \quad B(\rho) = -\int_{-\infty}^{\rho} l_0^*(\rho') \sin(\mu\rho') d\rho', \tag{4.21}$$

$$P_c(\rho) = \int_0^{\rho} \cos(\mu\rho') l^\#(\rho') d\rho', \quad P_s(\rho) = -\int_0^{\rho} \sin(\mu\rho') l^\#(\rho') d\rho'. \tag{4.22}$$

By expanding $l^\#$ in Fourier sine series and integrating term by term in (4.22), we find that the function P_s is bounded (and periodic) only if condition (4.14) holds. We point out that a second "nonresonance" condition for P_c is ruled out by the choice of

an odd $l^{\#}$. Then the right-hand side of (4.20) approaches a $2\pi/\mu$-periodic function for $\rho \to +\infty$; note that such a function is *not* odd, unless $\lim_{\rho\to+\infty} B(\rho) = 0$.

It remains to prove that (4.15) actually solves the problem. We point out that, by (4.10), the Fourier transform in $\mathcal{S}'$ of a function $l^* \in Z$ has the form

$$\hat{l}^*(p) = \hat{l}_0^*(p) - \sum_{n=1}^{\infty} b_n \hat{l}_n^*(p), \tag{4.23}$$

where $\hat{l}_0^*(p)$ is a smooth function, b_n are the coefficients of the Fourier series of $l^{\#}$, and $\hat{l}_n^*(p) = \lim_{\epsilon\to 0} n\mu/[(p-i\epsilon)^2 - (n\mu)^2]$ (the limit being in $\mathcal{S}'(\mathbf{R})$). Then we can define the one-parameter family of tempered distributions

$$\hat{\Psi}(p,\sigma) = \frac{\sinh[p(\sigma+H^*)]}{p\cosh(pH^*) - \nu_0^* \sinh(pH^*)} \hat{l}^*(p), \quad \sigma \in [-H, 0],$$

where the first factor on the right-hand side is regularized as in the discussion following (4.16). It is readily checked that

$$\lim_{\sigma\to 0} [\partial_\sigma \hat{\Psi}(p,\sigma) - \nu_0^* \hat{\Psi}(p,\sigma)] = \hat{l}^*(p)$$

in the distributional sense. Thus, by the properties of the Fourier transform $\mathcal{F}$ in $\mathcal{S}'$ (see [13], Thm. 7.15) we have that the inverse transform $\mathcal{F}^{-1}\hat{\Psi}$ solves (4.11)–(4.13) and is equal (a.e.) to the convolution (4.15). □

Remark 4.2. We remark that if (4.14) holds, the coefficient b_1 of the series on the right-hand side of (4.23) vanishes. As a consequence, in evaluating the inverse Fourier transform of $\hat{\Psi}$ by complex plane integration, we have only contributions of *simple poles*, which produce the oscillating terms of the solution at $+\infty$. If $b_1 \neq 0$, there are poles of order two at $p = \pm\mu$, generating a "resonance" term in Ψ, whose amplitude grows linearly for $\rho \to +\infty$, in agreement with the previous calculations (see (4.20) and (4.22)).

By taking l^* as in (4.9) we can now choose $\eta_2 = \Psi$ as the harmonic function satisfying (4.7) and (4.8). We now show that there is a unique value of ν^* such that (4.14) holds. For, by (4.9) and by the asymptotic properties of η_0, we have

$$l^{\#}(\rho) = l_{\#}(\rho) + \nu^* A_0 \sinh(\mu H^*) \sin(\mu\rho);$$

inserting in (4.14) we find

$$\nu^* = -\frac{1}{\pi A_0 \sinh(\mu H^*)} \int_{-\pi/\mu}^{\pi/\mu} l_{\#}(\rho) \sin(\mu\rho) d\rho. \tag{4.24}$$

4.2. Invertibility of the Frechet derivative. We can now state the main result of this section.

THEOREM 4.3. *Let $\nu_0^* > 1/H^*$, and assume that*

$$F(\varphi^m, \varphi^*) \equiv \int_{-1}^{1} f''(\varphi^*\rho + \varphi^m)\beta(\rho) d\rho \neq 0. \tag{4.25}$$

Then, for every pair $(k,l) \in Y$, there is a unique pair $(\bar{x}, \nu^*) \in \mathbf{R}^2$ such that Problem **L** *is uniquely solvable in X.*

Proof. We first note that, by Theorems 3.1 and 3.2, there is a unique harmonic function η_1 which satisfies (4.4)–(4.6) and the asymptotic condition (3.7). Then the function $\eta = \eta_1 + \eta_2$, with η_2 defined in section 4.1, satisfies (4.1)–(4.3) and vanishes (exponentially) at $-\infty$. Moreover, by choosing ν^* as in (4.24), we have that η is bounded and asymptotically $2\pi/\mu$-periodic for $\rho \to +\infty$. It remains to satisfy the symmetry condition. To this aim, we recall that by (4.18)–(4.22) the function η_2 approaches, for $\rho \to +\infty$, the sum of an odd (periodic) function with the function

$$B_+ \cos(\mu\rho) \sinh[\mu(\sigma + H^*)], \tag{4.26}$$

where B_+ is proportional to the limit for $\rho \to +\infty$ of the function $B(\rho)$ defined in (4.21). On the other hand, by denoting with k^* the right-hand side of (4.5), we have for large positive values of ρ

$$\eta_1 \approx [A_1 \sin(\mu\rho) + B_1 \cos(\mu\rho)] \sinh[\mu(\sigma + H^*)], \tag{4.27}$$

where, by recalling (3.10), $B_1 = \int_{-1}^{1} k^*(\rho)\beta(\rho)d\rho$.

Then η is asymptotically odd if

$$B_1 + B_+ = 0.$$

By the definition of k^* and by (4.25) the above equation is satisfied by choosing

$$\bar{x} = \frac{1}{F(\varphi^*, \varphi^m)} \left\{ B_+ + \int_{-1}^{1} [k(\rho) - \eta_2(\rho, 0)]\beta(\rho)d\rho \right\}. \tag{4.28}$$

It remains to prove uniqueness. Assume that the real numbers $\tilde{x}$ and $\tilde{\nu}^*$ and the function $\tilde{\eta}$ solve the *homogeneous* Problem **L**. By Proposition 4.1, there is a harmonic function $\tilde{\eta}_2$, vanishing for $\rho \to -\infty$, bounded by a linear function for $\rho \to +\infty$ (see Remark 4.2), and satisfying conditions (4.7), (4.8) with $l = 0$ and $\nu^* = \tilde{\nu}^*$. As a consequence, $\tilde{\eta}_1 = \tilde{\eta} - \tilde{\eta}_2$ solves (4.4)–(4.6) with $k = 0$ and with $\bar{x} = \tilde{x}$ and satisfies the same conditions at infinity. Then, by Theorems 3.1 and 3.2, such an $\tilde{\eta}_1$ is uniquely determined and satisfies (3.7); it follows that $\tilde{\eta}$ is bounded and asymptotically periodic only if $\tilde{\eta}_2$ has the same properties. By (4.24), this implies $\tilde{\nu}^* = 0$ so that $\tilde{\eta}_2 = 0$. In this case, condition (4.5) becomes $\tilde{\eta}_1(\rho, 0) = -\tilde{x} f''(\varphi^*\rho + \varphi^m)$, for $|\rho| < 1$. However, by condition (4.25), $\tilde{\eta} = \tilde{\eta}_1$ cannot approach an odd (periodic) function for $\rho \to +\infty$, unless $\tilde{x} = 0$. Then we also get $\tilde{\eta} = 0$.

We could now deduce local solvability of (2.25) by the implicit function theorem. We recall, however, that Theorem 4.3 has been proved by assuming a specific symmetry of the solutions at downstream infinity, starting from the additional condition that the right-hand side of (3.10) vanishes. We now show that the theorem holds without this extra assumption if one suitably modifies the definitions of the function spaces X, Y and the condition (4.25); at the end of the section, we will discuss the restrictions on the form of the cylinder's profile f for the validity of the latter condition.

Recalling the discussion at the end of section 3.1, if $\delta_0^* \neq 0$ in (3.12), we change the definition (3.13) of the space X by requiring that the limit function $\chi^\#$ satisfies the condition $\chi^\#(-\rho - \delta_0^*, \sigma) = \overline{\chi^\#(\rho - \delta_0^*, \sigma)}$. Similarly, in the definition (3.16) of the space $Y_{\rho_0,\bar{\rho}}$ we assume $l_\#(-\rho - \delta_0^*) = -l_\#(\rho - \delta_0^*)$; then we can formulate Problem **L** as before, referring to the new spaces X, Y. We can also modify in the obvious way the definition of the space Z below (4.9) and the properties of the function $l^\#$ in the decomposition (4.10). The first crucial remark is that a solution of the auxiliary

problem (see section 4.1) with datum l^* in the new space Z is simply a translation (with respect to ρ) of the solution Ψ given by Proposition 4.1, corresponding to the asymptotically odd datum $\rho \mapsto l^*(\rho - \delta_0^*)$. More precisely, Proposition 4.1 is now satisfied by the function $(\rho, \sigma) \mapsto \Psi(\rho + \delta_0^*, \sigma)$, provided the following condition holds:

$$\int_{-\pi/\mu}^{\pi/\mu} \sin[\mu(\rho + \delta_0^*)] l^{\#}(\rho) = 0. \tag{4.29}$$

We can now proceed as in section 4.1 and choose $\eta_2(\rho, \sigma) = \Psi(\rho + \delta_0^*, \sigma)$ to satisfy conditions (4.7), (4.8); by (4.9) and recalling (3.11), we easily check that there is again a unique value of ν^* such that (4.29) holds. Let us now turn to Theorem 4.3; we define as before $\eta = \eta_1 + \eta_2$, with η_1 solving (4.4)–(4.6) and satisfying the asymptotic condition (4.27); then, by a suitable translation of (4.26) we find that η belongs to the (new) space X if

$$(A_1 - B_+ \sin\delta_0)\sin(\mu\rho) + (B_1 + B_+ \cos\delta_0)\cos(\mu\rho) = C\sin(\mu\rho + \delta_0),$$

that is,

$$A_1 \sin\delta_0 - B_1 \cos\delta_0 = B_+. \tag{4.30}$$

Note that, for $\delta_0 = 0$, (4.30) reduces to the previous condition $B_1 + B_+ = 0$. Recalling (3.9), (3.10), if the assumption (4.25) is replaced by

$$F_0(\varphi^m, \varphi^*) \equiv \int_{-1}^{1} f''(\varphi^*\rho + \varphi^m)[\cos\delta_0\, \beta(\rho) - \sin\delta_0\, \alpha(\rho)]d\rho \neq 0, \tag{4.31}$$

we can solve (4.30) by choosing

$$\bar{x} = \frac{1}{F_0(\varphi^*, \varphi^m)} \left\{ B_+ + \int_{-1}^{1} [k(\rho) - \eta_2(\rho, 0)][\cos\delta_0\, \beta(\rho) - \sin\delta_0\, \alpha(\rho)]d\rho \right\}. \tag{4.32}$$

The rest of the proof of Theorem 4.3 now follows with obvious modifications. Now, with the new definitions of the space X (and Y) and by the implicit function theorem, we can state the following.

THEOREM 4.4. *Let $f \in \mathcal{C}^{3,1}$, $(\varphi^m, \varphi^*) \in \Phi \subset \mathbf{R}^2$, and $\mathcal{U} \subset X \times \mathbf{R}^2$ be given as in Theorem* 3.6*; moreover, assume that condition* (4.31) *holds, with δ_0 defined by the asymptotic condition* (3.11)*. Then there exists $\epsilon_0 > 0$ such that, for every $\epsilon \in [0, \epsilon_0)$, the equation $\mathbf{B}(\chi, \bar{x}, \nu^*, \varphi^m, \varphi^*, \epsilon) = 0$ has a unique solution*

$$(\chi(\varphi^m, \varphi^*, \epsilon), \nu(\varphi^m, \varphi^*, \epsilon), \bar{x}(\varphi^m, \varphi^*, \epsilon)) \in \mathcal{U}.$$

Moreover, the map $\epsilon \mapsto (\chi(\varphi^m, \varphi^*, \epsilon), \nu(\varphi^m, \varphi^*, \epsilon), \bar{x}(\varphi^m, \varphi^*, \epsilon))$ is differentiable.

In view of the discussion of the last conditions (2.13), (2.13′), it will be important to investigate the properties the function F_0 defined in (4.31) (see section 5 and the appendix). We remark here that the form of this function also depends on the data c_0, H, f'. □

5. Solution of the additional conditions. Theorem 4.4 provides, for a given pair of parameters φ^m, φ^*, a function χ holomorphic in A^*, satisfying the requested conditions on the boundary of A^* and the prescribed asymptotic behavior. We still

have to satisfy the continuity condition (2.13) and the additional condition (2.13′). These two conditions, when $\epsilon \to 0$, reduce to

$$\begin{cases} -\varphi^* \displaystyle\int_{-\infty}^{-1} \eta_0(s,0)ds = f(\varphi^m - \varphi^*), \\ \frac{c_0^2}{g}\xi_0(1,0) = f(\varphi^m + \varphi^*). \end{cases} \tag{5.1}$$

Here η_0 is the solution to Problem $\mathbf{L_0}$, and ξ_0 is its harmonic conjugate vanishing at infinity upstream. We shall prove that there exists a pair of numbers (φ^m, φ^*) satisfying system (5.1); then, by a continuity argument, we will deduce the existence, for small enough values of ϵ, of a pair solving (2.13) and (2.13′). Notice that, by integrating from $-\infty$ to -1 the boundary condition following (3.1), we get

$$-\varphi^* \int_{-\infty}^{-1} \eta_0(s,0)ds = \frac{c_0^2}{g}\xi_0(-1,0).$$

Then, by returning to the parameters $\varphi_\pm$ (see (2.14)), we can write system (5.1) in the form

$$\begin{cases} f(\varphi_-/c_0) = \frac{c_0^2}{g}\xi_0(-1,0;\varphi_-,\varphi_+), \\ f(\varphi_+/c_0) = \frac{c_0^2}{g}\xi_0(1,0;\varphi_-,\varphi_+). \end{cases} \tag{5.2}$$

Here we have stressed the dependence of ξ_0 on the unknowns $\varphi_\pm$. By arguments similar to those used in the proof of Proposition 4.1 in [2], one finds that the maps $(\varphi_-,\varphi_+) \mapsto \xi_0(\pm 1,0;\varphi_-,\varphi_+)$ are continuous on the second quadrant of the plane (φ_-,φ_+). We first show that system (5.2) has a solution. Let us fix $R>0$ and define $\mathcal{Q}_R = (-c_0R, 0)\times(0, c_0R)$; consider now the function

$$\mathbf{G} : \mathcal{Q}_R \to \mathbf{R}^2,$$

$$G_1(\varphi_-,\varphi_+) = f(\varphi_-/c_0) - \frac{c_0^2}{g}\xi_0(-1,0;\varphi_-,\varphi_+), \tag{5.3}$$

$$G_2(\varphi_-,\varphi_+) = f(\varphi_+/c_0) - \frac{c_0^2}{g}\xi_0(1,0;\varphi_-,\varphi_+). \tag{5.4}$$

Then we have the following.

LEMMA 5.1. *Let f be a function satisfying* (1.3) *and such that $f' \in W_p^{2-\frac{1}{p}}$ on any interval including the origin; assume further that f satisfies the growth condition $f(x) \approx C_0|x|^\alpha$ (C_0, α positive constants) for large $|x|$ and that this relation can be differentiated. Then there exists $R>0$ such that $G_1(-c_0R,\varphi_+) > 0$ for $0<\varphi_+<c_0R$ and $G_2(\varphi_-, c_0R) > 0$ for $-c_0R < \varphi_- < 0$.*

Proof. By recalling Remark 3.3, the quantities $|\xi_0(\pm 1,0;\varphi_-,\varphi_+)|$ can be bounded by a (local) Sobolev norm of the solution of Problem $\mathbf{L_0}$; then, mapping the problem in the plane of the scaled hodograph variables $(\varphi/c_0, \psi/c_0)$ (which equal the physical space variables at $\epsilon = 0$; see the discussion following (2.4), (2.5)) and using estimates on the solution (see, e.g., [1, equation (4.9)]), it can be proved that the absolute values of the right-hand sides of (5.2) are bounded by the Sobolev norm of f' in the interval $(\varphi_-/c_0, \varphi_+/c_0)$. Then the proposition follows by the assumptions on the growth of f and its derivatives. ∎

Remark 5.2. We note that the quantities $|\xi_0(\pm 1, 0; \varphi_-, \varphi_+)|$ decrease to zero for $\varphi_+ - \varphi_- \to 0$; this means that, in this limit, both components of $\mathbf{G}$ are strictly negative.

PROPOSITION 5.3. *Let f satisfy the assumptions of Lemma* 5.1*; suppose further that for every point in $\mathcal{Q}_R$ with $\varphi_- = 0$ or $\varphi_+ = 0$ we have*

$$f(0) < \frac{c_0^2}{g}\xi_0(1, 0; 0, \varphi_-), \quad f(0) < \frac{c_0^2}{g}\xi_0(-1, 0; \varphi_+, 0). \tag{5.5}$$

Then system (5.2) *has a solution.*

Proof. By Lemma 5.1, conditions (5.5) and definitions (5.3), (5.4) we have

$$G_1(-c_0 R, \varphi_+) > 0, \qquad G_1(0, \varphi_+) < 0$$

for $0 < \varphi_+ < c_0 R$ and

$$G_2(\varphi_-, c_0 R) > 0, \qquad G_2(\varphi_-, 0) < 0$$

for $0 < \varphi_- < c_0 R$. Now the statement that the map $\mathbf{G}$ has a zero in $\mathcal{Q}_R$ is equivalent to the Brouwer fixed point theorem. ∎

Remark 5.4. Regarding condition (5.5), we notice that, in the linear problem, the two quantities $\frac{c_0^2}{g}\xi_0(\pm 1, 0)$ represent the height of the free boundary at the contact points with the cylinder. Then, roughly speaking, the two inequalities in (5.5) state the (quite natural) requirement that the free surface reaches the cylinder's hull at least at the minimum of its profile (placed at $x = 0$). It can be shown that, for fixed $(\varphi_-/c_0, \varphi_+/c_0)$, the quantities $\frac{c_0^2}{g}\xi_0(\pm 1, 0)$ are vanishing for $c_0 \to 0$; hence, the assumptions of Proposition 5.3 hold (for a given f) if the velocity c_0 is small enough. This is in agreement with the physical intuition, since for small values of the velocity the perturbation of the free boundary from the line $y = 0$ should be small, even if compared to the width of a thin obstacle. It is also clear that, for $c_0 \ll 1$, the solutions of (5.2) are such that $\varphi_-/c_0 \approx a$ and $\varphi_+/c_0 \approx b$, where a, b are, respectively, the negative and positive solutions of $f = 0$ (see (1.3)).

Now, assuming that condition (4.31) holds for every pair (φ^*, φ_m) corresponding (through (2.14)) to a point of $\mathcal{Q}_R$, we get the solvability of the system (2.13), (2.13′) for small ϵ. In fact, by Theorem 4.4, we can define a map $\epsilon \mapsto \Omega(\varphi, 0; \varphi_+, \varphi_-, \epsilon)$ such that the composite maps appearing in (2.13), (2.13′) are continuous on $\mathcal{Q}_R$; moreover, (2.13), (2.13′) reduce to (5.1) (that is, (5.2)) when $\epsilon \to 0$.

Thus, as remarked at the end of the previous section, we should study the domain of validity of condition (4.31) for a given profile f and positive c_0, H satisfying (1.1); we note, however, that it is not strictly necessary to satisfy such a condition at every point of $\mathcal{Q}_R$. In fact, by homotopy with the linear map $(\varphi_-, \varphi_+) \mapsto (-\frac{2}{c_0 R}\varphi_- - 1, \frac{2}{c_0 R}\varphi_+ - 1)$, it follows that the topological degree $\deg(\mathcal{Q}_R, \mathbf{G}, \mathbf{0})$ is equal to 1; then, if the solutions of (5.2) are isolated points in $\mathcal{Q}_R$, there is at least one solution $\Phi^* = (\varphi_-^*, \varphi_+^*)$ with local mapping degree (or index $i(\mathbf{G}, \Phi^*, \mathbf{0})$) different from zero by the index sum theorem [15]. Now, assuming only that (4.31) holds at Φ^* and observing that the function F_0 appearing in this condition is continuous, we can still define a map $\epsilon \mapsto \Omega(\varphi, 0; \varphi_+, \varphi_-, \epsilon)$ in a suitable neighborhood of this point and write (2.13), (2.13′) as small perturbations of (5.1) in the same neighborhood; by the continuity property, the local mapping degree does not change for small enough ϵ so that (2.13), (2.13′) have a solution near Φ^*. Then we introduce the following definition:

DEFINITION 5.5. *We say that a pair of positive real numbers c_0, H satisfying* (1.1) *and a real function f (satisfying* (1.3)*) are* admissible data *for the nonlinear problem if the following conditions hold:*

(i) *f satisfies the assumptions of Theorem* 4.4 *and has polynomial growth;*
(ii) *relations* (5.5) *hold;*
(iii) *condition* (4.31) *is satisfied at least for one solution $\Phi^* = (\varphi_-^*, \varphi_+^*)$ of system* (5.2) *with index $i(\mathbf{G}, \Phi^*, \mathbf{0}) \neq 0$.*

In the appendix, by exploiting some qualitative properties of the function F_0, we will show, as an example, that all the conditions of the above definition are satisfied for hulls with parabolic profiles and for every subcritical value of the Froude number $c_0/\sqrt{gH}$.

Summing up the previous results, we can finally state our main result.

THEOREM 5.6. *Let f, c_0, H be admissible data as in Definition* 5.5. *Then one can find $\epsilon_0 > 0$ such that, for every $\epsilon \in [0, \epsilon_0)$, there exist a positive constant c ($c = c_0$ at $\epsilon = 0$), two real numbers $x_- < 0$, $x_+ > 0$, a real function $h(x)$ on $(-\infty, x_-) \cup (x_+, +\infty)$, and a complex function ω holomorphic in the domain S^* defined by* (1.5) *such that conditions* (1.6)–(1.12) *hold. Moreover, the free surface and the cylinder profile form a single $\mathcal{C}^1$ streamline, given in parametric form by*

$$\begin{cases} x(\rho) = \bar{x} + \varphi^* \displaystyle\int_{\rho^m}^{\rho} (1 + \epsilon\xi(s,0))ds, \\ y(\rho) = -\epsilon\varphi^* \displaystyle\int_{-\infty}^{\rho} \eta(s,0)ds, \end{cases} \qquad \rho \in \mathbf{R},$$

where $\varphi^* > 0$, ρ^m, and $\bar{x} \in (x_-, x_+)$ are known quantities (depending on ϵ) and the functions $\xi(\rho, 0)$, $\eta(\rho, 0)$ are now determined from Theorem 4.4. *By the properties of ξ and η, the free surface is exponentially vanishing for $x \to -\infty$ and is bounded and asymptotic to a $\frac{2\pi}{\mu_0}$-periodic function when $x \to +\infty$; here μ_0 is the positive solution of the equation*

$$\tanh(\mu_0 H) = \mu_0 \frac{c_0^2}{g}.$$

As a concluding remark, we observe that, in contrast with the situation encountered in the supercritical case, we are not able to give more detailed information on the location of the contact points $x_\pm$.

Appendix. In part I we prove Theorem 3.2 and property (iii) of Remark 3.3. Moreover, by exploiting some technical results obtained in the course of the proof (see Proposition A.4 below) in part II we write a more explicit form of the function F_0 defined in (4.31) and provide simple examples of data satisfying Definition 5.5.

I. To begin with, we need the following result, which is proved in [3, section 4].

PROPOSITION A.1. *For every $\nu_0^* > 1/H^*$ such that $\mu \neq n\pi/2$ (with μ the solution of* (3.6) *there are nontrivial harmonic functions ζ^s, ζ^c satisfying* (3.2), (3.4) *and the homogeneous condition* (3.3) *(i.e., with $f' = 0$) and with the following properties:*

$$\zeta^s(-\rho, \sigma) = -\zeta^s(\rho, \sigma), \qquad \zeta^c(-\rho, \sigma) = \zeta^c(\rho, \sigma), \tag{A.1}$$

$$\zeta^s(\rho, \sigma) = [\mathcal{A}_s \sin(\mu\rho) + \operatorname{sgn}(\rho)\mathcal{B}_s \cos(\mu\rho)] \sinh[\mu(\sigma + H^*)] + \zeta_0^s(\rho, \sigma), \tag{A.2}$$

$$\zeta^c(\rho, \sigma) = [\operatorname{sgn}(\rho)\mathcal{A}_c \sin(\mu\rho) + \mathcal{B}_c \cos(\mu\rho)] \sinh[\mu(\sigma + H^*)] + \zeta_0^c(\rho, \sigma)), \tag{A.3}$$

where the functions ζ_0^s, ζ_0^c are exponentially decreasing as $\rho \to \infty$ and the coefficients $\mathcal{A}_s$, $\mathcal{A}_c$, $\mathcal{B}_s$, $\mathcal{B}_c$ depend analytically on μ and are such that

$$\mathcal{A}_s\mathcal{B}_c - \mathcal{B}_s\mathcal{A}_c = \Lambda(\mu)\sin\mu\cos\mu, \tag{A.4}$$

with $\Lambda(\mu) > 0$. We stress that the functions ζ^s, ζ^c do not satisfy the asymptotic condition (3.5) and therefore are *not* solutions of the homogeneous Problem $\mathbf{L_0}$. However, they play a crucial role in the proof of unique solvability in Theorem 3.1; in fact, it is proved in [3, Theorem 4.7] that if $\mu \neq n\pi/2$ the solution of Problem $\mathbf{L_0}$ is uniquely determined and has the form

$$\eta_0 = \tilde{\eta}_0 + \alpha_0\zeta^s + \beta_0\zeta^c, \tag{A.5}$$

where $\tilde{\eta}_0$ is a suitably defined harmonic function satisfying the conditions of Problem $\mathbf{L_0}$ *except for the asymptotic condition* (it is in general oscillating at both limits $x \to \pm\infty$; see [3, Proposition 4.3]) and the coefficients α_0, β_0 are uniquely determined by imposing condition (3.5).

Now there are further properties of the functions ζ^s, ζ^c and of the coefficients α_0, β_0 which allow us to prove solvability when the relation $\mu = n\pi/2$ holds. Actually, we have the following.

PROPOSITION A.2. *Let η_0 be the solution of Problem $\mathbf{L_0}$ given by Theorem* 3.1. *Then there exists the limit of η_0 for $\mu \to n\pi/2$ and is still a solution of Problem $\mathbf{L_0}$ with the same regularity and asymptotic properties.*

Proof. Let η_0 be given by (A.5). By Proposition 4.3 of [3], the function $\tilde{\eta}_0$ on the right-hand side is defined for every positive value of μ and satisfies the regularity properties discussed in Remark 3.3; further results in [3] (see section 4 and the appendix) show that the coefficients $\mathcal{A}_s$, $\mathcal{B}_s$ and the function ζ_0^s in (A.2) are proportional to $\sin\mu$, while the coefficients $\mathcal{A}_c$, $\mathcal{B}_c$ and the function ζ_0^c in (A.3) are proportional to $\cos\mu$. Then the functions $\frac{1}{\sin\mu}\zeta^s$, $\frac{1}{\cos\mu}\zeta^c$ have well-defined *uniform limits* in the strip A^* for $\mu \to n\pi/2$, $n = 1, 2, \ldots$; such limits define nontrivial harmonic functions in the strip satisfying the same homogeneous boundary conditions. On the other hand, it follows by explicit calculation (see [3, equation (4.21)]) that for $\mu \neq n\pi/2$ the coefficients of ζ^s, ζ^c in (A.5) have the form

$$\alpha_0 = \frac{\Lambda_{f'}(\mu)}{\sin\mu}, \qquad \beta_0 = \frac{\Lambda_{f'}(\mu)}{\cos\mu},$$

where, for every positive μ, $\Lambda_{f'}(\mu)$ is a linear functional proportional to

$$\int_{-1}^{1} f'(\varphi^*\rho + \varphi^m)\Big[\frac{1}{\cos\mu}\zeta_\sigma^c(\rho,0) - \frac{1}{\sin\mu}\zeta_\sigma^s(\rho,0)\Big]d\rho. \tag{A.6}$$

(The traces of the derivatives ζ_σ^s, ζ_σ^c are continuous functions on $[-1,1]$ by the regularity results of [3]). From the previous discussion, we have that the uniform limits for $\mu \to n\pi/2$, $n = 1, 2, \ldots$, of the functions η_0 given by Theorem 3.1 exist and are solutions to Problem $\mathbf{L_0}$ corresponding to the values $n\pi/2$ of the solution of (3.6). ∎

In order to prove the uniqueness statement of Theorem 3.2, we investigate the relation between the coefficients A_0, B_0 in the asymptotic representation (3.7) and the functions ζ^s, ζ^c.

LEMMA A.3. *Let η_0 be a solution of Problem $\mathbf{L_0}$ with $\mu \neq n\pi/2$, and let ζ^s, ζ^c be defined by Proposition* A.1. *Then the following formulas hold:*

$$A_0 = \frac{K^*}{\mu\Lambda\sin\mu\cos\mu}\int_{-1}^{1} f'(\varphi^*\rho + \varphi^m)[\mathcal{A}_c\zeta_\sigma^s(\rho,0) - \mathcal{A}_s\zeta_\sigma^c(\rho,0)]d\rho, \tag{A.7}$$

$$B_0 = \frac{K^*}{\mu\Lambda \sin\mu\cos\mu}\int_{-1}^{1} f'(\varphi^*\rho+\varphi^m)[\mathcal{B}_c\zeta^s_\sigma(\rho,0)-\mathcal{B}_s\zeta^c_\sigma(\rho,0)]d\rho, \tag{A.8}$$

where Λ *is defined by* (A.4) *and*

$$K^* = \frac{2}{H^*}\Big(\frac{\sinh(2\mu H^*)}{2\mu H^*}-1\Big)^{-1}. \qquad \square$$

Proof. Apply Green's formula to η_0 and to each of the harmonic functions ζ^s, ζ^c in the bounded rectangle $(-R,R)\times(-H^*,0)$, with $R>1$; then, letting $R\to\infty$ and taking account of (3.7), (A.2), and (A.3) (which can be differentiated with respect to ρ) we get

$$\begin{aligned}
0 &= \int_{-1}^{1} f'(\varphi^*\rho+\varphi^m)\zeta^s_\sigma(\rho,0)d\rho \\
&+ \lim_{R\to+\infty}\int_{-H^*}^{0}\Big[\zeta^s(R,\sigma)\partial_\rho\eta_0(R,\sigma)-\eta_0(R,\sigma)\zeta^s_\rho(R,\sigma)\Big]d\sigma \\
&= \int_{-1}^{1} f'(\varphi^*\rho+\varphi^m)\zeta^s_\sigma(\rho,0)d\rho \\
&+ \mu(A_0\mathcal{B}_s - B_0\mathcal{A}_s)\int_{-H^*}^{0}\sinh^2[\mu^*(\sigma+H^*)]d\sigma,
\end{aligned}$$

$$\begin{aligned}
0 &= \int_{-1}^{1} f'(\varphi^*\rho+\varphi^m)\zeta^c_\sigma(\rho,0)d\rho \\
&+ \lim_{R\to+\infty}\int_{-H^*}^{0}\Big[\zeta^c(R,\sigma)\partial_\rho\eta_0(R,\sigma)-\eta_0(R,\sigma)\zeta^c_\rho(R,\sigma)\Big]d\sigma \\
&= \int_{-1}^{1} f'(\varphi^*\rho+\varphi^m)\zeta^c_\sigma(\rho,0)d\rho \\
&+ \mu(A_0\mathcal{B}_c - B_0\mathcal{A}_c)\int_{-H^*}^{0}\sinh^2[\mu^*(\sigma+H^*)]d\sigma.
\end{aligned}$$

Then, (A.7) and (A.8) follow by elementary calculations. $\square$

PROPOSITION A.4. *Let* $\mu\neq n\pi/2$*; then the relations* (3.9), (3.10) *are satisfied by choosing*

$$\alpha(\rho) = \frac{K^*}{\mu\Lambda\sin\mu\cos\mu}[\mathcal{A}_c\zeta^s_\sigma(\rho,0)-\mathcal{A}_s\zeta^c_\sigma(\rho,0)], \tag{A.9}$$

$$\beta(\rho) = \frac{K^*}{\mu\Lambda\sin\mu\cos\mu}[\mathcal{B}_c\zeta^s_\sigma(\rho,0)-\mathcal{B}_s\zeta^c_\sigma(\rho,0)]. \tag{A.10}$$

Moreover, the right-hand sides of (A.9), (A.10) *have limits for* $\mu\to n\pi/2$*, which verify* (3.9), (3.10) *for* $\mu = n\pi/2$.

Proof. From (A.7), (A.8), the above defined α, β verify relations (3.9), (3.10) for $\mu\neq n\pi/2$. On the other hand, recalling the proof of Proposition A.2, the proof of Lemma A.3 is also valid for $\mu = n\pi/2$ by replacing ζ^s, ζ^c with the limits for

$\mu \to n\pi/2$ of $\frac{1}{\sin\mu}\zeta^s$, $\frac{1}{\cos\mu}\zeta^c$, respectively. Furthermore, it is easily checked that also the right-hand sides of (A.9), (A.10) have finite uniform limits for $\mu \to n\pi/2$. Hence, the proposition follows. □

Proof of Theorem 3.2. Existence of a solution η_0 for the "singular values" of μ follows from Proposition A.2. Suppose that $\hat{\eta}_0$ is another solution corresponding to $\mu = n\pi/2$ and with the same boundary data; then by (3.9), (3.10) we have that $\hat{\eta}_0$ satisfies (3.7) with the same coefficients A_0, B_0 as η_0. Then $\eta_0 - \tilde{\eta}_0$ is a waveless solution of the *homogeneous* Problem $\mathbf{L_0}$; in particular, $\eta_0 - \hat{\eta}_0$ belongs to $H^1(A^*)$. By the uniqueness of a variational solution (see [14, Theorem 4.7]) we get $\eta_0 = \tilde{\eta}_0$.

II. *Discussion of condition* (4.31) *and an example of data satisfying Definition* 5.5. From (A.9), (A.10), the function F_0 in condition (4.31) can be written

$$F_0(\varphi^*, \varphi^m) = \int_{-1}^{1} f''(\varphi^*\rho + \varphi^m)\gamma_0(\rho)d\rho,$$

where $\gamma_0(\rho)$ is proportional to the function

$$\frac{1}{\sin\mu\cos\mu}\Big[(\cos\delta_0\,\mathcal{B}_c - \sin\delta_0\,\mathcal{A}_c)\zeta^s_\sigma(\rho,0) - (\cos\delta_0\mathcal{B}_s - \sin\delta_0\,\mathcal{A}_s)\zeta^c_\sigma(\rho,0)\Big].$$

We note that, by the proof of Lemma A.3 and by the definition of δ_0 in (3.11), the coefficients of ζ^s_σ, ζ^c_σ in the above expression are proportional to the scalar products

$$(f', \zeta^c_\sigma) = \int_{-1}^{1} f'(\varphi^*\rho + \varphi^m)\zeta^c_\sigma(\rho,0)d\rho,$$

$$(f', \zeta^s_\sigma) = \int_{-1}^{1} f'(\varphi^*\rho + \varphi^m)\zeta^s_\sigma(\rho,0)d\rho,$$

respectively; hence, we can write (4.31) in the form

$$\frac{1}{\sin\mu\cos\mu}\Big[(f', \zeta^c_\sigma)(f'', \zeta^s_\sigma) - (f'', \zeta^c_\sigma)(f', \zeta^s_\sigma)\Big] \neq 0, \tag{A.11}$$

where the scalar products involving f'' are defined in the same way. It can be shown that the above scalar products are analytic functions of the parameters ν_0^*, H^* of Problem $\mathbf{L_0}$; recalling (2.17) and (3.1), we conclude that the left-hand side of (A.11) is an analytic function of φ^* and φ^m if also f is analytic. We now show that this function does not vanish identically in the simple case of the parabolic profile $f(x) = x^2/2 - \gamma$ ($\gamma > 0$). Then we have $f'(\varphi^*\rho + \varphi^m) = \varphi^*\rho + \varphi^m$ and $f''(\varphi^*\rho + \varphi^m) = 1$, with $\rho \in [-1, 1]$; by the symmetry relations (A.1) we get from (A.11)

$$\frac{\varphi^*}{\sin\mu\cos\mu}\left(\int_{-1}^{1}\zeta^c_\sigma(\rho,0)d\rho\right)\left(\int_{-1}^{1}\rho\zeta^s_\sigma(\rho,0)d\rho\right) \neq 0. \tag{A.12}$$

We note that the left-hand side of (A.12) is independent of φ^m. By estimates of the above integrals which follow from the definitions of ζ^s, ζ^c (see [3, equations (4.13), (4.14)]) and by scaling $\rho \to \varphi^*\rho$, $\sigma \to \varphi^*\sigma$, one can show that (A.12) holds for $\varphi^* > 0$ small enough. Then (A.12) is satisfied for every positive value of φ^*, except possibly for a discrete set. We further remark that the integrals in the above condition depend on φ^* only through the parameters ν_0^*, H^*; therefore, *for a given Froude number*

$\frac{c_0}{\sqrt{gH}} = \frac{1}{\sqrt{\nu_0^* H^*}}$ we can say that (A.12) holds for any *fixed* value of the ratio ν_0^*/H^* outside a discrete set. Let us now discuss system (5.2) for the parabolic profile; we denote by $a_\pm$ and $b_\pm$ the values of $\xi_0(\pm 1, 0)$ corresponding to a solution of Problem $\mathbf{L_0}$, respectively, with the functions ρ and 1 on the right-hand side of (3.3). We stress that, for a given Froude number, the quantities $a_\pm$, $b_\pm$ depend only on the ratio ν_0^*/H^*; by linearity and the relation $\frac{c_0^2}{g} = \varphi^*/\nu_0^*$, system (5.2) takes the form (in terms of the variables φ^*, φ^m)

$$\text{(A.13)} \qquad \begin{cases} \frac{1}{2}(\varphi^m - \varphi^*)^2 - \gamma = \frac{a_-}{\nu_0^*}\varphi^{*2} + \frac{b_-}{\nu_0^*}\varphi^*\varphi^m, \\ \frac{1}{2}(\varphi^m + \varphi^*)^2 - \gamma = \frac{a_+}{\nu_0^*}\varphi^{*2} + \frac{b_+}{\nu_0^*}\varphi^*\varphi^m. \end{cases}$$

We will solve this system with respect to φ^*, φ^m for *given values* of ν_0^*, H^*; in terms of the physical parameters, this means that we do not fix c_0 and H but only the Froude number $\frac{c_0}{\sqrt{gH}}$. Now, again by scaling arguments, one can show that the four quantities $\frac{a_\pm}{\nu_0^*}$, $\frac{b_\pm}{\nu_0^*}$ are vanishing for $\nu_0^*/H^* \to \infty$ (at a fixed Froude number). Then, by elementary calculations, we find that (A.13) has a unique solution (with $\varphi^* > 0$) for any large enough values of ν_0^*/H^*; by the previous discussion, these values, except possibly for a discrete subset, also satisfy (A.12). ∎

REFERENCES

[1] C. D. Pagani and D. Pierotti, *On solvability of the nonlinear wave resistance problem for a surface-piercing symmetric cylinder*, SIAM J. Math. Anal., 32 (2000), pp. 214–233.

[2] C. D. Pagani and D. Pierotti, *The forward motion of an unsymmetric surface-piercing cylinder: The solvability of a nonlinear problem in the supercritical case*, Quart. J. Mech. Appl. Math., 54 (2001), pp. 85–106.

[3] D. Pierotti, *On unique solvability and regularity in the linearized two dimensional wave resistance problem*, Quart. Appl. Math., 61 (2003), pp. 639–655.

[4] N. G. Kuznetsov, V. G. Maz'ya, and V. Vainberg, *Linear Water Waves*, Cambridge University Press, Cambridge, UK, 2002.

[5] J. Asavanant and J. M. Vanden-Broeck, *Free surface flows past a surface-piercing object of finite length*, J. Fluid Mech., 273 (1994), pp. 109–124.

[6] C. J. Amick and J. F. Toland, *On solitary waves of finite amplitude*, Arch. Rational Mech. Anal., 76 (1981), pp. 9–95.

[7] K. Kirchgässner, *Nonlinearly resonant surface waves*, Adv. Appl. Mech., 26 (1988), pp. 135–181.

[8] E. Zeidler, *Nonlinear Functional Analysis and Its Applications. Volume* IV: *Applications to Mathematical Physics*, Springer-Verlag, New York, 1988.

[9] A. Mielke, *Reduction of quasilinear elliptic equations in cylindrical domains with applications*, Math. Methods Appl. Sci., 10 (1988), pp. 51–66.

[10] A. Mielke, *Steady flows of inviscid fluids under localized perturbations*, J. Differential Equations, 65 (1986), pp. 89–116.

[11] A. Ambrosetti and G. Prodi, *A Primer of Nonlinear Analysis*, Cambridge University Press, Cambridge, UK, 1993.

[12] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, London, 1985.

[13] W. Rudin, *Functional Analysis*, Tata McGraw-Hill, New Delhi, 1974.

[14] D. Pierotti, *The subcritical motion of a surface-piercing cylinder: Existence and regularity of waveless solutions of the linearized problem*, Adv. Differential Equations, 7 (2002), pp. 385–418.

[15] E. Zeidler, *Nonlinear Functional Analysis and Its Applications. Volume* I: *Fixed Point Theorems*, Springer-Verlag, New York, 1985.

SIAM J. MATH. ANAL.
Vol. 36, No. 1, pp. 94–120

SUPREMAL MULTISCALE SIGNAL ANALYSIS*

ULISSES BRAGA-NETO† AND JOHN GOUTSIAS‡

Abstract. We introduce a novel approach to nonlinear signal analysis, which is referred to as *supremal multiscale analysis*. The proposed approach provides a rigorous mathematical foundation for a class of nonlinear multiscale signal analysis schemes and leads to a decomposition that can effectively be used in signal processing and analysis. Moreover, it is related to the supremal scale-spaces proposed by Heijmans and van den Boomgaard and is similar in flavor to the well-known linear multiresolution theory of Mallat and Meyer. In this framework, linear concepts such as vector spaces, projections, and linear operators are replaced by conceptually analogous nonlinear notions.

We use supremal multiscale analysis to construct a multiscale image decomposition scheme based on two mathematical concepts that play a key role in the analysis and interpretation of images by vision systems, namely, regional maxima and connectivity. The resulting scheme is referred to as *skyline supremal multiscale analysis* and satisfies several useful properties desired by any multiscale image analysis tool. It is grayscale invariant, as well as translation and scale invariant. Moreover, it progressively removes connected components from the level sets of an image without introducing new ones. But, most importantly, it decomposes the regional maxima of an image in a natural causal hierarchy by gradually removing these maxima without introducing new ones.

Image decomposition by skyline supremal multiscale analysis can be used to construct nonlinear tools for image processing and analysis that provide solutions to problems where traditional linear techniques are ineffective. We discuss one such tool and illustrate its use in object-based extraction and denoising.

Key words. complete lattices, connectivity, connected operators, mathematical morphology, multiscale signal approximation, multiscale signal analysis, nonlinear signal analysis, scale-spaces, object-based image analysis

AMS subject classifications. 68U10, 94A08, 94A12

DOI. 10.1137/S0036141002409945

1. Introduction.

An important methodology for signal processing and analysis represents a signal at multiple scales. This methodology is based on the fundamental observation that information pertaining to features of interest in a signal is not confined to a particular scale, but it may span several scales. In order to effectively characterize such information, it is necessary to gradually simplify the signal, by means of a scale-dependent operator, which monotonically removes features of interest as the scale increases. The resulting evolution of a signal from fine to coarse scales is known as a *scale-space* (e.g., see [1, 2, 18, 21, 22, 42]).

Although early scale-space techniques were based on linear operators, it has been increasingly recognized that these techniques severely limit the capability of scale-spaces to accurately represent features of interest at coarser scales. For this reason, scale-space techniques based on nonlinear operators (or nonlinear partial differential equations) have appeared in the literature (e.g., see [1, 2, 32, 38, 41]). It is noticeable

*Received by the editors June 20, 2002; accepted for publication (in revised form) October 10, 2003; published electronically June 22, 2004. This work was supported by the Office of Naval Research, Mathematical, Computer, and Information Sciences Division, under ONR grants N00014-90-1345 and N00014-01-1-0027.

http://www.siam.org/journals/sima/36-1/40994.html

†Section of Clinical Cancer Genetics, University of Texas MD Anderson Cancer Center, Houston, TX 77030 and Department of Electrical Engineering, Texas A&M University, College Station, TX 77840 (ulisses@ee.tamu.edu). This author was supported by the CNPq Scholarship 200725196-3 of the Brazilian government.

‡Center for Imaging Science and Department of Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, MD 21218 (goutsias@jhu.edu).

that several of these techniques are based on *morphological operators* [1, 2, 3, 8, 9, 10, 15, 17, 28, 29, 37, 38].

On the other hand, a popular approach for multiscale signal processing and analysis is based on the multiresolution theory of Mallat [23, 24, 25] and Meyer [30]. According to this approach, approximations of a given signal at various scales (or resolutions) can be computed by means of orthogonal projections of the signal on a sequence of approximation spaces. The signal is then represented by means of a coarse approximation plus added details. The details are computed by means of orthogonal projections of the given signal on a sequence of detail spaces, with the detail spaces being orthogonal complements to the corresponding approximation spaces. At finer scales, the approximation error tends to zero, and a signal is spanned by spaces of successive details at all resolutions. This approach has naturally led to popular techniques for signal processing and analysis based on wavelet decompositions and filter banks (e.g., see [25, 39]).

The basic assumption behind the multiresolution theory of Mallat and Meyer is that signals reside in a vector space (namely, the space $L^2(\mathbb{R})$ of finite energy functions), with the approximation and detail spaces being subspaces of this vector space. Therefore, the theory is applied to linear multiscale tools for signal analysis. An attempt to conceptualize this approach in a nonlinear setting has appeared in [13, 14]. However, the discussion in [13, 14] on this issue is only preliminary.

To accomplish this goal, it is necessary (among other things) to extend linear concepts such as vector spaces, orthogonal projections, orthogonal spaces, and linear operators to a nonlinear setting. One way to do this is to assume that the signal space is a *complete lattice* (i.e., a nonempty collection of partially ordered elements such that any subcollection has a supremum and an infimum [4]). Complete lattices form the algebraic foundation of *mathematical morphology* [16], which assumes that signals are not combined by means of numerical addition and subtraction but by means of supremum and infimum. In mathematical morphology, an operator is "linear" if it commutes over suprema or infima. In the former case, the operator is called a *dilation*, whereas, in the latter case, it is called an *erosion*. Many linear concepts, such as convolution, can be recast in terms of suprema and infima (e.g., see [12, 26, 27]).

In this paper, we introduce a novel approach to nonlinear signal analysis that provides a rigorous mathematical foundation for a class of nonlinear multiscale signal analysis schemes and leads to a decomposition that can effectively be used for signal processing and analysis. The proposed approach, which we refer to as *supremal multiscale analysis*, is related to the supremal scale-spaces proposed by Heijmans and van den Boomgaard in [15] and is similar in flavor to the well-known linear multiresolution theory of Mallat and Meyer. In this framework, vector spaces are replaced by sup-closed spaces, projections are replaced by idempotent operators, orthogonal projections are replaced by sup-projections, orthogonal spaces are replaced by sup-orthogonal spaces, and linear operators are replaced by morphological operators.

We use supremal multiscale analysis to construct a multiscale image decomposition scheme, based on morphological reconstruction operators, which selectively removes regional maxima from a signal. Perhaps the most important feature of the proposed scheme, which is referred to as *skyline supremal multiscale analysis*, is its construction by means of two mathematical concepts that play a key role in the analysis and interpretation of images by vision systems, namely, regional maxima and connectivity (e.g., see [17, 20, 34]). This scheme represents a signal as the supremum of a coarse approximation and details. The coarse approximation preserves regional

maxima that are at some level σ or above, while it flattens the rest. In addition, the details preserve regional maxima with values in nonoverlapping subintervals of $(0, \sigma)$ and flatten the rest. The skyline supremal multiscale analysis is shown to satisfy a number of useful properties desired by any multiscale signal analysis tool. It is grayscale, translation, and scale invariant. Moreover, it progressively removes connected components from the level sets of a signal without introducing new ones. But, most importantly, it decomposes the regional maxima of a signal in a natural causal hierarchy by gradually removing these maxima without introducing new ones.

Image decomposition by skyline supremal multiscale analysis can be used to construct nonlinear tools for image processing and analysis that can provide solutions to some problems where traditional linear techniques are ineffective. We discuss one such tool and illustrate its effectiveness in object-based extraction and denoising.

This paper is structured as follows. In section 2, we provide a brief overview of basic mathematical concepts used throughout the paper and introduce our notation. In section 3, we introduce our framework for nonlinear multiscale analysis, which leads to the concepts of supremal multiscale approximation and supremal multiscale analysis. We also establish a relationship between the supremal multiscale approximation and scale-spaces and present two binary examples that illustrate these concepts. In section 4, we present the skyline supremal multiscale analysis scheme, constructed by means of morphological reconstruction operators, which decomposes the regional maxima of a signal in a natural causal hierarchy by selectively removing these maxima without introducing new ones. We show that the proposed scheme is indeed a supremal multiscale analysis, and we study its main properties. In section 5, we present examples that illustrate the use of the proposed multiscale approach in two image processing and analysis problems: object-based extraction and denoising. In the first case, the skyline supremal multiscale decomposition scheme is used to extract objects of interest, by placing them on individual frames, and enhance their presence by suppressing (flattening) surrounding details. In the second case, the scheme is used to restore an image corrupted by structured (more than a pixel thick) "pepper" noise. Finally, we summarize our conclusions in section 6.

2. Mathematical preliminaries. In this section, we review basic mathematical concepts and introduce our notation. For a more detailed exposition, the reader is referred to [4, 5, 6, 7, 16, 35, 36].

A *partially ordered set* or, briefly, a *poset*, is a nonempty set furnished with a binary *partial order relation* $\leq$ (i.e., a binary order relation that is reflexive, antisymmetric, and transitive). A *complete lattice* $(\mathcal{L}, \leq)$ is a poset such that every family $\mathcal{M} \subseteq \mathcal{L}$ has an infimum $\bigwedge \mathcal{M}$ and a supremum $\bigvee \mathcal{M}$ in $\mathcal{L}$. Every complete lattice $(\mathcal{L}, \leq)$ has a *least element* O and a *greatest element* I, given by $O = \bigwedge \mathcal{L}$ and $I = \bigvee \mathcal{L}$, respectively. In this paper, whenever we use the term "lattice" we mean "complete lattice." In addition, we often refer to "lattice $\mathcal{L}$" when there is no confusion as to the underlying partial order.

The following are some examples of lattices.

Example 1.

(a) The collection $\mathcal{P}(E)$ of all subsets of a set E, with set inclusion as the partial order. The infimum and supremum are set intersection and set union, respectively. This lattice is used as a mathematical model for binary images defined on E.

(b) The collection $\mathcal{G}(\mathbb{R}^d)$ of all open subsets of the Euclidean space $\mathbb{R}^d$, with set inclusion as the partial order. The infimum is the topological interior of set

intersection, whereas the supremum is set union. This lattice can also be used as a mathematical model for binary images on E. In this case, it is assumed that images do not include their boundary.

(c) The set $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$ of extended real numbers and the set $\overline{\mathbb{Z}} = \mathbb{Z} \cup \{-\infty, \infty\}$ of extended integers, as well as any closed subinterval of those, with the usual numerical ordering as the partial order. The infimum and supremum are given by the usual numerical infimum and supremum. These are *chains* (i.e., totally ordered lattices), which are used for modeling image values.

(d) The collection $\mathrm{Fun}(E, \mathcal{T})$ of all functions from a set E into a lattice $\mathcal{T}$, with the partial order $f \leq g$ if $f(v) \leq_{\mathcal{T}} g(v)$, for all $v \in E$, where "$\leq_{\mathcal{T}}$" is the partial order relation on $\mathcal{T}$. The infimum and supremum are the pointwise infimum and supremum, given, respectively, by $(\bigwedge f_\alpha)(v) = \bigwedge f_\alpha(v)$ and $(\bigvee f_\alpha)(v) = \bigvee f_\alpha(v)$, for all $v \in E$, where the infimum and supremum on the right-hand side are in $\mathcal{T}$. When $\mathcal{T}$ is a chain, this lattice is used as a mathematical model for grayscale images defined on E.

(e) The collection $\mathrm{Fun}_u(E, \mathcal{T})$ of all *upper semicontinuous* (u.s.c.) functions [19] from a topological space E into a lattice $\mathcal{T}$, with the partial order $f \leq g$ if $f(v) \leq_{\mathcal{T}} g(v)$, for $v \in E$. The infimum is the usual pointwise infimum, given by $(\bigwedge_u f_\alpha)(v) = \bigwedge f_\alpha(v)$, for $v \in E$. However, the supremum is given by $(\bigvee_u f_\alpha)(v) = \bigvee\{t \in \mathcal{T} \mid v \in \overline{\bigcup X_t(f_\alpha)}\}$, for $v \in E$, where $X_t(f) = \{v \in E \mid f(v) \geq t\}$ is the *level set* of f at level t, and $\overline{A}$ denotes the *closure* of a set A [7, Prop. 4.2.6]. Nevertheless, it can be shown that the supremum of any finite family of u.s.c. functions corresponds to the usual pointwise supremum. In general, whenever $\bigvee f_\alpha$ is u.s.c., then $\bigvee_u f_\alpha = \bigvee f_\alpha$ so that the supremum in lattice $\mathrm{Fun}_u(E, \overline{\mathbb{R}})$ can, and often does, reduce to the usual pointwise supremum. When $\mathcal{T}$ is a chain, this lattice is also used as a mathematical model for grayscale images defined on E. In this case, however, images are assumed to satisfy the property of upper semicontinuity.

The level sets of a function $f \in \mathrm{Fun}(E, \mathcal{T})$ satisfy the following properties: (a) $X_t(f) \subseteq X_s(f)$ if $t \geq s$; (b) $f \leq g$ if and only if $X_t(f) \subseteq X_t(g)$ for all $t \in \mathcal{T}$ (in particular, $f = g$ if and only if $X_t(f) = X_t(g)$ for all $t \in \mathcal{T}$); (c) for $t \in \mathcal{T}$, we have that $X_t(\bigwedge f_\alpha) = \bigcap X_t(f_\alpha)$, whereas $X_t(\bigvee f_\alpha) = \bigcup X_t(f_\alpha)$ if $\{f_\alpha\}$ is a finite family or if $\mathcal{T}$ is finite; (d) $f \in \mathrm{Fun}_u(E, \mathcal{T})$ if and only if the sets $X_t(f)$ are closed in E for all $t \in \mathcal{T}$.

Given a family $\mathcal{M} \subseteq \mathcal{L}$, we denote by $\langle\, \mathcal{M} \mid \vee \,\rangle$ the family sup-generated by $\mathcal{M}$, i.e., the family consisting of all elements of $\mathcal{L}$ that are obtained by taking suprema of elements of $\mathcal{M}$. The family $\mathcal{M}$ is said to be *sup-closed* if $\mathcal{M} = \langle\, \mathcal{M} \mid \vee \,\rangle$ (in particular, $\mathcal{M}$ must be nonempty, since $O = \bigvee \emptyset \in \mathcal{M}$).

A subset $\mathcal{S}$ of a lattice $\mathcal{L}$ is called a *sup-generating family* for $\mathcal{L}$ if every element of $\mathcal{L}$ can be written as the supremum of elements in $\mathcal{S}$; i.e., $\mathcal{L} = \langle\, \mathcal{S} \mid \vee \,\rangle$. An element of the sup-generating family $\mathcal{S}$ is called a *sup-generator*. It is assumed here that O is not a sup-generator; i.e., $O \notin \mathcal{S}$. For example, the lattice $\mathcal{P}(E)$ of binary images is sup-generated by the points in E. We define the family $\mathcal{S}(A) = \{x \in \mathcal{S} \mid x \leq A\}$ for $A \in \mathcal{L}$. Clearly, A is sup-generated by $\mathcal{S}(A)$.

An *operator* ψ on a lattice $\mathcal{L}$ is a mapping $\psi\colon \mathcal{L} \to \mathcal{L}$. The *invariance domain* of ψ is defined as $\mathrm{Inv}(\psi) = \{A \in \mathcal{L} \mid \psi(A) = A\}$. An operator ψ is said to be *increasing*, if $A \leq B \Rightarrow \psi(A) \leq \psi(B)$, for all $A, B \in \mathcal{L}$; *antiextensive*, if $\psi(A) \leq A$, for all $A \in \mathcal{L}$; *idempotent*, if $\psi\psi(A) = \psi(A)$, for all $A \in \mathcal{L}$. If ψ distributes over infima,

it is called an *erosion*, whereas, if it distributes over suprema, it is called a *dilation*. If ψ is increasing, antiextensive, and idempotent, it is called an *opening*. It can be shown (e.g., see [33]) that if $\{\gamma_\alpha\}$ is a family of openings, then $\bigvee \gamma_\alpha$ is an opening as well, with $\mathrm{Inv}(\bigvee \gamma_\alpha) = \langle \bigcup \mathrm{Inv}(\gamma_\alpha) \mid \vee \rangle$. Given a poset $\mathcal{K}$, the family of openings $\{\gamma_\alpha \mid \alpha \in \mathcal{K}\}$ is a *granulometry* if $\gamma_{\alpha_1} \leq \gamma_{\alpha_2}$ for $\alpha_1 \geq \alpha_2$.

The translation A_h of a set $A \in \mathcal{P}(E)$ is another set in $\mathcal{P}(E)$, given by $A_h = \{v + h \mid v \in A\}$. The *translation-invariant erosion* of $A \in \mathcal{P}(E)$ by a *structuring element* $B \in \mathcal{P}(E)$ is defined as $\epsilon_B(A) = A \ominus B = \{h \in E \mid B_h \subseteq A\}$. Similarly, the *translation-invariant dilation* of A by B is defined as $\delta_B(A) = A \oplus B = \bigcup\{B_h \mid h \in A\}$. It can be shown that the operator $\theta_B(A) = A \bigcirc B = (A \ominus B) \oplus B = \bigcup_{h \in E}\{B_h \mid B_h \subseteq A\}$ is an opening. This operator is referred to as a *structural opening*. If $A \in \mathrm{Inv}(\theta_B)$, we say that A is *B-open*.

An increasing operator ψ on $\mathcal{L}$ is said to be *$\downarrow$-continuous* if, for every totally ordered subset $\mathcal{K}$ of $\mathcal{L}$ that contains at most a countable number of elements, we have that

$$\psi\left(\bigwedge \mathcal{K}\right) = \bigwedge_{A \in \mathcal{K}} \psi(A).$$

If ψ is an $\downarrow$-continuous operator on a lattice $\mathcal{L}$ and if $\{A(s) \mid s \in \overline{\mathbb{R}}\}$ is a decreasing family of elements in $\mathcal{L}$, then [7, Prop. 2.2.10]

$$\psi\left(\bigwedge_{s<t} A(s)\right) = \bigwedge_{s<t} \psi(A(s)) \quad \forall\, t \in \overline{\mathbb{R}}. \tag{2.1}$$

Consider now a lattice $\mathcal{L}$, with a sup-generating family $\mathcal{S}$. A family $\mathcal{C} \subseteq \mathcal{L}$ is called a *connectivity class* in $\mathcal{L}$ if the following conditions are satisfied:

(i) $O \in \mathcal{C}$;
(ii) $\mathcal{S} \subseteq \mathcal{C}$;
(iii) for a family $\{C_\alpha\}$ in $\mathcal{C}$ such that $\bigwedge C_\alpha \neq O$, we have that $\bigvee C_\alpha \in \mathcal{C}$.

The family $\mathcal{C}$ generates a *connectivity* in $\mathcal{L}$, and the elements in $\mathcal{C}$ are said to be *connected*.

Classical topological and graph-theoretic connectivities correspond to connectivity classes, and so do several examples of fuzzy connectivity [5, 7]. Moreover, based on the notion of connectivity class, many new interesting examples of connectivity can be defined [5, 6, 7, 35, 36].

We say that C is a *connected component* of $A \in \mathcal{L}$ if $C \in \mathcal{C}$, $C \leq A$, and there is no $C' \in \mathcal{C}$ different from C such that $C \leq C' \leq A$. In other words, a connected component of an object is a maximal connected part of the object. The set of connected components of A is denoted by $\mathcal{C}(A)$.

We can define an operator $\gamma_x(A)$ that extracts connected components from elements $A \in \mathcal{L}$ by

$$\gamma_x(A) = \bigvee\{C \in \mathcal{C} \mid x \leq C \leq A\}, \quad A \in \mathcal{L}, \quad x \in \mathcal{S}.$$

It can be seen that this operator is an opening; it is called the *connectivity opening* associated with $\mathcal{C}$. It can also be checked that $\gamma_x(A) \in \mathcal{C}$. As a matter of fact, $\gamma_x(A)$ is the connected component C of A *marked* by x (i.e., such that $x \leq C$).

It is natural to extend connectivity openings to operators that extract connected components marked by arbitrary markers, not just sup-generators. This gives rise

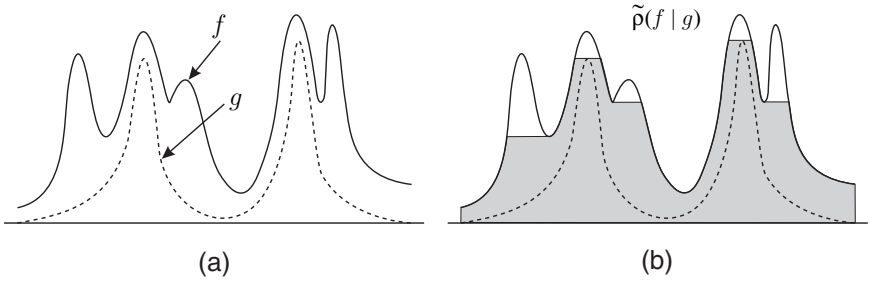

FIG. 1. (a) *Original image f and a marker g.* (b) *The grayscale reconstruction $\widetilde{\rho}(f \mid g)$, according to the usual topological connectivity of the Euclidean real line.*

to the *reconstruction operator* associated with a connectivity class $\mathcal{C}$. For a marker $M \in \mathcal{L}$, the *reconstruction* $\rho(A \mid M)$ of a given $A \in \mathcal{L}$ from M is defined by

$$\rho(A \mid M) = \bigvee_{x \in \mathcal{S}(M)} \gamma_x(A) = \bigvee \{C \in \mathcal{C}(A) \mid C \wedge M \neq O\}.$$

The second equality above can be easily verified [6, 7]. Hence, the reconstruction operator $\rho(A \mid M)$ extracts the connected components of A that "intersect" marker M. Being a supremum of openings, the operator $\rho(\cdot \mid M)$ is an opening on $\mathcal{L}$ for a fixed marker $M \in \mathcal{L}$. When M reduces to a sup-generator x, the reconstruction $\rho(A \mid x)$ reduces to the connectivity opening $\gamma_x(A)$, provided that $x \leq A$.

Given a connectivity class $\mathcal{C}$ in the binary lattice $\mathcal{P}(E)$ and the associated reconstruction operator ρ: $\mathcal{P}(E) \times \mathcal{P}(E) \to \mathcal{P}(E)$, we can define an operator $\widetilde{\rho}$: $\mathrm{Fun}(E, \mathcal{T}) \times \mathrm{Fun}(E, \mathcal{T}) \to \mathrm{Fun}(E, \mathcal{T})$ by

$$\widetilde{\rho}(f \mid g)(v) = \bigvee \{t \in \mathcal{T} \mid v \in \rho(X_t(f) \mid X_t(g))\}, \quad v \in E. \tag{2.2}$$

It can be shown that $\widetilde{\rho}(\cdot \mid g)$ is an opening on $\mathrm{Fun}(E, \mathcal{T})$ for a fixed marker $g \in \mathrm{Fun}(E, \mathcal{T})$. If we assume that $\mathcal{T}$ is a chain, then the operator $\widetilde{\rho}(f \mid g)$ in (2.2) is known as the *grayscale reconstruction* of f from marker g associated with the connectivity class $\mathcal{C}$. The grayscale reconstruction is a very useful operator in applications [40]. Figure 1 illustrates the grayscale reconstruction operator in the one-dimensional case.

3. Supremal scale-spaces and multiscale analysis. In this section, we introduce a framework for nonlinear multiscale signal analysis, which is related to the supremal scale-spaces introduced by Heijmans and van den Boomgaard [15]. The proposed framework is referred to as *supremal multiscale analysis* and leads to a nonlinear multiscale signal representation scheme that decomposes a signal into the supremum of a coarse approximation and details. We show that supremal multiscale analysis satisfies a number of properties, which are similar in flavor to properties satisfied by the well-known linear multiresolution signal analysis scheme of Mallat [23, 24, 25] and Meyer [30]. As a matter of fact, we derive the supremal multiscale analysis scheme by using nonlinear analogues of certain linear concepts (e.g., vector spaces, orthogonal projections, and orthogonal complements).

We assume that signals of interest reside in a complete lattice $(\mathcal{V}, \leq)$. An operator ϕ on $\mathcal{V}$ is said to be a *projection* if it is idempotent [14, 31]. Furthermore, we say that ϕ is a projection on $\mathcal{U} \subseteq \mathcal{V}$ if $\mathrm{Ran}(\phi) = \mathcal{U}$ and ϕ is idempotent on $\mathcal{U}$ (in which case $\mathrm{Inv}(\phi) = \mathrm{Ran}(\phi) = \mathcal{U}$), where $\mathrm{Ran}(\phi)$ denotes the *range* of operator ϕ. In the linear

multiresolution framework proposed in [23, 24, 25, 30], approximations of signals at various scales are computed by means of orthogonal projections on a sequence of approximation spaces. An orthogonal projection of a signal $f \in \mathcal{V}$ on a vector subspace $\mathcal{U} \subseteq \mathcal{V}$ is defined as the signal $\phi(f) \in \mathcal{U}$ that minimizes the norm $\|f - g\|$ over all signals $g \in \mathcal{U}$. Note that the range of this operator is $\mathcal{U}$ and that the operator is idempotent; therefore, it is a projection on $\mathcal{U}$. In the nonlinear framework proposed here, $\mathcal{V}$ is not a vector space in general. Therefore, we introduce the alternative notion of sup-projection, which is conceptually analogous to an orthogonal projection.

DEFINITION 3.1. *Let $\mathcal{U} \subseteq \mathcal{V}$ such that $\mathcal{U}$ is sup-closed in $\mathcal{V}$. The operator*

$$\phi(f) = \bigvee \{g \in \mathcal{U} \mid g \leq f\}, \quad f \in \mathcal{V}, \tag{3.1}$$

defines the sup-projection *of f on $\mathcal{U}$.*

The sup-closure requirement and (3.1) imply that $\mathrm{Ran}(\phi) = \mathcal{U}$ and that ϕ is idempotent on $\mathcal{U}$; therefore, ϕ is a projection on $\mathcal{U}$. The requirement that $\mathcal{U}$ must be sup-closed is analogous to the linear requirement that $\mathcal{U}$ must be a vector space (i.e., closed under linear combinations). Note that (3.1) implies that $\phi(f)$ is the "closest" element to f in $\mathcal{U}$, in the sense of the underlying partial order. Hence, a sup-projection is a nonlinear analogue of an orthogonal projection.

A fundamental aspect of (linear) multiresolution analysis is that distinct signal approximations can be obtained from each other by means of scaling (this is known as the "dilation" property). In order to formulate this idea in a nonlinear setting, we use a general definition of scaling, proposed in [15]. In what follows, id denotes the identity operator.

DEFINITION 3.2. *A family $S = \{s_t \mid t \in (0, \infty)\}$ of operators on a lattice $\mathcal{V}$ is a* scaling *if*

(i) $s_1 = \mathsf{id}$,
(ii) $s_r s_t = s_{rt}$ *for* $r, t \in (0, \infty)$.

This definition implies that S is a commutative group, where the inverse s_t^{-1} of s_t is given by $s_t^{-1} = s_{1/t}$ for $t \in (0, \infty)$. Moreover, if $S = \{s_t \mid t \in (0, \infty)\}$ is a scaling on $\mathcal{V}$, then so is $S^p = \{s_{t^p} \mid t \in (0, \infty)\}$, $p \in \mathbb{R}$ [15].

Example 2.

(a) For $\mathcal{V} = \mathcal{P}(\mathbb{R}^d)$, the scaling $\{tA \mid t \in (0, \infty)\}$, where $tA = \{tv \mid v \in A\}$, for $A \in \mathcal{V}$, is known as the *spatial scaling*.
(b) For $\mathcal{V} = \mathrm{Fun}(E, \overline{\mathbb{R}})$, the scalings $\{tf(\cdot) \mid t \in (0, \infty)\}$, $\{f(\cdot/t) \mid t \in (0, \infty)\}$, and $\{tf(\cdot/t) \mid t \in (0, \infty)\}$ are known as the *gray-level*, *spatial*, and *umbral scalings*, respectively.

In practice, useful scalings consist of increasing operators. We refer to these as *increasing scalings*. For example, all scalings considered in Example 2 are increasing. The following result shows that scalings are increasing if and only if they consist of dilations.

PROPOSITION 3.3. *A scaling $S = \{s_t \mid t \in (0, \infty)\}$ on a lattice $\mathcal{V}$ is increasing if and only if s_t is a dilation for every $t \in (0, \infty)$.*

Proof. The reverse implication follows trivially from the fact that every dilation is an increasing operator [16]. We show the direct implication. Given $t \in (0, \infty)$ and $\{f_\alpha\} \subseteq \mathcal{V}$, we have that $s_t(\bigvee f_\alpha) \geq \bigvee s_t(f_\alpha)$, since s_t is increasing. To show the reverse inequality, note that $s_t^{-1}(\bigvee s_t(f_\alpha)) \geq \bigvee s_t^{-1} s_t(f_\alpha) = \bigvee f_\alpha$, since $s_t^{-1} = s_{1/t}$ is increasing. Applying s_t on both sides of this inequality gives $s_t s_t^{-1}(\bigvee s_t(f_\alpha)) = \bigvee s_t(f_\alpha) \geq s_t(\bigvee f_\alpha)$. Therefore, $s_t(\bigvee f_\alpha) = \bigvee s_t(f_\alpha)$, and s_t is a dilation on $\mathcal{V}$. ∎

We now introduce an axiomatic formulation for supremal multiscale approximations, which compute approximations of signals at various scales by means of sup-projections on a sequence of approximation spaces.

DEFINITION 3.4. *Let S be a scaling on a lattice $\mathcal{V}$. A family $\{\mathcal{V}_\sigma \mid \sigma \in (0,\infty)\}$ of sup-closed subsets of $\mathcal{V}$ is said to be a* supremal multiscale S-approximation *of $\mathcal{V}$ if the following properties are satisfied:*

1. *The sequence $\{\mathcal{V}_\sigma \mid \sigma \in (0,\infty)\}$ is decreasing; i.e.,*

$$\mathcal{V}_\tau \subseteq \mathcal{V}_\sigma \quad \text{for} \quad \tau \geq \sigma. \tag{3.2}$$

   This implies that an approximation at scale σ contains all necessary information to compute an approximation at a coarser scale $\tau \geq \sigma$.

2. *The sequence $\{\mathcal{V}_\sigma \mid \sigma \in (0,\infty)\}$ "converges" to $\mathcal{V}$, as $\sigma \to 0^+$, in the sense that*

$$\lim_{\sigma \to 0^+} \mathcal{V}_\sigma \triangleq \left\langle \bigcup_{\sigma \in (0,\infty)} \mathcal{V}_\sigma \mid \vee \right\rangle = \mathcal{V}. \tag{3.3}$$

   This implies that any signal can be recovered by the supremum of its approximations at sufficiently small scales.

3. *(S-invariance). We have that*

$$f \in \mathcal{V}_\sigma \quad \Leftrightarrow \quad s_{\tau/\sigma}(f) \in \mathcal{V}_\tau \quad \text{for} \quad \sigma, \tau \in (0,\infty). \tag{3.4}$$

   This means that an approximation space $\mathcal{V}_\sigma$ can be obtained from another approximation space $\mathcal{V}_\tau$, and vice-versa, by means of scaling.

The previous properties are similar in flavor to properties satisfied by the linear multiresolution analysis scheme of Mallat and Meyer. The approximation of a signal $f \in \mathcal{V}$ at scale $\sigma \in (0,\infty)$ is given by the sup-projection of f on the approximation space $\mathcal{V}_\sigma$, which therefore must be sup-closed. This is the nonlinear analogue of the assumption that the approximation spaces are vector subspaces. The inclusion property specified by (3.2) is also true in the linear case. The convergence requirement specified by (3.3) is similar to the one in the linear case, except that linear closure is replaced by sup-closure. Finally, the scaling requirement specified by (3.4) is similar to the one in the linear case. However, a very important property of the linear case is the existence of vector bases for the approximation spaces. This is an inherently linear property and has no counterpart in a nonlinear setting.

Note that the sup-closure assumption implies that, for $\sigma \in (0,\infty)$, $\mathcal{V}_\sigma$ is a complete lattice under the partial order of $\mathcal{V}$ [16, Prop. 2.12]. The proof of the following result is straightforward.

PROPOSITION 3.5. *The family $\{\mathcal{V}_\sigma \mid \sigma \in (0,\infty)\}$ is a supremal multiscale S-approximation of $\mathcal{V}$ if and only if the family $\{\mathcal{V}_{\sigma^p} \mid \sigma \in (0,\infty)\}$, $p \in \mathbb{R}$, is a supremal multiscale S^p-approximation of $\mathcal{V}$.*

For each $\sigma \in (0,\infty)$, let us define the *approximation operator* ϕ_σ on $\mathcal{V}$ as the operator that maps an element $f \in \mathcal{V}$ to its sup-projection on $\mathcal{V}_\sigma$; i.e.,

$$\phi_\sigma(f) = \bigvee \{g \in \mathcal{V}_\sigma \mid g \leq f\}, \quad f \in \mathcal{V}, \quad \sigma \in (0,\infty). \tag{3.5}$$

The operator ϕ_σ provides the approximation of a signal in $\mathcal{V}$ at scale σ. The following fundamental result implies that a supremal multiscale approximation can be specified by its approximation operators.

PROPOSITION 3.6. *Let $\{\mathcal{V}_\sigma \mid \sigma \in (0,\infty)\}$ be a supremal multiscale S-approximation of $\mathcal{V}$, where S is an increasing scaling. The family $\{\phi_\sigma \mid \sigma \in (0,\infty)\}$ of approximation operators is such that $\mathrm{Inv}(\phi_\sigma) = \mathcal{V}_\sigma$, for $\sigma \in (0,\infty)$, and*

(i) *the family $\{\phi_\sigma \mid \sigma \in (0,\infty)\}$ is a granulometry on $\mathcal{V}$,*
(ii) $\bigvee_{\sigma\in(0,\infty)} \phi_\sigma = \mathsf{id}$,
(iii) $\phi_\sigma = s_{\sigma/\tau}\, \phi_\tau\, s^{-1}_{\sigma/\tau}$ *for $\sigma, \tau \in (0,\infty)$.*

Conversely, if $\{\phi_\sigma \mid \sigma \in (0,\infty)\}$ is a family of operators on $\mathcal{V}$ that satisfies the previous properties (i)–(iii)*, then $\{\mathcal{V}_\sigma = \mathrm{Inv}(\phi_\sigma) \mid \sigma \in (0,\infty)\}$ is a supremal multiscale S-approximation of $\mathcal{V}$, with approximation operators that coincide with $\{\phi_\sigma \mid \sigma \in (0,\infty)\}$.*

Proof. For $f \in \mathrm{Inv}(\phi_\sigma)$, we have that $f = \phi_\sigma(f) = \bigvee\{g \in \mathcal{V}_\sigma \mid g \le f\} \Rightarrow f \in \langle \mathcal{V}_\sigma \mid \vee \rangle = \mathcal{V}_\sigma$, since $\mathcal{V}_\sigma$ is sup-closed, so that $\mathrm{Inv}(\phi_\sigma) \subseteq \mathcal{V}_\sigma$. The reverse inclusion follows easily from (3.5); hence, $\mathrm{Inv}(\phi_\sigma) = \mathcal{V}_\sigma$. We now show (i). From (3.5), it is clear that ϕ_σ is increasing and antiextensive. This implies that $\phi_\sigma\phi_\sigma \le \phi_\sigma$. On the other hand, (3.5) implies that $g \in \mathcal{V}_\sigma$, $g \le f \Rightarrow g \le \phi_\sigma(f)$ so that $\phi_\sigma(f) = \bigvee\{g \in \mathcal{V}_\sigma \mid g \le f\} \le \bigvee\{g \in \mathcal{V}_\sigma \mid g \le \phi_\sigma(f)\} = \phi_\sigma\phi_\sigma(f)$ for $f \in \mathcal{V}$. Therefore, ϕ_σ is idempotent, and hence it is an opening. For $\tau \ge \sigma$, we have that $\mathrm{Inv}(\phi_\tau) = \mathcal{V}_\tau \subseteq \mathcal{V}_\sigma = \mathrm{Inv}(\phi_\sigma)$, which implies that $\phi_\tau \le \phi_\sigma$ [16, Thm. 3.24]. Therefore, $\{\phi_\sigma \mid \sigma \in (0,\infty)\}$ is a granulometry on $\mathcal{V}$. To show (ii), we use the fact that the supremum $\bigvee \theta_\alpha$ of openings is an opening, with $\mathrm{Inv}(\bigvee \theta_\alpha) = \langle \bigcup \mathrm{Inv}(\theta_\alpha) \mid \vee \rangle$. Hence, $\mathrm{Inv}(\bigvee_{\sigma\in(0,\infty)} \phi_\sigma) = \langle \bigcup_{\sigma\in(0,\infty)} \mathrm{Inv}(\phi_\sigma) \mid \vee \rangle = \langle \bigcup_{\sigma\in(0,\infty)} \mathcal{V}_\sigma \mid \vee \rangle = \mathcal{V} = \mathrm{Inv}(\mathsf{id})$. But, since id is an opening and two openings are equal if and only if their domains of invariance are equal [16, Thm. 3.24], we get that $\bigvee_{\sigma\in(0,\infty)} \phi_\sigma = \mathsf{id}$. We now show (iii). For $f \in \mathcal{V}$, we have that $\phi_\sigma(f) = \bigvee\{g \in \mathcal{V}_\sigma \mid g \le f\} = \bigvee\{s_{\sigma/\tau}(h) \mid h \in \mathcal{V}_\tau,\ s_{\sigma/\tau}(h) \le f\} = s_{\sigma/\tau}(\bigvee\{h \mid h \in \mathcal{V}_\tau,\ h \le s^{-1}_{\sigma/\tau}(f)\}) = s_{\sigma/\tau}\, \phi_\tau\, s^{-1}_{\sigma/\tau}(f)$, since $s_{\sigma/\tau}$ is a dilation (see Proposition 3.3). Therefore, $\phi_\sigma = s_{\sigma/\tau}\, \phi_\tau\, s^{-1}_{\sigma/\tau}$ for $\sigma, \tau \in (0,\infty)$.

We now show the converse implication. Note that, for each $\sigma \in (0,\infty)$, $\mathcal{V}_\sigma = \mathrm{Inv}(\phi_\sigma)$ is sup-closed, since ϕ_σ is an opening. Equation (3.2) follows from the fact that $\phi_\sigma \le \phi_\tau$ for $\sigma \ge \tau$ [16, Thm. 3.24]. Equation (3.3) follows from $\mathcal{V} = \mathrm{Inv}(\mathsf{id}) = \mathrm{Inv}(\bigvee_{\sigma\in(0,\infty)} \phi_\sigma) = \langle \bigcup_{\sigma\in(0,\infty)} \mathcal{V}_\sigma \mid \vee \rangle$. To verify (3.4), note that $f \in \mathcal{V}_\sigma \Leftrightarrow f = \phi_\sigma(f) = s_{\sigma/\tau}\, \phi_\tau\, s^{-1}_{\sigma/\tau}(f) \Leftrightarrow s_{\tau/\sigma}(f) = \phi_\tau s_{\tau/\sigma}(f) \Leftrightarrow s_{\tau/\sigma}(f) \in \mathcal{V}_\tau$ for $\sigma, \tau \in (0,\infty)$. Finally, if $\{\phi'_\sigma \mid \sigma \in (0,\infty)\}$ are the approximation operators associated with $\{\mathcal{V}_\sigma \mid \sigma \in (0,\infty)\}$, then $\mathrm{Inv}(\phi'_\sigma) = \mathcal{V}_\sigma = \mathrm{Inv}(\phi_\sigma) \Leftrightarrow \phi'_\sigma = \phi_\sigma$ for $\sigma \in (0,\infty)$ (see [16, Thm. 3.24]). ∎

The following proposition shows that we can build supremal multiscale approximations by using unions of existing ones.

PROPOSITION 3.7. *Let S be an increasing scaling. If, for each α, $\{\mathcal{V}^\alpha_\sigma \mid \sigma \in (0,\infty)\}$ is a supremal multiscale S-approximation of $\mathcal{V}$, with approximation operators $\{\phi^\alpha_\sigma \mid \sigma \in (0,\infty)\}$, then $\{\langle \bigcup_\alpha \mathcal{V}^\alpha_\sigma \mid \vee \rangle \mid \sigma \in (0,\infty)\}$ is a supremal multiscale S-approximation of $\mathcal{V}$ as well, with approximation operators $\{\bigvee_\alpha \phi^\alpha_\sigma \mid \sigma \in (0,\infty)\}$.*

Proof. Equations (3.2) and (3.3) are easy to show; therefore, we show only (3.4). Let $\mathcal{V}_\sigma = \langle \bigcup_\alpha \mathcal{V}^\alpha_\sigma \mid \vee \rangle$ for $\sigma \in (0,\infty)$. For $f \in \mathcal{V}_\sigma$, we have that $f = \bigvee f_\beta$, where each f_β belongs to some $\mathcal{V}^\alpha_\sigma$. Therefore, $s_{\tau/\sigma}(f_\beta) \in \mathcal{V}_\tau$, for each β and $\tau \in (0,\infty)$. Since $\mathcal{V}_\tau$ is sup-closed and since $s_{\tau/\sigma}$ is a dilation (see Proposition 3.3), we have that $s_{\tau/\sigma}(f) = s_{\tau/\sigma}(\bigvee f_\beta) = \bigvee s_{\tau/\sigma}(f_\beta) \in \mathcal{V}_\tau$, as required. Now let ϕ_σ be the approximation operator associated with $\mathcal{V}_\sigma$. Since $\mathrm{Inv}(\phi^\alpha_\sigma) = \mathcal{V}^\alpha_\sigma$, we have that $\mathrm{Inv}(\phi_\sigma) = \mathcal{V}_\sigma = \langle \bigcup_\alpha \mathcal{V}^\alpha_\sigma \mid \vee \rangle = \mathrm{Inv}(\bigvee_\alpha \phi^\alpha_\sigma)$. Since two openings are equal if and only if their domains of invariance are equal [16, Thm. 3.24] and since $\bigvee_\alpha \phi^\alpha_\sigma$ is an opening, we conclude that $\phi_\sigma = \bigvee_\alpha \phi^\alpha_\sigma$, as required. ∎

To illustrate the concept of supremal multiscale approximation, we now provide two binary examples. A grayscale supremal multiscale approximation scheme will be discussed in the next section.

Example 3.

(a) Let $\mathcal{V} = \mathcal{G}(\mathbb{R}^d)$ be the lattice of open subsets of the Euclidean space, and consider the spaces

$$\mathcal{V}_\sigma = \{A \in \mathcal{V} \mid A \text{ is } \sigma B\text{-open}\}, \quad \sigma \in (0, \infty), \tag{3.6}$$

where $B \in \mathcal{V}$ is a bounded convex structuring element (e.g., an open ball of unit radius). In (3.6), $\sigma B = \{\sigma b \mid b \in B\}$. The family $\{\mathcal{V}_\sigma \mid \sigma \in (0, \infty)\}$ is a supremal multiscale S-approximation of $\mathcal{V}$, where S is the spatial scaling: $\mathcal{V}_\sigma$ is sup-closed, for $\sigma \in (0, \infty)$, and (3.2) and (3.4) are clearly satisfied, whereas (3.3) follows from the facts that $\mathcal{B} = \{(\sigma B)_v \mid v \in \mathbb{R}^d, \sigma \in (0, \infty)\} \subset \bigcup_{\sigma \in (0,\infty)} \mathcal{V}_\sigma$ and $\mathcal{B}$ is a basis for the Euclidean topology. In this case, the approximation operators are the structural openings

$$\phi_\sigma(A) = A \bigcirc \sigma B, \quad A \in \mathcal{V}, \quad \sigma \in (0, \infty).$$

(b) Let $\mathcal{V} = \mathcal{G}(\mathbb{R}^d)$, furnished with a connectivity class $\mathcal{C}$, and consider the *opening by reconstruction operators*:

$$\phi_\sigma(A) = \rho(A \mid A \bigcirc \sigma B), \quad A \in \mathcal{V}, \quad \sigma \in (0, \infty), \tag{3.7}$$

where $B \in \mathcal{V}$ is a bounded structuring element. It can be shown that the invariance domain of ϕ_σ is given by

$$\mathcal{V}_\sigma(A) = \{A \in \mathcal{V} \mid C \bigcirc \sigma B \neq \emptyset \ \forall\, C \in \mathcal{C}(A)\}, \quad \sigma \in (0, \infty). \tag{3.8}$$

The family $\{\mathcal{V}_\sigma \mid \sigma \in (0, \infty)\}$ is a supremal multiscale S-approximation of $\mathcal{V}$, where S is the spatial scaling. Properties (i) and (iii) of Proposition 3.6 are clearly satisfied. Now we have that $\phi_\sigma(A) = \bigcup\{C \in \mathcal{C}(A) \mid C \cap (A \ominus \sigma B) \neq \emptyset\} = \bigcup\{C \in \mathcal{C}(A) \mid \exists\, v \in C \text{ such that } (\sigma B)_v \subseteq A\}$. Since A is open, for any $C \in \mathcal{C}(A)$ and $v \in C$, we can find a σ such that $(\sigma B)_v \subseteq A$ so that $C \subseteq \phi_\sigma(A)$. It then follows that $A = \bigcup_{\sigma \in (0,\infty)} \phi_\sigma(A)$, which shows property (ii). The approximation operators $\{\phi_\sigma \mid \sigma \in (0, \infty)\}$ are, of course, given by (3.7).

We now show that supremal multiscale approximations and scale-spaces are related. The following definition introduces the notion of supremal scale-space in the terminology of [15].

DEFINITION 3.8. *Let $\mathcal{V}$ be a lattice, and let S be a scaling on $\mathcal{V}$. A family $\{\phi_\sigma \mid \sigma \in (0, \infty)\}$ of operators on $\mathcal{V}$ is said to be a* supremal S-scale-space *if*

(i) $\phi_\sigma \phi_\tau = \phi_{\sigma \vee \tau}$ *for* $\sigma, \tau \in (0, \infty)$,

(ii) $\phi_\sigma s_\sigma = s_\sigma \phi_1$ *for* $\sigma \in (0, \infty)$.

We have the following result.

PROPOSITION 3.9. *Let $\{\mathcal{V}_\sigma \mid \sigma \in (0, \infty)\}$ be a supremal multiscale S-approximation of $\mathcal{V}$, where S is an increasing scaling. The family $\{\phi_\sigma \mid \sigma \in (0, \infty)\}$ of approximation operators, given by* (3.5)*, is a supremal S-scale-space.*

Proof. Properties (i) and (ii) of a supremal scale-space follow directly from properties (i) and (iii) in Proposition 3.6. □

Therefore, given a signal $f \in \mathcal{V}$, its approximations $\{\phi_\sigma(f) \mid \sigma \in (0, \infty)\}$ form a scale-space, where increasing scale corresponds to an "evolution" of f towards decreasing levels of "detail." Several scale-spaces that coincide with or are similar to the

two binary supremal multiscale approximation schemes of Example 3 have appeared in [3, 8, 9, 10, 15, 17, 28, 29, 37].

We now proceed to derive a nonlinear multiscale signal analysis scheme based on supremal multiscale approximations.

In linear orthogonal wavelet decomposition schemes, given two approximation spaces $\mathcal{V}_{\sigma+1} \subseteq \mathcal{V}_\sigma$, one defines a *detail space* $\mathcal{W}_{\sigma+1}$ (also called a *wavelet space*) as the orthogonal complement of $\mathcal{V}_{\sigma+1}$ in $\mathcal{V}_\sigma$, given by

$$\mathcal{W}_{\sigma+1} = \{f \in \mathcal{V}_\sigma \mid f \perp \mathcal{V}_{\sigma+1}\}. \tag{3.9}$$

From the fact that $\mathcal{V}_\sigma$ and $\mathcal{V}_{\sigma+1}$ are vector spaces, it follows that $\mathcal{W}_{\sigma+1}$ is a vector space as well. Moreover, $\mathcal{W}_{\sigma+1} \subseteq \mathcal{V}_\sigma$ and $\mathcal{W}_{\sigma+1} \perp \mathcal{V}_{\sigma+1}$.

In vector analysis, a space $\mathcal{V}$ is said to be the *direct sum* of two subspaces $\mathcal{V}_1$ and $\mathcal{V}_2$, which is denoted by $\mathcal{V} = \mathcal{V}_1 \oplus \mathcal{V}_2$, if

$$\mathcal{V} = \{f + g \mid f \in \mathcal{V}_1 \text{ and } g \in \mathcal{V}_2\} \quad \text{and} \quad \mathcal{V}_1 \cap \mathcal{V}_2 = \{O\}. \tag{3.10}$$

A fundamental property of linear wavelet analysis is that $\mathcal{V}_\sigma = \mathcal{V}_{\sigma+1} \oplus \mathcal{W}_{\sigma+1}$; i.e., the approximation space at scale σ is the direct sum of the approximation and detail spaces at scale $\sigma + 1$ (which, in this case, is also an orthogonal sum, since $\mathcal{W}_{\sigma+1} \perp \mathcal{V}_{\sigma+1}$).

In order to formulate similar ideas in a nonlinear setting, we need to define nonlinear analogues of the notions of "orthogonal complement" and "direct sum." A signal f is said to be *sup-orthogonal* to a sup-closed space $\mathcal{V}$ if its sup-projection on $\mathcal{V}$ is O. Therefore, a signal f is sup-orthogonal to an approximation space $\mathcal{V}_\sigma$ if and only if $\phi_\sigma(f) = O$. A space $\mathcal{W}$ is said to be sup-orthogonal to $\mathcal{V}$ if every signal in $\mathcal{W}$ is sup-orthogonal to $\mathcal{V}$. Note that this implies that $\mathcal{W}$ and $\mathcal{V}$ cannot have any common elements other than $\{O\}$. Given two approximation spaces $\mathcal{V}_\sigma$ and $\mathcal{V}_\tau$, with $\tau > \sigma$, we define a detail space $\mathcal{W}_{\sigma,\tau}$ as the *sup-orthogonal complement* of $\mathcal{V}_\tau$ in $\mathcal{V}_\sigma$, given by

$$\mathcal{W}_{\sigma,\tau} = \{f \in \mathcal{V}_\sigma \mid \phi_\tau(f) = O\}, \quad \tau > \sigma. \tag{3.11}$$

Note that this is the nonlinear analogue of (3.9). Moreover, we say that a space $\mathcal{V}$ is the *direct sup-sum* of two subspaces $\mathcal{V}_1$ and $\mathcal{V}_2$, which we denote by $\mathcal{V} = \mathcal{V}_1 \bigvee \mathcal{V}_2$, if

$$\mathcal{V} = \langle\, \mathcal{V}_1 \cup \mathcal{V}_2 \mid \vee \,\rangle \quad \text{and} \quad \mathcal{V}_1 \cap \mathcal{V}_2 = \{O\}.$$

This is the analogue of (3.10), where vector summation is replaced by supremum.

We now have the following definition.

DEFINITION 3.10. *Let $\{\mathcal{V}_\sigma \mid \sigma \in (0,\infty)\}$ be a supremal multiscale S-approximation of $\mathcal{V}$. If the detail spaces $\mathcal{W}_{\sigma,\tau}$, given by* (3.11)*, satisfy the property*

$$\mathcal{V}_\sigma = \mathcal{V}_\tau \bigvee \mathcal{W}_{\sigma,\tau} \quad \textit{for} \quad \sigma \in (0,\infty),\ \tau \in (\sigma,\infty), \tag{3.12}$$

then $\{\mathcal{V}_\sigma,\ \mathcal{W}_{\sigma,\tau} \mid \sigma \in (0,\infty),\ \tau \in (\sigma,\infty)\}$ is said to be a supremal multiscale S-analysis *of $\mathcal{V}$.*

It follows that, in a supremal multiscale analysis, for every $\sigma \in (0,\infty)$, $\tau \in (\sigma,\infty)$, we have that

(a) $\mathcal{V}_\tau, \mathcal{W}_{\sigma,\tau} \subseteq \mathcal{V}_\sigma$,
(b) $\mathcal{W}_{\sigma,\tau}$ is sup-orthogonal to $\mathcal{V}_\tau$,
(c) $\mathcal{V}_\sigma$ is the direct sup-sum of $\mathcal{V}_\tau$ and $\mathcal{W}_{\sigma,\tau}$.

These properties are the nonlinear analogues of similar properties satisfied by the approximation and detail spaces in linear orthogonal wavelet analysis.

Next, we define the notion of detail operator.

DEFINITION 3.11. *Let* $\{\mathcal{V}_\sigma, \mathcal{W}_{\sigma,\tau} \mid \sigma \in (0,\infty), \tau \in (\sigma,\infty)\}$ *be a supremal multiscale S-analysis of* $\mathcal{V}$. *If* $\psi_{\sigma,\tau}$ *is a projection on* $\mathcal{W}_{\sigma,\tau}$ *and*

$$\phi_\sigma = \phi_\tau \vee \psi_{\sigma,\tau} \quad \textit{for} \quad \sigma \in (0,\infty),\ \tau \in (\sigma,\infty), \tag{3.13}$$

then $\{\psi_{\sigma,\tau} \mid \sigma \in (0,\infty), \tau \in (\sigma,\infty)\}$ *is a family of* detail operators *of the supremal multiscale S-analysis.*

From (3.13), it follows that an approximation $\phi_\sigma(f)$ of a signal $f \in \mathcal{V}$ can be decomposed, in a unique way, as the supremum of a sup-projection $\phi_\tau(f)$ on $\mathcal{V}_\tau$ and a projection $\psi_{\sigma,\tau}(f)$ on $\mathcal{W}_{\sigma,\tau}$. Furthermore, $\psi_{\sigma,\tau}(f)$ is sup-orthogonal to $\mathcal{V}_\tau$; i.e., $\phi_\tau\psi_{\sigma,\tau}(f) = O$. Therefore, the approximation of f at scale σ has a unique decomposition as the supremum of the approximation signal at scale τ and a sup-orthogonal detail signal, which contains information about f that is present at scale σ but is removed at the coarser scale τ. Finally, by applying the supremum $\bigvee_{\sigma\in(0,\tau)}$ on both sides of (3.13) and by using properties (i) and (ii) of Proposition 3.6, we get

$$f = \phi_\tau(f) \vee \bigvee_{\sigma\in(0,\tau)} \psi_{\sigma,\tau}(f) \quad \text{for} \quad \tau \in (0,\infty). \tag{3.14}$$

This shows that a signal $f \in \mathcal{V}$ can be uniquely decomposed in terms of a scaled signal $\phi_\tau(f)$ at scale $\tau \in (0,\infty)$ and detail signals $\psi_{\sigma,\tau}(f)$, $\sigma \in (0,\tau)$. It is worthwhile noticing that the decomposition suggested by (3.14) is conceptually analogous to the well-known wavelet decomposition.

Note that, for $\tau' \geq \tau$, we have $\phi_{\tau'}\psi_{\sigma,\tau}(f) \leq \phi_\tau\psi_{\sigma,\tau}(f) = O \Rightarrow \phi_{\tau'}\psi_{\sigma,\tau}(f) = O$. Therefore, a detail signal $\psi_{\sigma,\tau}(f)$ is sup-orthogonal to all approximation spaces $\mathcal{V}_{\tau'}$, $\tau' \geq \tau$.

In practice, a multiscale signal decomposition scheme can be constructed by selecting initial and final approximation scales σ and τ and a set of intermediary scales $\sigma_0 = \sigma, \sigma_1, \ldots, \sigma_{N-1}, \sigma_N = \tau$ such that $\sigma_k < \sigma_{k+1}$ for $k = 0, 1, \ldots, N-1$. Then, by repeatedly applying (3.13), we get

$$\phi_\sigma(f) \;=\; \phi_\tau(f) \;\vee \bigvee_{0\leq k\leq N-1} \psi_{\sigma_k,\sigma_{k+1}}(f), \quad \tau > \sigma, \tag{3.15}$$

which provides a decomposition of the approximation $\phi_\sigma(f)$ of a signal f into the sequence $\{\phi_\tau(f), \psi_{\sigma,\sigma_1}(f), \psi_{\sigma_1,\sigma_2}(f), \ldots, \psi_{\sigma_{N-1},\tau}(f)\}$. Clearly, all detail signals are sup-orthogonal to the final approximation space $\mathcal{V}_\tau$; i.e., $\phi_\tau\psi_{\sigma_k,\sigma_{k+1}}(f) = O$ for $k = 0, 1, \ldots, N-1$.

Example 4.

(a) Let $\mathcal{V} = \mathcal{G}(\mathbb{R}^d)$, and consider the supremal multiscale approximation of $\mathcal{V}$ given in Example 3(a). Using (3.6) and (3.11), we get

$$\mathcal{W}_{\sigma,\tau} = \{A \in \mathcal{V} \mid A \bigcirc \sigma B = A \ \text{ and } \ A \bigcirc \tau B = \emptyset\}, \quad \tau > \sigma.$$

Clearly, $\mathcal{V}_\sigma = \langle\, \mathcal{V}_\tau \cup \mathcal{W}_{\sigma,\tau} \mid \vee \,\rangle$ and $\mathcal{V}_\sigma \cap \mathcal{W}_{\sigma,\tau} = \{\emptyset\}$. Therefore, (3.12) is satisfied so that $\{\mathcal{V}_\sigma, \mathcal{W}_{\sigma,\tau} \mid \sigma \in (0,\infty), \tau \in (\sigma,\infty)\}$ is a supremal multiscale S-analysis of $\mathcal{V}$, where S is the spatial scaling. If B is a structuring element that contains the origin, then the operators

$$\psi_{\sigma,\tau}(A) = A \bigcirc \sigma B \smallsetminus A \ominus \tau B, \quad \tau > \sigma,$$

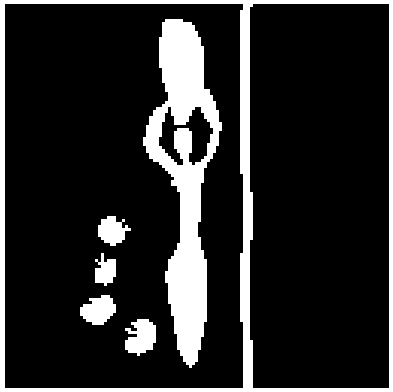

FIG. 2. *A binary "Matisse" image.*

are projections on $\mathcal{W}_{\sigma,\tau}$ that satisfy (3.13). Therefore, $\{\psi_{\sigma,\tau} \mid \sigma \in (0,\infty),\ \tau \in (\sigma,\infty)\}$ is a family of detail operators associated with the supremal multiscale analysis. As an illustration, this scheme is applied on the binary "Matisse" image[1] depicted in Figure 2. The result is depicted in Figure 3. Note that

$$\psi_{\sigma,\tau}(A) = (A \ominus \sigma B) \oplus \sigma B \smallsetminus (A \ominus \sigma B) \ominus (\tau - \sigma)B, \quad \tau > \sigma.$$

This shows that the detail signal $\psi_{\sigma,\tau}(A)$ is obtained by applying a *morphological gradient* on the erosion $A \ominus \sigma B$ (a morphological gradient is an operator of the form $A \oplus tB \smallsetminus A \ominus sB$, where B is a structuring element that contains the origin—see [12, 16]).

(b) Let $\mathcal{V} = \mathcal{G}(\mathbb{R}^d)$, and consider the supremal multiscale approximation of $\mathcal{V}$ given in Example 3(b). Using (3.8) and (3.11), we get

$$\mathcal{W}_{\sigma,\tau} = \{A \in \mathcal{V} \mid C \bigcirc \sigma B \neq \emptyset \ \text{ and } \ C \bigcirc \tau B = \emptyset \ \forall\, C \in \mathcal{C}(A)\}, \quad \tau > \sigma,$$

for $\sigma \in (0,\infty)$ and $\tau \in (\sigma,\infty)$. Again, $\mathcal{V}_\sigma = \langle\, \mathcal{V}_\tau \cup \mathcal{W}_{\sigma,\tau} \mid \vee \,\rangle$ and $\mathcal{V}_\sigma \cap \mathcal{W}_{\sigma,\tau} = \{\emptyset\}$. Therefore, (3.12) is satisfied so that $\{\mathcal{V}_\sigma,\ \mathcal{W}_{\sigma,\tau} \mid \sigma \in (0,\infty),\ \tau \in (\sigma,\infty)\}$ is a supremal multiscale S-analysis of $\mathcal{V}$, where S is the spatial scaling. The operators

$$\psi_{\sigma,\tau}(A) = \rho(A \mid A \bigcirc \sigma B) \smallsetminus \rho(A \mid A \bigcirc \tau B), \quad \tau > \sigma,$$

are projections on $\mathcal{W}_{\sigma,\tau}$ that satisfy (3.13). Therefore, $\{\psi_{\sigma,\tau} \mid \sigma \in (0,\infty),\ \tau \in (\sigma,\infty)\}$ is a family of detail operators associated with the supremal multiscale analysis. The detail signal $\psi_{\sigma,\tau}(A)$ contains the connected components of A whose "size" is between σ and τ. The resulting supremal multiscale S-analysis scheme is a *discrete size transform* based on openings by reconstruction (see [12, 16] for the notion of the discrete size transform and [10] for such a decomposition). This scheme is illustrated in Figure 4.

The previous examples are binary. In the following, we present an important example of supremal multiscale analysis based on a reconstructive scheme that selectively removes regional maxima from a grayscale signal.

[1] Henri Matisse: *Woman with Amphora and Pomegranates*, 1952—Paper on canvas.

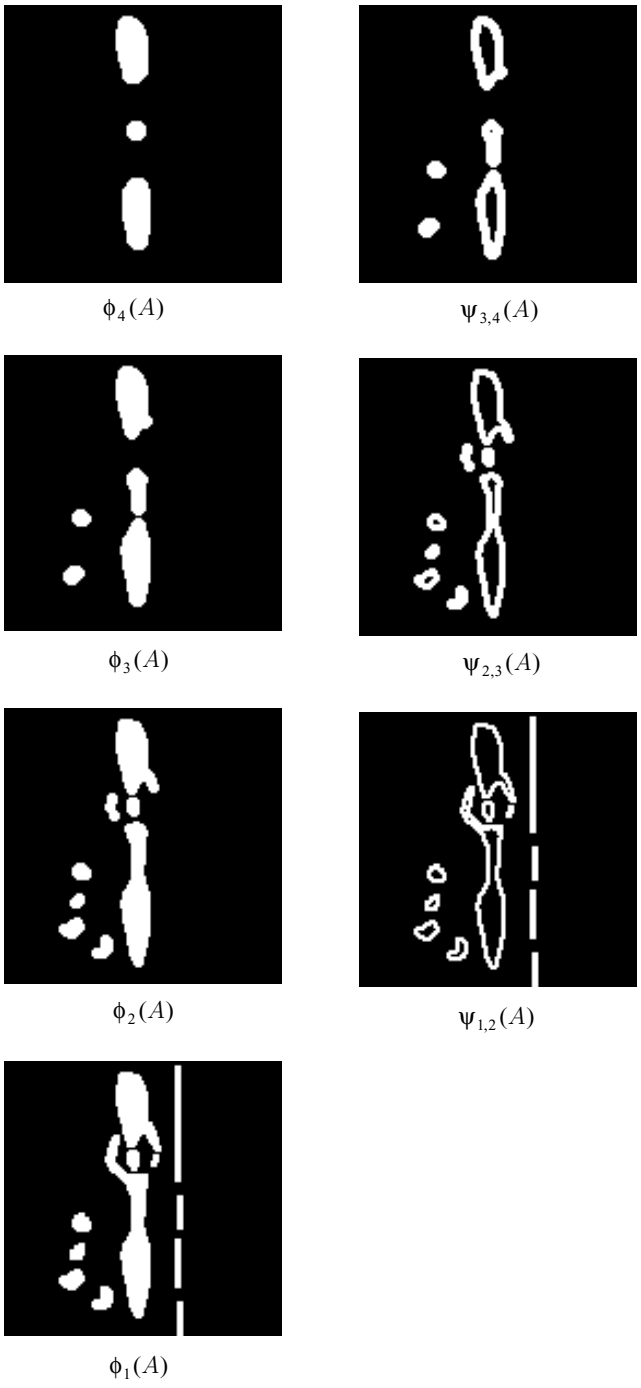

FIG. 3. *An illustration of supremal multiscale analysis of a binary "Matisse" image A depicted in Figure* 2 *based on structural openings. Note that* $\phi_k(A) = \phi_{k+1}(A) \cup \psi_{k,k+1}(A)$*, for* $k = 1, 2, 3$*, and* $\phi_1(A) = \phi_4(A) \cup \psi_{1,2}(A) \cup \psi_{2,3}(A) \cup \psi_{3,4}(A)$*, in accordance with* (3.13) *and* (3.15)*. In this example, B is a disk structuring element of unit radius.*

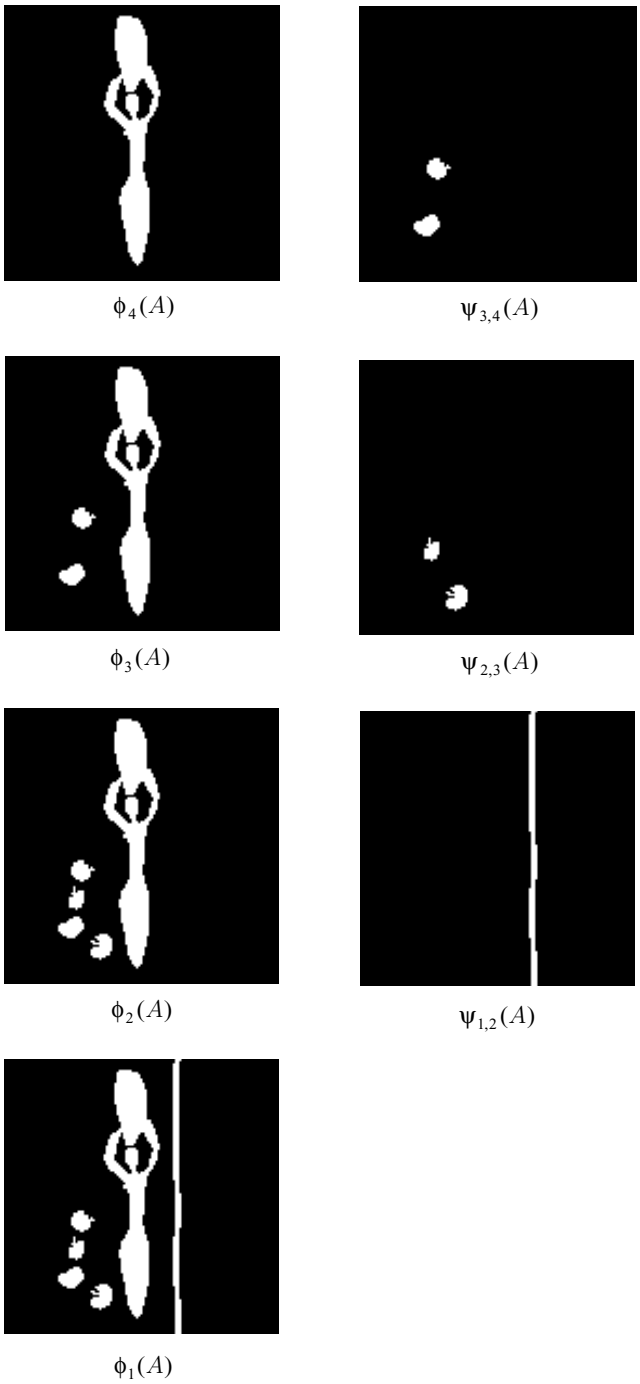

FIG. 4. *An illustration of supremal multiscale analysis of the binary "Matisse" image A depicted in Figure* 2 *based on openings by reconstruction. Note that* $\phi_k(A) = \phi_{k+1}(A) \cup \psi_{k,k+1}(A)$*, for* $k = 1, 2, 3$*, and* $\phi_1(A) = \phi_4(A) \cup \psi_{1,2}(A) \cup \psi_{2,3}(A) \cup \psi_{3,4}(A)$*, in accordance with* (3.13) *and* (3.15)*. In this example, B is a disk structuring element of unit radius.*

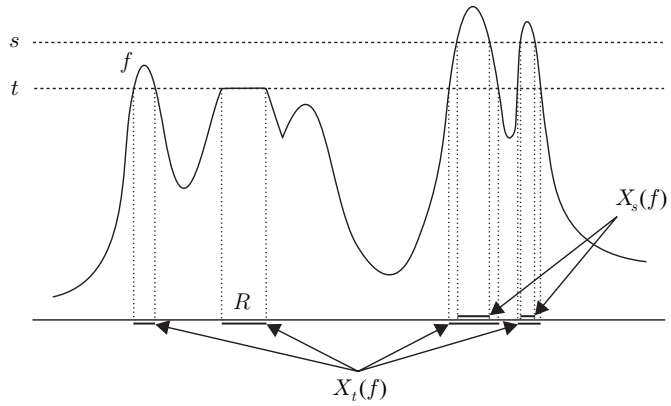

FIG. 5. *A signal f with a regional maximum R at level t. Note that R is a connected component of $X_t(f)$, that $R \cap X_s(f) = \emptyset$, for $s > t$, and that f is constant over R. The usual topological connectivity of the Euclidean real line is assumed.*

4. Skyline supremal multiscale analysis. Recall the lattice $\mathrm{Fun}_u(E, \mathcal{T})$ of u.s.c. functions, discussed in Example 1(e). Here we adopt as the lattice $\mathcal{V}$ of signals of interest the lattice $\mathcal{V} = \mathrm{Fun}_u(E, \overline{\mathbb{R}}_+)$ of nonnegative u.s.c. real-valued functions defined on a topological space E. We are making the following basic assumption.

Assumption 1. We assume that E is a compact Hausdorff space with a countable basis. Moreover, we assume that E is furnished with a connectivity class $\mathcal{C} \subseteq \mathcal{P}(E)$ such that we have the following:

(a) $A \in \mathcal{C}$ implies that $\overline{A} \in \mathcal{C}$ (in this case, the connectivity class $\mathcal{C}$ is said to be *compatible* with the topology of E [6]).
(b) The connectivity openings $\{\gamma_x \mid x \in E\}$, associated with $\mathcal{C}$, are $\downarrow$-continuous operators on $\mathcal{F}(E)$ (i.e., on the collection of all closed subsets of E).
(c) For each $A \in \mathcal{F}(E)$, $\gamma_x(A)$ is an u.s.c. function from A into $\mathcal{F}(E)$.

For example, one may assume E to be a connected, closed, and bounded subset of $\mathbb{R}^d$, with the Euclidean topology, and take $\mathcal{C}$ to be the connectivity class consisting of the usual Euclidean connected subsets of E. It has been shown in [6] that this choice satisfies all conditions stated in Assumption 1.

Next, we give a precise definition of a regional maximum of a signal in $\mathcal{V}$.

DEFINITION 4.1. *A set $R \subseteq E$ is a* regional maximum of $f \in \mathrm{Fun}_u(E, \overline{\mathbb{R}}_+)$ at level $t \in \overline{\mathbb{R}}_+$ *if R is a connected component of $X_t(f)$ and $R \cap X_s(f) = \emptyset$ for all $s > t$.*

Therefore, regional maxima depend on the underlying connectivity assumed. See Figure 5 for an illustration. A regional maximum is always a closed set, since $X_t(f)$ is closed and $\mathcal{C}$ is compatible [7]. It is easy to see that a signal $f \in \mathrm{Fun}_u(E, \overline{\mathbb{R}}_+)$ is constant over a regional maximum R; we denote this constant value by $f(R)$. In addition, we denote by $\mathcal{R}(f)$ the set of all regional maxima of a signal f and by $\mathcal{R}_t(f)$ the set of all regional maxima of f at level t or above; i.e., $\mathcal{R}_t(f) = \{R \in \mathcal{R}(f) \mid f(R) \geq t\}$ for $t \in \overline{\mathbb{R}}_+$.

We have the following result regarding regional maxima.

PROPOSITION 4.2.

(a) *Any function $f \in \mathrm{Fun}_u(E, \overline{\mathbb{R}}_+)$ has at least one regional maximum.*
(b) *A function $f \in \mathrm{Fun}_u(E, \overline{\mathbb{R}}_+)$ has exactly one regional maximum if and only if $X_t(f) \in \mathcal{C}$ for all $t \in \overline{\mathbb{R}}_+$.*

Proof. (a) From Weierstrass's theorem of real analysis [19] and the facts that E is compact and f is an u.s.c. function, f achieves its supremum in E; i.e., there is a

point $x_0 \in E$ such that $f(x_0) = \bigvee\{f(x) \mid x \in E\}$. It is clear that $X_t(f) = \emptyset$ for all $t > f(x_0)$. Hence, $R = \gamma_{x_0}(X_{f(x_0)}(f))$ is a regional maximum of f at level $f(x_0)$.

(b) We show that f has two or more regional maxima if and only if $X_t(f) \notin \mathcal{C}$, for some $t \in \overline{\mathbb{R}}_+$, which is the contrapositive of the assertion. To show the direct implication, assume that R_1 and R_2 are two regional maxima of f. If $f(R_1) = f(R_2) = t$, then $X_t(f) \notin \mathcal{C}$. Otherwise, let $f(R_1) = t_1 > t_2 = f(R_2)$. We have that $R_1 \subseteq X_{t_1}(f) \subseteq X_{t_2}(f)$. But $R_2 \cap X_{t_1}(f) = \emptyset \Rightarrow R_1 \cap R_2 = \emptyset$ so that R_2 must be a strict subset of $X_{t_2}(f)$, which implies that $X_{t_2}(f) \notin \mathcal{C}$. To show the converse implication, assume that $X_t(f) \notin \mathcal{C}$, for some $t \in \overline{\mathbb{R}}_+$, and let C_1 and C_2 be two connected components of $X_t(f)$. Sets C_1 and C_2 are closed subsets of the compact space E; thus C_1 and C_2 are themselves compact [11]. Hence, the restrictions f_1 and f_2 of f to C_1 and C_2, respectively, are u.s.c. functions defined on compact sets so that each achieves its supremum, say, at points $x_1 \in R_1$ and $x_2 \in R_2$. Clearly, the corresponding regional maxima of f_1 and f_2 at $f(x_1)$ and $f(x_2)$, respectively, are distinct regional maxima of f. □

Part (a) of the previous proposition shows that the set $\mathcal{R}(f)$ of regional maxima of f is nonempty, whereas part (b) indicates that the notions of regional maxima and connectivity of level sets are closely related.

Recall from section 2 the grayscale reconstruction operator associated with a connectivity class $\mathcal{C} \subseteq \mathcal{P}(E)$. This operator will be central for our purposes. The next fundamental result shows that a signal $f \in \mathrm{Fun}_u(E, \overline{\mathbb{R}}_+)$ can be "reconstructed" from the grayscale reconstructions of f "marked" by each of its regional maxima. Before that, we need the following definition: A *cylinder* $h_{A,t}$ of base $A \subseteq E$ and height $t \in \overline{\mathbb{R}}_+$ is a function in $\mathrm{Fun}_u(E, \overline{\mathbb{R}}_+)$ defined by

$$h_{A,t}(x) = \begin{cases} t & \text{if } x \in A, \\ 0 & \text{otherwise} \end{cases} \quad \text{for } x \in E.$$

PROPOSITION 4.3. *Let $f \in \mathrm{Fun}_u(E, \overline{\mathbb{R}}_+)$. For each $R \in \mathcal{R}(f)$, we have that $g = \widetilde{\rho}(f \mid h_{R,f(R)}) \in \mathrm{Fun}_u(E, \overline{\mathbb{R}}_+)$, and $\mathcal{R}(g) = \{R\}$, with $g(R) = f(R)$. Moreover,*

$$f = \bigvee_u \{\widetilde{\rho}(f \mid h_{R,f(R)}) \mid R \in \mathcal{R}(f)\}. \tag{4.1}$$

Proof. From the definition of $\widetilde{\rho}$ in (2.2), we can write

$$g(v) = \widetilde{\rho}(f \mid h_{R,f(R)})(v) = \bigvee\{t \in \overline{\mathbb{R}}_+ \mid v \in \rho(\, X_t(f) \mid X_t(h_{R,f(R)}))\}, \quad v \in E. \tag{4.2}$$

Note that $X_t(h_{R,f(R)}) = R$, if $t \leq f(R)$, and $X_t(h_{R,f(R)}) = \emptyset$ if $t > f(R)$. Also, $X_t(f) \cap R = \emptyset$ for $t > f(R)$. Hence, $\rho(X_t(f) \mid X_t(h_{R,f(R)})) = \rho(X_t(f) \mid R)$ for all $t \in \overline{\mathbb{R}}_+$. Moreover, R is connected so that it must be contained in one of the connected components of $X_t(f)$, and, therefore, $\rho(X_t(f) \mid R) = \gamma_x(X_t(f))$ for some $x \in R$. Thus, (4.2) becomes $g(v) = \bigvee\{t \in \overline{\mathbb{R}}_+ \mid v \in \gamma_x(X_t(f))\}$ for $v \in E$. Hence, $X_t(g) = \bigcap_{s<t} \gamma_x(X_s(f)) = \gamma_x\left(\bigcap_{s<t} X_s(f)\right) = \gamma_x(X_t(f))$, for all $t \in \overline{\mathbb{R}}_+$, from the $\downarrow$-continuity of γ_x on $\mathcal{F}(E)$ and (2.1). In other words, $X_t(g)$ is a closed (by the compatibility of $\mathcal{C}$) connected set, for all $t \in \overline{\mathbb{R}}_+$, so that, by Proposition 4.2(b), g is u.s.c. and has a single regional maximum. In addition, we have that $X_t(g) = R$, for $t = f(R)$, and $X_t(g) = \emptyset$, for $t > f(R)$, so that R is the only regional maximum of g at level $g(R) = f(R)$. This shows the first part of the result. Note that the right-hand side of (4.1) makes sense, since $\widetilde{\rho}(f \mid h_{R,f(R)})$ is a function in $\mathrm{Fun}_u(E, \overline{\mathbb{R}}_+)$ for each $R \in \mathcal{R}(f)$. Let C be a connected component of any nonempty level set $X_t(f)$ of f. It

follows from the fact that any closed subset of a compact space is compact and from the compatibility of $\mathcal{C}$ that C is compact. In addition, the restriction of f to C is an u.s.c. function; hence, C contains some regional maximum $R \in \mathcal{R}_t(f)$. Moreover, the definition of regional maximum implies that each $R \in \mathcal{R}_t(f)$ must be contained in some component C of $X_t(f)$. Since $X_t(f)$ equals the union of its components, we conclude that $X_t(f) = \bigcup_{R\in\mathcal{R}_t(f)} \rho(X_t(f) \,|\, R)$. But, by definition, any $R \in \mathcal{R}(f) \setminus \mathcal{R}_t(f)$ does not intersect $X_t(f)$. Hence, $X_t(f) = \bigcup_{R\in\mathcal{R}(f)} \rho(X_t(f) \,|\, R)$. In addition, from our previous discussion, we have that $X_t(\,\widetilde{\rho}(f \mid h_{R,f(R)})) = \rho(X_t(f) \,|\, R)$ for all $t \in \overline{\mathbb{R}}_+$. It follows from the last two equations and the fact $X_t(f) = \bigcap_{s<t} \overline{\bigcup X_s(f_\alpha)}$ that [7]

$$\begin{aligned} X_t\left(\bigvee_u \{\widetilde{\rho}(f \mid h_{R,f(R)}) \mid R \in \mathcal{R}(f)\}\right) &= \bigcap_{s<t} \overline{\bigcup_{R\in\mathcal{R}(f)} X_t(\,\widetilde{\rho}(f \mid h_{R,f(R)}))} \\ &= \bigcap_{s<t} \overline{\bigcup_{R\in\mathcal{R}(f)} \rho(X_s(f) \,|\, R)} \\ &= \bigcap_{s<t} \overline{X_s(f)} = \bigcap_{s<t} X_s(f) = X_t(f), \end{aligned}$$

for all $t \in \overline{\mathbb{R}}_+$, which implies (4.1). □

Now consider the subsets $\mathcal{V}_\sigma$ of $\mathcal{V}$ given by

$$\mathcal{V}_\sigma = \{O\} \cup \{f \in \mathcal{V} \mid \mathcal{R}(f) = \mathcal{R}_\sigma(f)\}, \quad \sigma \in (0,\infty). \tag{4.3}$$

In other words, $\mathcal{V}_\sigma$ consists of the least signal O and all signals whose regional maxima are at level σ or above. The following is a fundamental result for our purposes.

PROPOSITION 4.4. *The space $\mathcal{V}_\sigma$ is sup-closed in* $\mathrm{Fun}_u(E, \overline{\mathbb{R}}_+)$ *for* $\sigma \in (0,\infty)$.

Proof. First, note that $\bigvee \emptyset = O \in \mathcal{V}_\sigma$ for every $\sigma \in (0,\infty)$. For a given $\sigma \in (0,\infty)$, let $\{f_\alpha\}$ be a family of functions in $\mathrm{Fun}_u(E, \overline{\mathbb{R}}_+)$ such that $\{f_\alpha\} \subseteq \mathcal{V}_\sigma$. We can assume, without loss of generality, that $f_\alpha \neq O$ for all α. Hence, $\mathcal{R}_\sigma(f_\alpha) = \mathcal{R}(f_\alpha) \neq \emptyset$, for each f_α, which implies that $X_t(f_\alpha) \neq \emptyset$ for all $t \leq \sigma$. Let $f = \bigvee_u f_\alpha$. We have that $X_t(f) = \bigcap_{s<t} \overline{\bigcup X_s(f_\alpha)} \supseteq \bigcup X_t(f_\alpha)$ [7]. Therefore, $X_t(f) \neq \emptyset$ for all $t \leq \sigma$. Suppose that R is a regional maximum of f at a level $r < \sigma$. By definition, we have that $R \cap X_t(f) = \emptyset$ for all $t > r$. Therefore, the sets R and $T = X_\sigma(f)$ are closed nonempty disjoint sets. Moreover, since E is a compact Hausdorff space, there exist disjoint open sets U and V such that $R \subset U$ and $T \subset V$ [11]. Now, given $x \in R$, we have that $R = \gamma_x(X_r(f)) = \gamma_x(\bigcap_{s<r} \overline{\bigcup X_s(f_\alpha)}) = \bigcap_{s<r} \gamma_x(\overline{\bigcup X_s(f_\alpha)})$ from the $\downarrow$-continuity of γ_x on $\mathcal{F}(E)$ and (2.1). Let $C(s) = \gamma_x(\overline{\bigcup X_s(f_\alpha)})$ for $s < r$. Note that $\{C(s)\}_{s<r}$ is a decreasing family of nonempty closed sets in the compact space E, and $\bigcap_{s<r} C(s) \subset U$. It follows that there is some $p < r$ such that $C(p) \subset U$ [7, Prop. 2.3.7]. Since $\gamma_x(A)$ is an u.s.c. function from A into $\mathcal{F}(E)$, we can apply Proposition 4.1.14 in [7] to conclude that there is some connected component C of $\bigcup X_p(f_\alpha)$ such that $C \subset U$. Clearly, this implies that there is some index α' such that a connected component C' of $X_p(f_{\alpha'})$ is contained in U. This follows from the fact that each component of $\bigcup A_\alpha$ must contain at least one component of some $A_{\alpha'}$. However, note that $T = X_\sigma(f) \supseteq \bigcup X_\sigma(f_\alpha)$ implies that $X_\sigma(f_\alpha) \subset V$ for all α. Hence, $C' \cap X_\sigma(f_{\alpha'}) = \emptyset$ so that function $f_{\alpha'}$ has a regional maximum inside C' at some level below t, which is a contradiction. Therefore, $f = \bigvee_u f_\alpha$ must not have any regional maxima below level σ; i.e., $f \in \mathcal{V}_\sigma$, as required. □

We can now use the previous result to show that the family $\{\mathcal{V}_\sigma \mid \sigma \in (0,\infty)\}$ is a supremal multiscale approximation of $\mathcal{V}$.

PROPOSITION 4.5. *The family* $\{\mathcal{V}_\sigma \mid \sigma \in (0,\infty)\}$*, given by* (4.3)*, is a supremal multiscale* S*-approximation of* $\mathcal{V}$ *for the gray-level scaling* $S = \{tf(\cdot) \mid t \in (0,\infty)\}$.

Proof. From Proposition 4.4, $\mathcal{V}_\sigma$ is sup-closed in $\mathcal{V}$ for each $\sigma \in (0,\infty)$. In addition, (3.2) is clearly satisfied, whereas (3.4) is a direct consequence of the fact that $X_\tau(f) = X_{t\tau}(tf)$, which implies that R is a regional maximum of f at level τ if and only if R is a regional maximum of tf at level $t\tau$. To show (3.3), note that Proposition 4.3 implies that, for a given $f \in \mathcal{V}$, $\widetilde{\rho}(f \mid h_{R,f(R)}) \in \mathcal{V}_{f(R)}$ for each $R \in \mathcal{R}(f)$. Moreover, it implies that $f = \bigvee_u\{\widetilde{\rho}(f \mid h_{R,f(R)}) \mid R \in \mathcal{R}(f)\} \in \langle \bigcup \mathcal{V}_\sigma \mid \vee_u \rangle$, from which we obtain the desired result. ∎

The next result provides an expression for the associated approximation operators.

PROPOSITION 4.6. *Let* $\{\mathcal{V}_\sigma \mid \sigma \in (0,\infty)\}$ *be the supremal multiscale approximation of* $\mathcal{V}$*, given by* (4.3)*. The associated approximation operators are given by*

$$\phi_\sigma(f) = \bigvee_u\{\,\widetilde{\rho}(f \mid h_{R,f(R)}) \mid R \in \mathcal{R}_\sigma(f)\}, \quad f \in \mathcal{V}, \quad \sigma \in \overline{\mathbb{R}}_+. \tag{4.4}$$

Proof. Let $\sigma \in (0,\infty)$, and consider the operator $\theta(f) = \bigvee_u\{\,\widetilde{\rho}(f \mid h_{R,f(R)}) \mid R \in \mathcal{R}_\sigma(f)\}$ for $f \in \mathcal{V}$. Note that Proposition 4.3 guarantees that θ is an operator on $\mathcal{V}$. We show that $\phi_\sigma(f) = \bigvee_u\{g \in \mathcal{V}_\sigma \mid g \leq f\} = \theta(f)$ for $f \in \mathcal{V}$. First, we show that θ is an increasing operator. Let $f, g \in \mathcal{V}$ such that $f \leq g$. Consider a regional maximum $R \in \mathcal{R}_\sigma(f)$ at level $t = f(R)$. Since $R \in \mathcal{C}$ and $R \subseteq X_t(f) \subseteq X_t(g)$, we must have that $R \subseteq C$ for some connected component C of $X_t(g)$. As argued in the proof of Proposition 4.3, there is a regional maximum $R' \in \mathcal{R}_\sigma(g)$ such that $R' \subseteq C$. For any $s \leq t$, it is clear that $\rho(X_s(g) \mid R) = \rho(X_s(g) \mid R')$, since both R and R' are contained in the same connected component of $X_s(g)$ that contains C. This implies that $X_s(\,\widetilde{\rho}(f \mid h_{R,f(R)})) = \rho(X_s(f) \mid R) \subseteq \rho(X_s(g) \mid R) = \rho(X_s(g) \mid R') = X_s(\,\widetilde{\rho}(g \mid h_{R',g(R')}))$, for all $s \leq f(R)$, where we have used the fact that $\rho(\cdot \mid R)$ is an opening and thus is increasing. Since $X_s(\,\widetilde{\rho}(f \mid h_{R,f(R)})) = \emptyset$, for $s > f(R)$, we conclude that $\widetilde{\rho}(f \mid h_{R,f(R)}) \leq \widetilde{\rho}(g \mid h_{R',g(R')})$. This implies that $\theta(f) \leq \theta(g)$ so that θ is increasing. Now let $f \in \mathcal{V}$. If $\mathcal{R}_\sigma(f) = \emptyset$, then clearly $\theta(f) = \phi_\sigma(f) = O$. Hence, we can assume that $\mathcal{R}_\sigma(f) \neq \emptyset$. We have that $\phi_\sigma(f) \in \mathcal{V}_\sigma$; hence $\mathcal{R}_\sigma(\phi_\sigma(f)) = \mathcal{R}(\phi_\sigma(f))$. It follows from Proposition 4.3 that $\phi_\sigma(f) = \theta(\phi_\sigma(f))$. But, since θ is increasing and ϕ_σ is antiextensive, we have that $\theta(\phi_\sigma(f)) \leq \theta(f)$. Therefore, $\phi_\sigma(f) \leq \theta(f)$. To show the converse inequality, note that Proposition 4.3 implies that $\widetilde{\rho}(f \mid h_{R,f(R)}) \in \mathcal{V}_\sigma$ for each $R \in \mathcal{R}_\sigma(f)$. Since $\mathcal{V}_\sigma$ is sup-closed, we must have $\theta(f) \in \mathcal{V}_\sigma$. Combined with the fact that $\theta(f) \leq f$, this implies that $\theta(f) \leq \phi_\sigma(f)$. Hence, $\phi_\sigma(f) = \theta(f)$. ∎

Given a signal $f \in \mathcal{V}$, its approximation $\phi_\sigma(f)$, obtained from f by means of (4.4), preserves the regional maxima of f that are at level σ or above, while it flattens the rest. As the scale σ increases, only the highest peaks in the signal survive. In this scale-space, evolution towards decreasing levels of detail is akin to viewing a city skyline as one moves away from it: near the city, the shorter buildings are visible, but far away only the tallest buildings can be discerned. This is illustrated in Figure 6. For this reason, we refer to this scheme as a *skyline supremal multiscale approximation*, whereas the associated scale-space is referred to as a *skyline supremal scale-space*.

In addition to being grayscale, translation, and scale invariant, the most striking property of the skyline supremal scale-space is that, by construction, it decomposes the regional maxima of a function f in a natural causal hierarchy. As σ increases, the scaling operator ϕ_σ removes regional maxima from f without introducing new ones. Moreover, as σ increases, the scaling operator ϕ_σ progressively removes connected

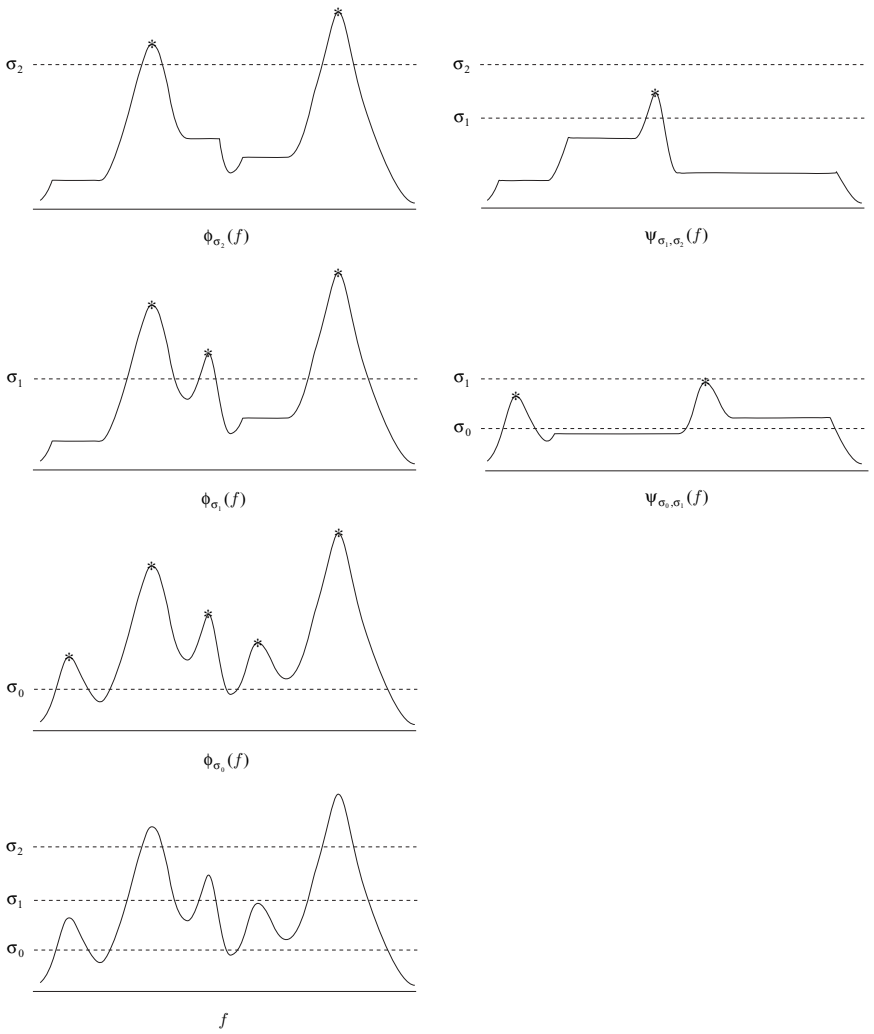

FIG. 6. *Skyline supremal multiscale analysis of a one-dimensional signal f. Note that the scaling $\phi_\sigma(f)$ preserves the regional maxima (depicted by $*$) of f that are at level σ or above, while it flattens the rest. Moreover, the detail signal $\psi_{\sigma,\tau}(f)$ preserves the regional maxima of f with values in $[\sigma,\tau)$ and flattens the rest. Finally, $\phi_{\sigma_k}(f) = \phi_{\sigma_{k+1}}(f) \vee \psi_{\sigma_k,\sigma_{k+1}}(f)$, for $k = 0, 1$, and $\phi_{\sigma_0}(f) = \phi_{\sigma_2}(f) \vee \psi_{\sigma_0,\sigma_1}(f) \vee \psi_{\sigma_1,\sigma_2}(f)$, in accordance with* (3.13) *and* (3.15)*, respectively.*

components from the level sets $X_t(f)$ of f without introducing new ones. These properties are much desired by any useful scale-space scheme [3, 18, 21, 22, 41].

We now derive the corresponding supremal multiscale analysis. From (3.11), (4.3), and (4.4), we have that

$$\mathcal{W}_{\sigma,\tau} = \{O\} \cup \{f \in \mathcal{V} \mid \mathcal{R}(f) = \mathcal{R}_\sigma(f) \smallsetminus \mathcal{R}_\tau(f)\}, \quad \tau > \sigma.$$

It is easy to check that (3.12) is satisfied so that $\{\mathcal{V}_\sigma, \mathcal{W}_{\sigma,\tau} \mid \sigma \in (0,\infty), \tau \in (\sigma,\infty)\}$ is a supremal multiscale S-analysis of $\mathcal{V}$, where S is the gray-level scaling. This is referred to as the *skyline supremal multiscale analysis* of $\mathcal{V}$. The operators

$$\psi_{\sigma,\tau}(f) = \bigvee_u \{\widetilde{\rho}(f \mid h_{R,f(R)}) \mid R \in \mathcal{R}_s(f) \smallsetminus \mathcal{R}_\tau(f)\}, \quad \tau > \sigma,$$

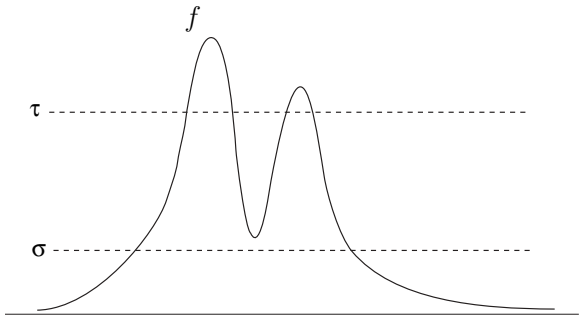

FIG. 7. *The function $f \in \mathcal{V}_\sigma$ is level-σ connected, but it is not level-τ connected. The usual topological connectivity of the Euclidean real line is assumed for the underlying binary connectivity class $\mathcal{C}$.*

are projections on $\mathcal{W}_{\sigma,\tau}$. Therefore, these are detail operators associated with the skyline supremal multiscale analysis scheme. The detail signal $\psi_{\sigma,\tau}(f)$ preserves the regional maxima of f with values in $[\sigma, \tau)$ and flattens the rest. This is illustrated in Figure 6.

From our previous discussion, it is clear that (4.3) satisfies (3.2) and (3.3), regardless of the choice of scaling. Whether or not (3.4) is satisfied depends on the choice of scaling and the choice of the connectivity class $\mathcal{C}$, since the concept of regional maximum depends on the underlying connectivity class. In the case of gray-level scaling, our results hold true for any choice of connectivity class $\mathcal{C}$. However, in the cases of spatial and umbral scalings, the results are valid if $\mathcal{C}$ is invariant to spatial scalings, i.e., if $A \in \mathcal{C} \Leftrightarrow tA \in \mathcal{C}$, for all $A \in \mathcal{F}(E)$ and $t \in (0, \infty)$. Topological connectivity clearly satisfies this property.

Additional insight can be gained by realizing that each approximation space $\mathcal{V}_\sigma$ constitutes a complete lattice, under the partial order of $\mathcal{V}$, with supremum $\bigvee^\sigma$ and infimum $\bigwedge^\sigma$, given by

$$\bigvee\nolimits^\sigma f_\alpha = \bigvee_u f_\alpha,$$

$$\bigwedge\nolimits^\sigma f_\alpha = \phi_\sigma\Big(\bigwedge f_\alpha\Big) = \bigvee_u \Big\{\widetilde{\rho}\Big(\bigwedge f_\alpha \mid h_{R,(\bigwedge f_\alpha)(R)}\Big) \mid R \in \mathcal{R}_\sigma\Big(\bigwedge f_\alpha\Big)\Big\}.$$

In this framework, $\bigwedge^\sigma f_\alpha = O$ if and only if $\bigwedge f_\alpha$ has no regional maxima at level σ or above. Hence, even if the signals $\{f_\alpha\}$ have nonzero pointwise infimum, they can still have zero infimum in $\mathcal{V}_\sigma$.

It has been shown in [7] that the family

$$\mathcal{S}_\sigma = \{\delta_{v,t} \mid t \geq \sigma\} \cup \{f \in \mathcal{V} \mid \mathcal{R}(f) = \{R\},\ f(R) = \sigma\}$$

is sup-generating in $\mathcal{V}_s$. Moreover, assuming this sup-generating family, we can define a connectivity class $\mathcal{C}_\sigma$ on $\mathcal{V}_\sigma$, given by [7]

$$\mathcal{C}_\sigma = \{f \in \mathcal{V}_\sigma \mid X_t(f) \in \mathcal{C}\ \forall\, t \leq \sigma\}.$$

We call this the *level-σ connectivity class*. In this framework, a function $f \in \mathcal{V}_\sigma$ is level-σ connected if all level sets below level σ are connected, according to the connectivity class $\mathcal{C}$. Loosely speaking, this means that f is not allowed to have any "disconnecting dips" below level σ. See Figure 7 for an illustration.

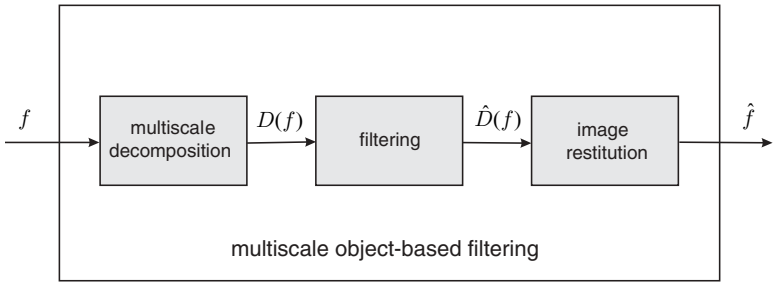

FIG. 8. *Block diagram for multiscale object-based filtering.*

The connected components of a function $f \in \mathcal{V}_\sigma$ are associated with the regional maxima of f contained in the connected components of the level set $X_\sigma(f)$. For each $C \in \mathcal{C}(X_\sigma(f))$, there corresponds a grayscale level-σ connected component f_C of f, given by

$$f_C = \bigvee\nolimits_u \{\widetilde{\rho}(f \mid h_{R,f(R)}) \mid R \in \mathcal{R}_\sigma(f) \text{ and } R \subseteq C\}.$$

Therefore, we can write the approximation signal $\phi_\sigma(f)$ as the supremum of grayscale connected components; i.e.,

$$\phi_\sigma(f) = \bigvee\nolimits_u \{f_C \mid C \in \mathcal{C}(X_\sigma(f))\}, \quad \sigma \in (0, \infty).$$

These grayscale connected components are "mutually disjoint," in the sense that, for $\alpha \neq \beta$, we have that $\phi_\sigma(f_{C_\alpha} \wedge f_{C_\beta}) = O$, which says that the infimum $f_{C_\alpha} \wedge f_{C_\beta}$ has no regional maxima above level σ. This is similar to the linear case, in which the orthogonal projection of a function f over a linear approximation space $\mathcal{V}_\sigma$ is obtained with an expansion in terms of the orthogonal scaling basis [25].

5. Multiscale object-based filtering. Several image processing and analysis tasks are geared towards identifying objects of interest and manipulating those objects to achieve a desired result. For example, if we want to remove certain objects from a scene, we should first identify those objects and then extract them from the scene with operators that do not affect other objects. This task is referred to as *object-based filtering* and can be effectively implemented by the three-step multiscale approach depicted in Figure 8. The first step performs a multiscale decomposition of an image f into a finite collection $D(f) = \{f_1, f_2, \ldots, f_N\}$ of images that contain objects of interest in f at various scales such that f can be uniquely reconstructed from $D(f)$. The images in $D(f)$ are then processed individually by the filtering step. This produces a new multiscale decomposition $\widehat{D}(f)$, which is then used to restitute the filtered image $\hat{f}$. Note that $D(\hat{f}) = \widehat{D}(f)$.

In this paper, we assume that objects of interest are identified by their intensity distribution and, more precisely, by the regional maxima of such intensities. Moreover, we assume that regional maxima associated with similar objects have similar values. In this case, we are interested in a technique that identifies the regional maxima of an image f and decomposes f into a finite collection $D(f) = \{f_1, f_2, \ldots, f_N\}$, with each image f_k containing all regional maxima of f with similar values, such that f can be uniquely reconstructed from $D(f)$. This naturally leads to the previously discussed skyline supremal multiscale analysis scheme.

We specify a finite collection $\{\sigma_k \mid k \in I\}$ of scales, where $I = \{0, 1, \ldots, N\}$, such that $\phi_{\sigma_0}(f) = f$ and $\sigma_k < \sigma_{k+1}$, and decompose the grayscale image f into the

collection $D(f) = \{\psi_{\sigma_k,\sigma_{k+1}}(f) \mid k \in I\}$, where $\psi_{\sigma_N,\sigma_{N+1}}(f) \triangleq \phi_{\sigma_N}(f)$. The image f can be uniquely reconstructed from such decomposition, since (recall (3.15))

$$f = \phi_{\sigma_0}(f) = \bigvee_{k\in I} \psi_{\sigma_k,\sigma_{k+1}}(f).$$

Recall that $\psi_{\sigma_k,\sigma_{k+1}}(f)$ contains the regional maxima of f with values in $[\sigma_k, \sigma_{k+1})$ with all other regional maxima suppressed (flattened), whereas $\phi_{\sigma_N}(f)$ contains the regional maxima of f that are above level σ_N with all other regional maxima suppressed.

During the filtering step, a subset $J \subseteq I$ is determined, and then the images $\{\psi_{\sigma_j,\sigma_{j+1}}(f) \mid j \in J\}$ are processed to produce a new collection $\{\widehat{\psi}_{\sigma_j,\sigma_{j+1}}(f) \mid j \in J\}$. The output of the filtering step depicted in Figure 8 is given by $\widehat{D}(f) = \{\psi_{\sigma_k,\sigma_{k+1}}(f), \widehat{\psi}_{\sigma_j,\sigma_{j+1}}(f) \mid k \in I \smallsetminus J, j \in J\}$, and the new filtered image $\widehat{f}$ is obtained by means of

$$\widehat{f} = \bigvee_{k\in I\smallsetminus J} \psi_{\sigma_k,\sigma_{k+1}}(f) \vee \bigvee_{j\in J} \widehat{\psi}_{\sigma_j,\sigma_{j+1}}(f).$$

We illustrate the previous filtering approach with two examples. Figure 9(a) depicts a grayscale MRI "tumor" image f that contains several objects, including a large tumor on the right-hand side and a small tumor slightly above it.[2] Our objective is to extract the tumors and place them on two different image frames. Moreover, we would like to enhance their presence by flattening surrounding details. We set $\sigma_k = k + 1$, for $k = 0, 1, \ldots, N - 1$, where N is the maximum grayscale value in f (in this case, $N = 255$). The skyline supremal multiscale decomposition of the "tumor" image f reveals that most information related to the small tumor is contained in the detail images $\psi_{k,k+1}(f)$, $153 \leq k \leq 173$, whereas most information related to the large tumor is contained in the detail images $\psi_{k,k+1}(f)$, $206 \leq k \leq 212$. This observation leads to a "filtering" step in Figure 8 that preserves the previous detail images and sets the rest equal to zero. The images $\hat{f}$, obtained by the "restitution" step of Figure 8, are depicted in Figures 9(b) and (c). The results indicate that, as expected, the skyline supremal multiscale decomposition scheme successfully extracts the two tumors and flattens surrounding details.

Figure 10 depicts a grayscale "boat" image f that has been corrupted by "pepper" noise. The noise consists of black spots (that may be more than one pixel thick), which are randomly distributed over the entire image. Our objective is to remove the noise from the image depicted in Figure 10(b) and recover a sufficiently good approximation of the original image depicted in Figure 10(a). This is the classical problem of *image denoising*.

As before, we set $\sigma_k = k + 1$, for $k = 0, 1, \ldots, N - 1$, where $N = 255$. Skyline supremal multiscale analysis of the noisy "boat" image f depicted in Figure 10(b) reveals that most information related to noise is contained in image $\phi_N(N - f)$, since the black spots in f show as narrow bright peaks of amplitude N in the negative image $N - f$. This observation leads to a "filtering" step in Figure 8 that preserves all detail images but replaces $\phi_N(N - f)$ with its grayscale reconstruction $\widetilde{\rho}(\phi_N(N - f) \mid \phi_N(N - f) \bigcirc B)$, where B is a disk structuring element of radius 4. The structural opening $\phi_N(N - f) \bigcirc B$ removes most peaks in $\phi_N(N - f)$ due to noise and provides

[2]The image is courtesy of Christos Davatzikos.

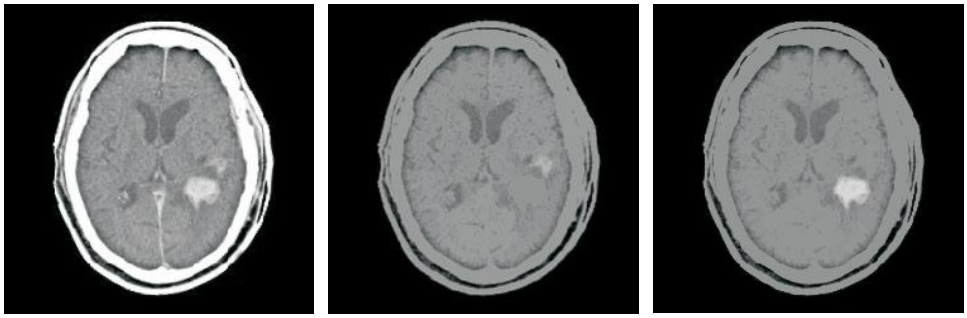

FIG. 9. (a) *A grayscale MRI "tumor" image.* (b) *Extraction of the small tumor and flattening of surrounding details.* (c) *Extraction of the large tumor and flattening of surrounding details.*

a marker for the reconstruction of the "noise-free" part of $\phi_N(N - f)$. The image $\hat{f}$, obtained by subtracting the result of the "restitution" step of Figure 8 from N, is depicted in the first row of Figure 11(a).

On the other hand, the first row of Figure 11(b) depicts the result obtained from a conventional morphological denoising approach that subtracts the grayscale reconstruction $\widetilde{\rho}(N - f \mid (N - f) \bigcirc B)$, applied on the negative noisy image $N - f$, from N. Although, at first glance, the two results seem to be similar, the details depicted in the second row of Figure 11 reveal that they are different in quality. Although noise has been equally suppressed in both cases, the result depicted in Figure 11(b) shows that direct application of grayscale reconstruction on the noisy image may result in excessive smoothing of important features (e.g., the masts and the letters on the stern). Clearly, a denoising approach based on skyline supremal multiscale analysis is more preferable in this case.

6. Conclusion. In this paper, we have presented a new approach to nonlinear multiscale signal analysis. The proposed scheme is related to the concept of supremal scale-spaces, introduced by Heijmans and van den Boomgaard, and is referred to as supremal multiscale analysis. To develop this approach, we have extended (among other things) the concepts of (orthogonal) vector spaces, (orthogonal) projections, and linear operators to a nonlinear setting. We have accomplished this by employing the theory of complete lattices in conjunction with mathematical morphology and by replacing numerical addition with supremum. We have also proposed a particular supremal multiscale analysis scheme that is based on morphological reconstruction operators. This approach, which is referred to as skyline supremal multiscale analysis, decomposes the regional maxima of a signal in a natural causal hierarchy by gradually removing these maxima without introducing new ones. More precisely, the skyline supremal multiscale analysis scheme represents a signal as the supremum of a coarse approximation and details. The coarse approximation preserves the regional maxima above some level σ, while it flattens the rest. On the other hand, the details preserve regional maxima with values in nonoverlapping subintervals of $(0, \sigma)$ and flatten the rest. We show that this scheme is grayscale, translation, and scale invariant, and it progressively removes connected components from the level sets of a signal without introducing new ones. We believe that skyline supremal multiscale analysis can be effectively used for multiscale signal decomposition, representation, and analysis.

FIG. 10. (a) *An original grayscale "boat" image.* (b) *A noisy copy of the image depicted in* (a).

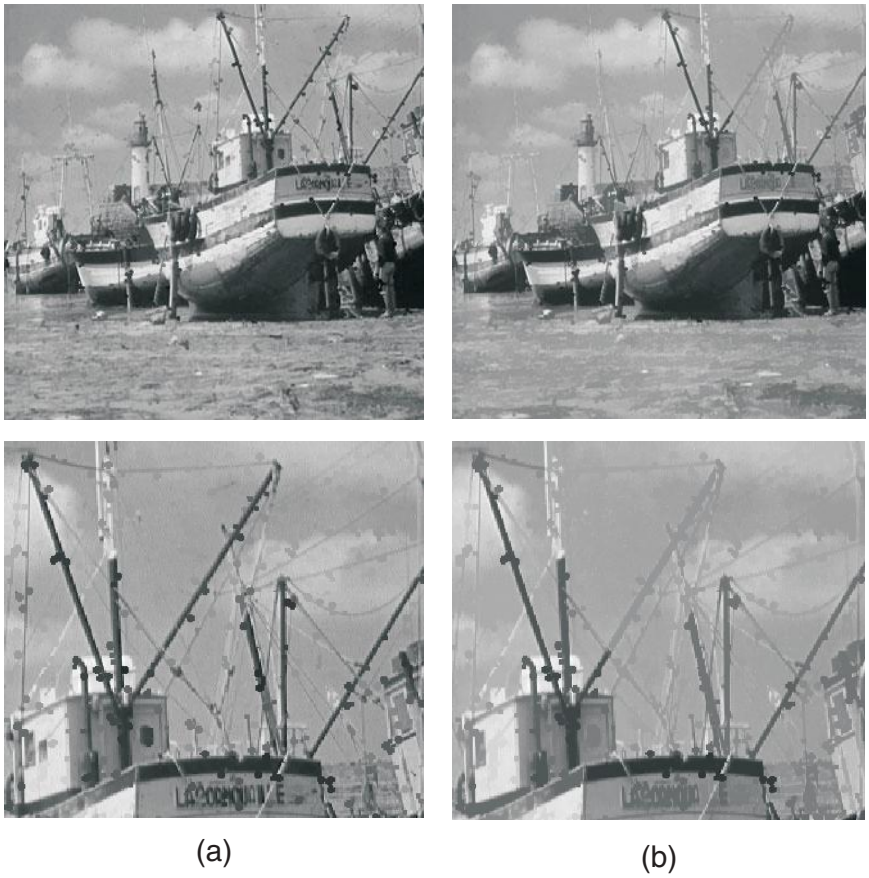

FIG. 11. *Denoising results obtained:* (a) *by skyline supremal multiscale analysis and grayscale reconstruction of* $\phi_N(N-f)$ *from its structural opening* $\phi_N(N-f)\circ B$, *and* (b) *by grayscale reconstruction of* $N-f$ *from its structural opening* $(N-f)\circ B$.

REFERENCES

[1] L. Alvarez, F. Guichard, P.-L. Lions, and J.-M. Morel, *Axioms and fundamental equations of image processing*, Arch. Rational Mech. Anal., 123 (1993), pp. 199–257.

[2] L. Alvarez and J.-M. Morel, *Formalization and computational aspects of image analysis*, in Acta Numerica, Cambridge University Press, Cambridge, UK, 1994, pp. 1–59.

[3] J. A. Bangham, P. D. Ling, and R. Harvey, *Scale-space from nonlinear filters*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 18 (1996), pp. 520–528.

[4] G. Birkhoff, *Lattice Theory*, 3rd ed., Amer. Math. Soc. Colloq. Publ. 25, AMS, Providence, RI, 1967.

[5] U. M. Braga-Neto and J. Goutsias, *Connectivity on complete lattices: New results*, Computer Vision and Image Understanding, 85 (2002), pp. 22–53.

[6] U. M. Braga-Neto and J. Goutsias, *A theoretical tour of connectivity in image processing and analysis*, J. Math. Imaging Vision, 19 (2003), pp. 5–31.

[7] U. M. Braga-Neto, *Connectivity in Image Processing and Analysis: Theory, Multiscale Extensions, and Applications*, Ph.D. thesis, Center for Imaging Science and Department of Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, MD, 2001.

[8] R. W. Brockett and P. Maragos, *Evolution equations for continuous-scale morphological filtering*, IEEE Trans. Signal Process., 42 (1994), pp. 3377–3386.

[9] M.-H. Chen and P.-F. Yan, *A multiscaling approach based on morphological filtering*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 11 (1989), pp. 694–700.

[10] A. Doulamis, N. Doulamis, and P. Maragos, *Generalized multiscale connected operators with applications to granulometric image analysis*, in Proceedings of the International Conference on Image Processing, Thessaloniki, Greece, 2001, pp. 684–687.

[11] J. Dugundji, *Topology*, Allyn and Bacon, Boston, MA, 1966.

[12] J. Goutsias and S. Batman, *Morphological methods for biomedical image analysis*, in Handbook of Medical Imaging. Volume 2. Medical Image Processing and Analysis, M. Sonka and J. M. Fitzpatrick, eds., SPIE Press, Bellingham, WA, 2000, pp. 175–272.

[13] J. Goutsias and H. J. A. M. Heijmans, *Nonlinear multiresolution signal decomposition schemes – Part* I: *Morphological pyramids*, IEEE Trans. Image Process., 9 (2000), pp. 1862–1876.

[14] H. J. A. M. Heijmans and J. Goutsias, *Nonlinear multiresolution signal decomposition schemes – Part* II: *Morphological wavelets*, IEEE Trans. Image Process., 9 (2000), pp. 1897–1913.

[15] H. J. A. M. Heijmans and R. van den Boomgaard, *Algebraic framework for linear and morphological scale-spaces*, Journal of Visual Communication and Image Representation, 13 (2002), pp. 269–301.

[16] H. J. A. M. Heijmans, *Morphological Image Operators*, Academic Press, Boston, MA, 1994.

[17] P. T. Jackway and M. Deriche, *Scale-space properties of the multiscale morphological dilation-erosion*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 18 (1996), pp. 38–51.

[18] J. J. Koenderink, *The structure of images*, Biol. Cybernet., 50 (1984), pp. 363–370.

[19] A. Kolmogorov and S. Fomin, *Introductory Real Analysis*, Dover, New York, 1975.

[20] L. M. Lifshitz and S. M. Pizer, *A multiresolution hierarchical approach to image segmentation based on intensity extrema*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 12 (1990), pp. 529–540.

[21] T. Lindeberg, *Scale-space theory: A basic tool for analysing structures at different scales*, J. Appl. Stat., 21 (1994), pp. 225–270.

[22] T. Lindeberg, *Scale-Space Theory in Computer Vision*, Kluwer Academic Publishers, Boston, MA, 1994.

[23] S. G. Mallat, *Multiresolution approximations and wavelet orthonormal bases of* $L^2(\mathbb{R})$, Trans. Amer. Math. Soc., 315 (1989), pp. 69–87.

[24] S. G. Mallat, *A theory for multiresolution signal decomposition: The wavelet representation*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 11 (1989), pp. 674–693.

[25] S. Mallat, *A Wavelet Tour of Signal Processing*, 2nd ed., Academic Press, San Diego, CA, 1999.

[26] P. Maragos and R. W. Schafer, *Morphological filters - Part* I: *Their set-theoretic analysis and relations to linear shift-invariant filters*, IEEE Transactions on Acoustics, Speech, and Signal Processing, 35 (1987), pp. 1153–1169.

[27] P. Maragos and R. W. Schafer, *Morphological systems for multidimensional signal processing*, Proceedings of the IEEE, 78 (1990), pp. 690–710.

[28] P. Maragos, *Pattern spectrum and multiscale shape representation*, IEEE Transactions on

Pattern Analysis and Machine Intelligence, 11 (1989), pp. 701–716.
[29] F. Meyer and P. Maragos, *Morphological scale-space representation with levelings*, in Scale-Space Theories in Computer Vision, Lecture Notes in Comput. Sci. 1682, M. Nielsen, P. Johansen, O. F. Olsen, and J. Weickert, eds., Springer-Verlag, Berlin, 1999, pp. 187–198.
[30] Y. Meyer, *Ondelettes et Opérateurs*, Herman, Paris, 1990.
[31] A. W. Naylor and G. R. Sell, *Linear Operator Theory in Engineering and Science*, Springer-Verlag, New York, 1982.
[32] P. Perona and J. Malik, *Scale-space and edge detection using anisotropic diffusion*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 12 (1990), pp. 629–639.
[33] C. Ronse, *Openings: Main Properties, and How to Construct Them*, unpublished manuscript, 1991.
[34] P. Salembier, A. Oliveras, and L. Garrido, *Antiextensive connected operators for image and sequence processing*, IEEE Trans. Image Process., 7 (1998), pp. 555–570.
[35] J. Serra, *Connectivity on complete lattices*, J. Math. Imaging Vision, 9 (1998), pp. 231–251.
[36] J. Serra, *Connections for sets and functions*, Fund. Inform., 41 (2000), pp. 147–186.
[37] C. S. Tzafestas and P. Maragos, *Shape connectivity: Multiscale analysis and application to generalized granulometries*, J. Math. Imaging Vision, 17 (2002), pp. 109–129.
[38] R. van den Boomgaard and A. Smeulders, *The morphological structure of images: The differential equations of morphological scale-space*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 16 (1994), pp. 1101–1113.
[39] M. Vetterli and J. Kovačević, *Wavelets and Subband Coding*, Prentice–Hall, Englewood Cliffs, NJ, 1995.
[40] L. Vincent, *Morphological grayscale reconstruction in image analysis: Applications and efficient algorithms*, IEEE Trans. Image Process., 2 (1993), pp. 176–201.
[41] J. Weickert, *Anisotropic Diffusion in Image Processing*, Teubner-Verlag, Stuttgart, 1998.
[42] A. P. Witkin, *Scale-space filtering*, in Proceedings of the 8th International Joint Conference on Artificial Intelligence, Palo Alto, CA, 1983, Morgan Kaufmann, San Francisco, 1983, pp. 1019–1022.

SIAM J. MATH. ANAL.
Vol. 36, No. 1, pp. 121–171

REGULARITY FOR THE VLASOV–POISSON SYSTEM IN A CONVEX DOMAIN*

HYUNG JU HWANG†

Abstract. We consider the initial-boundary value problem in a convex domain for the Vlasov–Poisson system. Boundary effects play an important role in such physical problems that are modeled by the Vlasov–Poisson system. We establish the global existence of classical solutions with regular initial boundary data under the absorbing boundary condition. We also prove that regular symmetric initial data lead to unique classical solutions for all time in the specular reflection case.

Key words. regularity, global existence, initial-boundary value problem, convexity, Vlasov–Poisson system

AMS subject classifications. 35A05, 35B65, 78A35

DOI. 10.1137/S0036141003422278

1. Introduction. The behavior of a confined hot plasma is governed by the Vlasov–Maxwell system with boundary conditions. A simpler model is the Vlasov–Poisson system where the speed of light is treated as infinity and the magnetic field is neglected. For the absorbing case, we consider $\Pi = [0,T] \times \Omega \times \mathbb{R}^3$, where Ω is a smooth bounded convex domain in $\mathbb{R}^3$ and $T > 0$ is arbitrary, while we restrict to the unit ball $\Omega = B$ in the case of the specular reflection. We denote by n_x the outward normal vector at a boundary point $x \in \partial\Omega$. The Vlasov–Poisson system describes a collisionless plasma electrostatic:

$$\begin{aligned} &f_t + v \cdot \partial_x f + \nabla\varphi \cdot \partial_v f = 0, \\ &\Delta\varphi = \rho = 4\pi \int_{\mathbb{R}^3} f(t,x,v)\, dv, \\ &f|_{t=0} = f_0, \end{aligned} \tag{1.1}$$

where $f(t,x,v)$ represents the distribution of an electron gas, and φ is the electrostatic potential. The particles have the same sign of charge inside the region Ω, and $\nabla\varphi(t,x)$ is the self-consistent electric field. Boundary effects play an important role in such physical problems as tokamaks, diodes, and electron guns. Particles can be either *absorbed* at the boundary or *reflected* specularly at the boundary. For the absorbing boundary case, at $\{v \cdot n_x < 0\}$, with n_x the outward normal at $x \in \partial\Omega$, we have

$$f(t,x,v) = g(t,x,v), \tag{1.2}$$

where g is a given function. In the case of the specular reflection, at $\{v \cdot n_x < 0\}$, we have

$$f(t,x,v) = f(t,x,v_*), \tag{1.3}$$

where $v_* = v - 2(v \cdot n_x) n_x$.

*Received by the editors January 31, 2003; accepted for publication (in revised form) October 3, 2003; published electronically June 22, 2004.
http://www.siam.org/journals/sima/36-1/42227.html
†Department of Mathematics, Duke University, Box 93020, Durham, NC 27708 (hjhwang@math.duke.edu).

In this article, we construct classical solutions for the nonlinear Vlasov–Poisson system in a three-dimensional smooth bounded convex domain. We demonstrate our results of regularity as follows.

THEOREM 1.1 (absorbing case). *Assume the absorbing condition* (1.2) *for the Vlasov and the Dirichlet boundary condition for the Poisson. Let* $f_0 \geq 0$, $g \geq 0$ *be smooth with compact supports and* f_0 *be not identically zero. Let* f_0 *and* g *satisfy some compatibility conditions. Moreover, assume some vanishing condition for* g *at* $\{x \cdot n_x = 0\}$. *Then there exists a unique smooth solution* f *and* φ *of* (1.1) *with* (1.2), *where* f *has compact support for* v.

THEOREM 1.2 (specular reflection case). *Assume the specular boundary condition* (1.3) *for the Vlasov and the Dirichlet boundary condition for the Poisson. Assume there is an* $\omega_0 > 0$ *such that* $f_0(x,v)$ *is constant for* $\left(1-|x|^2\right)^2 + (2v \cdot x)^2 \leq \omega_0$.

(a) *Assume* $f_0 \in C^1$. *Let* f_0 *have compact support and satisfy the compatibility conditions. Let* f_0 *be spherically symmetric. Then there exists a unique spherically symmetric solution* (f, φ) *of* (1.1) *with* (1.3) *such that* $f \in W^{1,\infty}$ *with compact support.*

(b) *Assume* $f_0 \in C^{1,\eta}$ *for some* $\eta > 0$. *Let* f_0 *have compact support and satisfy the compatibility conditions. Let* f_0 *be spherically symmetric. Then there exists a unique spherically symmetric solution* (f, φ) *of* (1.1) *with* (1.3) *such that* $f \in C^{1,\mu}$, $\varphi \in C^{3,\mu}$ *for some* $0 < \mu < \eta$, *with compact support.*

Much effort and fruitful achievement have been made for the Cauchy problem for the Vlasov–Poisson system during the last few decades. Many mathematicians have made their contributions to the Vlasov–Poisson system in the whole three space dimensions without boundary conditions. In particular, in [22], [18], [23], and [16], global classical solutions for the Vlasov–Poisson system have been constructed by different methods, provided the initial data is regular.

However, the boundary-value problem is much more complicated since the boundary is always characteristic. In a half space with a flat boundary [7], [8], it is known that singularities of distribution function are expected, forming from the boundaries, *unless* the electric field has the correct sign. The global classical solutions for the full Vlasov–Poisson system have been constructed for a half space with a flat boundary in [7], [8] for one dimension and three dimensions, respectively.

This article extends the work of Guo to a three-dimensional smooth convex domain. We note that convexity plays an important role in obtaining regularity of the solutions of the Vlasov equation with boundary conditions. We refer the reader to [8] for a simple counterexample. We begin by generalizing the linear $C^{1,\alpha}$ and $W^{1,p}$ estimates in [7] and [8] to a general smooth bounded convex domain where a new geometric part comes in. As in the half space case, the main difficulty lies in the estimation of the particles moving slowly in the normal direction near the boundary. This can be overcome via the geometric velocity lemma with an extra factor coming from the geometry of the convexity. We still require the outwardness of the electric field E at the boundary and the flatness of the initial density f_0 to ensure the regularity in the linear problem. In the absorbing case, we adopt the high-moment technique in [18] in order to establish the existence of global classical solutions for the absorbing boundary condition. The key step to get control of large velocities is to represent the macrocharge density in the presence of the boundary condition. We are able to attain a representation for the charge density in spite of the complex particle paths by an *exact cancellation* at the boundary. This cancellation demands a new computation. Unfortunately, neither this high-moment method nor the technique first invented by

Pfaffelmoser worked for the specular reflection case. The central difficulty comes from the fact that we cannot avoid so many repeated bounces of the particles near the boundary with very small tangential angles if the particles are allowed to reflect at the boundary. This accelerates the hindrance to the control on the behavior of the particles near the boundary in addition to the difficulty from large velocities. In fact, the number of bounces of a particle near the boundary with constant velocity v and its tangential angle θ is proportional to $|v|/\theta$. So even if the particle moves slowly near the boundary, we easily lose the control on the number of bounces, because the particle moves almost tangentially with the very small θ. However, the invariance of the angular momentum in the spherically symmetric case enables us to treat the particles with small tangential angles since the angular momentum of the particles near the boundary with small tangential velocity amounts approximately to the full velocity. This leads to a global bound on the increase in velocity, employing the idea in [14].

This article is arranged as follows. From section 2 to section 4, we study the linear problem. In section 2, we establish the velocity lemma for a convex domain, followed by the study of the bouncing trajectories. The absorbing case is discussed in section 3. We deal with the linear estimates for the specular reflection in section 4. In section 5, we treat the fully nonlinear Vlasov–Poisson system with the absorbing boundary condition and get its regularity. Finally, in section 6, the nonlinear Vlasov–Poisson system, endowed with the specular boundary condition, obtains the regularity for the spherically symmetric case.

2. Bouncing trajectories. Let the boundary γ of Π consist of

$$\begin{aligned} \gamma^+ &= \{(t,x,v)\,|0 \le t \le T, x \in \partial\Omega, v \cdot n_x < 0\}, \\ \gamma^- &= \{(t,x,v)\,|0 \le t \le T, x \in \partial\Omega, v \cdot n_x > 0\}, \\ \gamma^0 &= \{(t,x,v)\,|0 \le t \le T, x \in \partial\Omega, v \cdot n_x = 0\}. \end{aligned} \tag{2.1}$$

Let $\Pi_s = \{t = s\} \cap \Pi$, $\gamma_s = \{t = s\} \cap \gamma$, $\gamma_s^+ = \{t = s\} \cap \gamma^+$, and $\gamma_s^- = \{t = s\} \cap \gamma^-$ for $0 \le s \le t$.

Let the unique trajectory of

$$\frac{d}{d\tau}X = V, \quad \frac{d}{d\tau}V = E \tag{2.2}$$

such that $X(t;t,x,v) = x$, $V(t;t,x,v) = v$ be the following:

$$\Gamma(\tau;t,x,v) = (\tau; X(\tau;t,x,v), V(\tau;t,x,v)), \tag{2.3}$$

where $E(t,x) = \nabla\varphi(t,x)$ is the given electric field.

We consider in this section the initial-boundary problem for the linear Vlasov equation

$$\begin{aligned} f_t + v \cdot \partial_x f + E \cdot \partial_v f = 0, \\ f|_{t=0} = f_0, \quad f|_{\gamma^+} = g, \end{aligned} \tag{2.4}$$

where the given electric field $E(t,x)$ satisfies $E(t,x) \cdot n_x \ge \delta > 0$ at the boundary, for a fixed $\delta > 0$.

In the following, we establish a generalized velocity lemma [8] for our convex domain.

Let $x_0 \in \partial\Omega$; there exist a neighborhood V of x_0 and a smooth convex function $\phi(x_1, x_2)$ such that, after proper translation and rotation, $x_0 = (0,0,0)$, $\partial\Omega \cap V = \{x = (x_1, x_2, x_3) \,|x_1 = \phi(x_2, x_3)\}$, and $\Omega \cap V = \{x = (x_1, x_2, x_3) \,|x_1 > \phi(x_2, x_3)\}$. Straightening out the portion of the boundary by the diffeomorphism $\Phi(x_1, x_2, x_3) = (x_1 - \phi(x_2, x_3), x_2, x_3)$ with the inverse $\Psi(x_1, x_2, x_3) = (x_1 + \phi(x_2, x_3), x_2, x_3)$, we may assume that near the point $x_0 = (0,0,0)$, $\partial\Omega = \{x_1 = 0\}$, $\Omega = \{x_1 > 0\}$. Now we consider the Vlasov–Poisson system in the new coordinates

$$\begin{aligned}
\tilde{t} := t, \quad \tilde{x} := \Phi(x), \quad \tilde{v} =: \partial\Phi(x)\, v,\\
\tilde{E}(\tilde{t}, \tilde{x}) := \partial\Phi(x)\, E(t,x) = \partial\Phi(\Psi(\tilde{x}))\, E(\tilde{t}, \Psi(\tilde{x})),\\
\tilde{f}(\tilde{t}, \tilde{x}, \tilde{v}) := f(t,x,v) = f(\tilde{t}, \Psi(\tilde{x}), \partial\Psi(\tilde{x})\,\tilde{v}).
\end{aligned}$$

Then we have

$$\begin{aligned}
0 &= f_t + v \cdot \partial_x f + E \cdot \partial_v f\\
&= \tilde{f}_{\tilde{t}} + \tilde{v} \cdot \partial_{\tilde{x}} \tilde{f} + \left[\tilde{E}(\tilde{t}, \tilde{x}) + v\partial^2\Phi(x)\, v\right] \cdot \partial_{\tilde{v}} \tilde{f}.
\end{aligned}$$

Notice that the outward normal $\tilde{n}_{\tilde{x}} = n_x \partial\Psi(\tilde{x})$, and so the boundary set γ^+ corresponds to $\tilde{\gamma}^+$, γ^0 corresponds to $\tilde{\gamma}^0$, and γ^- corresponds to $\tilde{\gamma}^-$, respectively, under this change of variables since

$$\tilde{n}_{\tilde{x}} \cdot \tilde{v} = [n_x \partial\Psi(\tilde{x})] \cdot [\partial\Phi(x)\, v] = n_x \cdot v.$$

Furthermore, the assumption on the electric field is invariant under the change of variables for the same reason. We shall now look at the sign on $v\partial^2\Phi^1(x)\, v$ as follows:

$$\begin{aligned}
v \cdot \partial^2\Phi^1 \cdot v &= \begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0\\ 0 & -\partial_{22}\phi & -\partial_{23}\phi\\ 0 & -\partial_{32}\phi & -\partial_{33}\phi \end{bmatrix} \begin{bmatrix} v_1\\ v_2\\ v_3 \end{bmatrix}\\
&= \begin{bmatrix} v_2 & v_3 \end{bmatrix} \begin{bmatrix} -\partial_{22}\phi & -\partial_{23}\phi\\ -\partial_{32}\phi & -\partial_{33}\phi \end{bmatrix} \begin{bmatrix} v_2\\ v_3 \end{bmatrix} \le 0
\end{aligned}$$

since ϕ is a convex function.

We can thus reduce locally our case to the half space case with a different equation. For our convenience, we will use the notation without a tilde, indicating things with a tilde, throughout this section. We now consider locally the following system in the upper half space in the time interval $[\tilde{t}, \tilde{t} + \varepsilon]$:

$$f_t + v \cdot \partial_x f + [E(t,x) + J(x,v)] \cdot \partial_v f = 0,$$

where $E_1(t,x) \le -\delta < 0$, $J_1(x,v) \le 0$ for all v, at $\gamma = \{(t,x,v) \,|x_1 = 0\}$.

LEMMA 2.1 (velocity lemma). *Suppose $E_1(t,0,\bar{x}) \le -\delta$ and $J_1(x,v) \le 0$. Let $E \in C^1$ and $J \in C^1$. Let $(\tau, X(\tau), V(\tau)) \in \bar{\Pi}$ for small time interval $[\tilde{t}, \tilde{t} + \varepsilon]$. Then for $\tilde{t} \le s \le t \le \tilde{t} + \varepsilon$,*

$$e^{-C(t-s)}\alpha(s) \le \alpha(t) \le e^{C(t-s)}\alpha(s), \tag{2.5}$$

where

$$\alpha(t) = X_1^2(t) + V_1^2(t) - 2\left[E_1\left(t; 0, \bar{X}(t)\right) + J_1(X,V)\right] X_1,$$

$(\tau, X(\tau), V(\tau))$ *is a trajectory* (2.2)*, and* C *depends on* Ω, $\|E\|_{C^1}$, δ, *and* $\sup_{s\le\tau\le t}|V(\tau)|$.

Proof. We follow closely the proof of Lemma 1.1 in [7]. Expanding $E_1(t,x)$ around $x_1 = 0$, we get from (2.2)

$$
\begin{aligned}
(2.6)\qquad & X_1^\bullet = V_1,\\
& V_1^\bullet = E_1\left(\tau; 0, \bar{X}(\tau)\right) + \partial_{x_1} E\left(\tau; \zeta, \bar{X}(\tau)\right) X_1(\tau) + J_1\left(X(\tau), V(\tau)\right),
\end{aligned}
$$

where $\bullet$ means the τ derivative, and $0 \le \zeta \le X_1(\tau)$. We multiply the first of (2.6) with $X_1(\tau)$ and the second of (2.6) with $V_1(\tau)$. Then there is a C large, depending on $\|\nabla_x E\|_\infty$, such that, along the trajectory,

$$
\left[e^{C\tau}\left(X_1^2(\tau) + V_1^2(\tau)\right)\right]^\bullet \ge 2\left[E_1\left(\tau; 0, \bar{X}(\tau)\right) + J_1\left(X(\tau), V(\tau)\right)\right] e^{C\tau} V_1(\tau)
$$

for $s \le \tau \le t$. Notice that

$$
\begin{aligned}
(2.7)\qquad & \left[E_1\left(\tau; 0, \bar{X}(\tau)\right) + J_1\left(X(\tau), V(\tau)\right)\right] e^{C\tau} V_1(\tau)\\
&= \left\{\left[E_1\left(\tau; 0, \bar{X}(\tau)\right) + J_1\left(X(\tau), V(\tau)\right)\right] e^{C\tau} X_1(\tau)\right\}^\bullet\\
&\quad - C\left[E_1\left(\tau; 0, \bar{X}(\tau)\right) + J_1\left(X(\tau), V(\tau)\right)\right] e^{C\tau} X_1(\tau)\\
&\quad - \frac{d}{d\tau}\left[E_1\left(\tau; 0, \bar{X}(\tau)\right) + J_1\left(X(\tau), V(\tau)\right)\right] e^{C\tau} X_1(\tau).
\end{aligned}
$$

Integrating (2.7) from s to t, for C large enough, we get the left-hand side (LHS) of (2.5) as

$$
\begin{aligned}
& e^{Ct}\left(X_1^2(\tau) + V_1^2(\tau)\right) - e^{Cs}\left(X_1^2(\tau) + V_1^2(\tau)\right)\\
&\ge 2\left[E_1\left(t; 0, \bar{X}(t)\right) + J_1\left(X(t), V(t)\right)\right] e^{Ct} X_1(t)\\
&\quad - 2\left[E_1\left(s; 0, \bar{X}(s)\right) + J_1\left(X(s), V(s)\right)\right] e^{Cs} X_1(s)\\
&\quad - \int_s^t \frac{d}{d\tau}\left[E_1\left(\tau; 0, \bar{X}(\tau)\right) + J_1\left(X(\tau), V(\tau)\right)\right] e^{C\tau} X_1(\tau)\, d\tau\\
&\quad - \int_s^t C\left[E_1\left(\tau; 0, \bar{X}(\tau)\right) + J_1\left(X(\tau), V(\tau)\right)\right] e^{C\tau} X_1(\tau)\, d\tau\\
&\ge 2\left[E_1\left(t; 0, \bar{X}(t)\right) + J_1\left(X(t), V(t)\right)\right] e^{Ct} X_1(t)\\
&\quad - 2\left[E_1\left(s; 0, \bar{X}(s)\right) + J_1\left(X(s), V(s)\right)\right] e^{Cs} X_1(s).
\end{aligned}
$$

We have used the fact that

$$
\begin{aligned}
& \left|\frac{d}{d\tau}\left[E_1\left(\tau; 0, \bar{X}(\tau)\right) + J_1\left(X(\tau), V(\tau)\right)\right]\right|\\
&\le \left|E_{1t}\left(\tau; 0, \bar{X}(\tau)\right) + \nabla_{\bar{x}} E_1\left(\tau; 0, \bar{X}(\tau)\right)\cdot \bar{V}(\tau) + \nabla_x J\left(X(\tau), V(\tau)\right)\cdot V(\tau)\right.\\
&\qquad \left. + \nabla_v J_1\left(X(\tau), V(\tau)\right)\cdot\left[E\left(\tau; X(\tau)\right) + J\left(X(\tau), V(\tau)\right)\right]\right|\\
&\le C\left(1 + \|E\|_{C^1}\right),
\end{aligned}
$$

and $E_1\left(\tau; 0, \bar{X}(\tau)\right) + J_1\left(X(\tau), V(\tau)\right) \le -\delta < 0$. Similarly, we establish the right-hand side (RHS) of (2.5). □

Corollary 2.2. *Suppose that* $E(t,x)\cdot n_x \ge \delta > 0$ *for all* $x \in \partial\Omega$ *for some fixed constant* $\delta > 0$. *Let* (t,x,v) *with* $x \notin \partial\Omega$ *connect to* (t_0, x_0, v_0) *with* $x \in \partial\Omega$ *through*

a trajectory, where $t_0 < t$ and the trajectory stays in the domain Ω in $(t_0, t]$. Then we have $(t_0, x_0, v_0) \in \gamma^+$.

Proof. We may assume without loss of generality that x is near the boundary so that we can localize near the point x. Then, by Lemma 2.1, we deduce that $(t_0, x_0, v_0) \in \gamma^+$. □

The following corollary gives a better estimate for C in the velocity lemma when $(d/dt) E_1|_\gamma \equiv 0$. We refer the reader to [7], [8] for the proof in the half space case with the flat boundary. It is important for the nonlinear specular case.

COROLLARY 2.3. *If $E \in C^{0,1}_{t,x}$ and $E_1(t, 0, \bar{x}) \equiv E_0(0, \bar{x}) < 0$ for all $0 \le t \le T$, then for $\tilde{t} \le s \le t \le \tilde{t} + \varepsilon$,*

$$e^{-C(t-s)}\beta(s) \le \beta(t) \le e^{C(t-s)}\beta(s),$$

where

$$\beta(t) = V_1^2(t) - 2\left[E_0\left(0, \bar{X}(t)\right) + J_1(X(t), V(t))\right] X_1(t),$$

and C depends only on $\sup_{0\le t\le T} \|E\|_{C^{0,1/2}(\Omega)}(t)$, E_0, and $\sup_{s\le\tau\le t} |V(\tau)|$.

In many physical problems, particles may have the complex behavior of bouncing off the boundary repeatedly. In order to describe such phenomena and to study especially the specular reflection case, we will investigate trajectories which bounce many times at the boundary. We call such a particle path which ends at a given point a "back-time cycle" as in [8]. Notice that the density is constant along these kinds of generalized trajectories.

DEFINITION 2.4. *$v_* = v - 2(v \cdot n_x) n_x$ is said to be the reflected velocity of v.*

DEFINITION 2.5 (back-time cycles). *Given a C^1 field $E(t, x)$, by an l-cycle, we mean the trajectories in $\bar{\Pi}$ which connect $(t, x, v) = \left(t^l, x^l, v^l\right)$ with $\left(t^{l-1}, x^{l-1}, v^{l-1}\right)$, $\left(t^{l-1}, x^{l-1}, v_*^{l-1}\right)$ with $\left(t^{l-2}, x^{l-2}, v^{l-2}\right)$, $\ldots$, $\left(t^i, x^i, v_*^i\right)$ with $\left(t^{i-1}, x^{i-1}, v^{i-1}\right)$, $\ldots$, $\left(t^1, x^1, v_*^1\right)$ with $(0, x_0, v_0)$, where $t^i > t^{i-1}$, $x^i \in \partial\Omega$ for $1 \le i \le l-1$, $v^i \cdot n_x \ge 0$, $1 \le i \le l$.*

We rewrite the velocity lemma, Lemma 2.1, involving our geometry. Let $\xi(x)$ be a smooth function which defines the boundary such that

$$\partial\Omega = \{\xi(x) = 0\}, \quad \Omega = \{\xi(x) > 0\}; \tag{2.8}$$

then $n_x = -\nabla\xi(x) / |\nabla\xi(x)|$ is the outward normal at each point x at the boundary. For instance, $\xi(x) = 1 - |x|^2$ for the unit ball.

LEMMA 2.6 (geometric velocity lemma). *Let $E(t, x) \cdot n_x \ge \delta > 0$ for all $x \in \partial\Omega$ with $E \in C^1$. If the trajectory stays away from the origin, i.e., $|X(\tau)| \ge \sigma$ for $s \le \tau \le t$, for any small fixed $\sigma > 0$, then*

$$e^{-C(t-s)}\alpha(s) \le \alpha(t) \le e^{C(t-s)}\alpha(s), \tag{2.9}$$

where

$$\begin{aligned}\alpha(t) = {} & \xi^2(X(t)) + [V(t) \cdot \nabla\xi(X(t))]^2 \\ & - 2\left[E\left(t, \bar{X}(t)\right) \cdot \nabla\xi\left(\bar{X}(t)\right) + V(t) \cdot \nabla^2\xi(X(t)) \cdot V(t)\right] \xi(X(t))\end{aligned}$$

and where C depends on $\|E\|_{C^1}$, $\sup_{0\le\tau\le t}[|X(\tau)| + |V(\tau)|]$, δ, and σ.

Proof. Let $\bar{X}$ be the point at the boundary which lies on the half-line from the point X in the direction $-\nabla\xi(X)$. Then we expand $E \cdot \nabla\xi(X)$ around $E \cdot \nabla\xi\left(\bar{X}\right)$ to

get

$$\begin{aligned}(E\cdot\nabla\xi)(X) &= (E\cdot\nabla\xi)\left(\bar{X}\right)+\nabla_x(E\cdot\nabla\xi)\cdot\left(X-\bar{X}\right)\\ &= E\left(t,\bar{X}\right)\cdot\nabla\xi\left(\bar{X}\right)+\left[\nabla_x(E\cdot\nabla\xi)\cdot\frac{\left(X-\bar{X}\right)}{\xi(X)}\right]\xi(X).\end{aligned}$$

From the fact that $\xi(X)=\xi\left(\bar{X}\right)+\nabla\xi(\theta)\cdot\left(X-\bar{X}\right)=\nabla\xi(\theta)\cdot\left(X-\bar{X}\right)$ for some point θ on the line segment connecting X and $\bar{X}$, which implies that $\nabla\xi(\theta)\cdot\frac{(X-\bar{X})}{\xi(X)}=1$, we can easily see that $\left|\frac{(X-\bar{X})}{\xi(X)}\right|\le C$. (Near the boundary where $\xi(X)\approx 0$, we have $|\nabla\xi(\theta)|\ge c>0$, and $\nabla\xi(\theta)$ is almost parallel to $X-\bar{X}$.) Along the trajectory, there is a C so large that

$$\begin{aligned}&\left\{e^{C\tau}\left[\xi^2(X)+(V\cdot\nabla\xi)^2\right]\right\}^{\bullet}\\ &= e^{C\tau}\Big[C\xi^2(X)+C(V\cdot\nabla\xi)^2+2\xi(X)(V\cdot\nabla\xi)\\ &\qquad +2(V\cdot\nabla\xi)\left(E\cdot\nabla\xi(X)+V\cdot\nabla^2\xi\cdot V\right)\Big]\\ &= e^{C\tau}\left[C\xi^2(X)+C(V\cdot\nabla\xi)^2+2\left(1+\nabla_x(E\cdot\nabla\xi)\cdot\frac{\left(X-\bar{X}\right)}{\xi(X)}\right)\xi(X)(V\cdot\nabla\xi)\right.\\ &\qquad \left.+2(V\cdot\nabla\xi)\left(E\left(t,\bar{X}\right)\cdot\nabla\xi\left(\bar{X}\right)+V\cdot\nabla^2\xi\cdot V\right)\right]\\ &\ge e^{C\tau}\left[2(V\cdot\nabla\xi)\left(E\left(t,\bar{X}\right)\cdot\nabla\xi\left(\bar{X}\right)+V\cdot\nabla^2\xi\cdot V\right)\right].\end{aligned}$$

Assuming that $E(t,x)\cdot n_x\ge\delta>0$, we notice that for a large C,

$$\begin{aligned}&e^{C\tau}\left[2(V\cdot\nabla\xi)\left(E\left(t,\bar{X}\right)\cdot\nabla\xi\left(\bar{X}\right)+V\cdot\nabla^2\xi\cdot V\right)\right]\\ &= \left\{2e^{C\tau}\xi(X)\left[E\left(t,\bar{X}\right)\cdot\nabla\xi\left(\bar{X}\right)+V\cdot\nabla^2\xi\cdot V\right]\right\}^{\bullet}\\ &\quad -2e^{C\tau}C\left[E\left(t,\bar{X}\right)\cdot\nabla\xi\left(\bar{X}\right)+V\cdot\nabla^2\xi\cdot V\right]\xi(X)\\ &\quad -2e^{C\tau}\left[E_t\left(t,\bar{X}\right)\cdot\nabla\xi\left(\bar{X}\right)+\bar{X}^{\bullet}\cdot\nabla_{\bar{x}}E\left(t,\bar{X}\right)\cdot\nabla\xi\left(\bar{X}\right)\right.\\ &\qquad\qquad +E\left(t,\bar{X}\right)\cdot\nabla^2\xi\left(\bar{X}\right)\cdot\bar{X}^{\bullet}+2E\cdot\nabla^2\xi(X)\cdot V\\ &\qquad\qquad \left.+V\cdot\left(\nabla^3\xi(X)\cdot V\right)\cdot V\right]\xi(X)\\ &\ge \left\{2e^{C\tau}\xi(X)\left[E\left(t,\bar{X}\right)\cdot\nabla\xi\left(\bar{X}\right)+V\cdot\nabla^2\xi\cdot V\right]\right\}^{\bullet},\end{aligned}$$

where $\xi(X)\ge 0$, and $\left|\bar{X}^{\bullet}\right|\le C$ (since $|X(\tau)|\ge\sigma>0$), $|X(\tau)|\le C$, $|V(\tau)|\le C$, and $|E|_{C^1}\le C$. This proves the LHS of (2.9). Similarly, we get the RHS of (2.9) to complete the proof of the lemma. □

The following lemma shows that if a particle initially has a nonzero normal velocity, then its normal velocity of a particle remains bounded away from 0, and the bound is independent of the number of the bounces.

LEMMA 2.7. *Let $E(t,x)\cdot n_x\ge\delta>0$ for all $x\in\partial\Omega$. Consider the back-time cycle of (t,x,v). Then there exist C_1 and C_2 such that*

$$\begin{aligned}C_1\left[\xi(x)+(v\cdot\nabla\xi(x))^2\right] &\le \left(v^i\cdot\nabla\xi\left(x^i\right)\right)^2\le C_2\left[\xi(x_0)+(v_0\cdot\nabla\xi(x_0))^2\right],\\ C_1\left[\xi(x_0)+(v_0\cdot\nabla\xi(x_0))^2\right] &\le \left(v^i\cdot\nabla\xi\left(x^i\right)\right)^2\le C_2\left[\xi(x)+(v\cdot\nabla\xi(x))^2\right],\end{aligned}$$

where $1 \le i \le l$, *and* C_1 *and* C_2 *are independent of* l *and dependent on* $\|E\|_{C^1}$, δ, *and the bound for* $|V(\tau)|$.

Proof. We may assume that Ω contains the origin without loss of generality. We first consider small balls B_σ with radius $\sigma > 0$ small and let (s, y, w) connect with $(\tilde{t}, \tilde{x}, \tilde{v})$ through a trajectory in the ball B_σ. Since $|X(\tau)| \le \sigma$, we have $\xi(X(\tau)) \ge C(\sigma)$. Then there is a constant D such that

$$\begin{aligned}\left\{e^{D\tau}\left[\xi(X) + (V \cdot \nabla\xi(X))^2\right]\right\}^\bullet &= e^{D\tau}\Big[D\xi(X) + D(V \cdot \nabla\xi(X))^2 + V \cdot \nabla\xi(X) \\ &\qquad + 2(V \cdot \nabla\xi(X))\left(E \cdot \nabla\xi(X) + V \cdot \nabla^2\xi \cdot V\right)\Big] \\ &\ge 0,\end{aligned}$$

where D depends on $\|E\|_\infty$, Ω, σ, and the bound for $|V(\tau)|$. Hence we get

$$e^{Ds}\left[\xi(y) + (w \cdot \nabla\xi(y))^2\right] \le e^{D\tilde{t}}\left[\xi(\tilde{x}) + (\tilde{v} \cdot \nabla\xi(\tilde{x}))^2\right]. \tag{2.10}$$

Next, let (s, y, w) connect with $(\tilde{t}, \tilde{x}, \tilde{v})$ through a trajectory which goes through the σ-ball, where both y and $\tilde{x}$ are at the boundary. Let (s', y', w') and $(\tilde{t}', \tilde{x}', \tilde{v}')$ be the two points with $|y'| = |\tilde{x}'| = \sigma$ on the trajectory connecting (s, y, w), $(\tilde{t}, \tilde{x}, \tilde{v})$. Then by the geometric velocity lemma and by (2.10), there exist a C and D such that

$$\begin{aligned}e^{Ds}(w \cdot \nabla\xi(y))^2 &\le e^{Ds'}\Big[\xi^2(y') + \left(w' \cdot \nabla\xi(y')\right)^2 \\ &\qquad - 2\left\{(E \cdot \nabla\xi)(s', \bar{y}') + w' \cdot \nabla^2\xi(y') \cdot w'\right\}\xi(y')\Big] \\ &\le Ce^{Ds'}\left[\xi(y') + \left(w' \cdot \nabla\xi(y')\right)^2\right] \\ &\le Ce^{D\tilde{t}'}\left[\xi(\tilde{x}') + \left(\tilde{v}' \cdot \nabla\xi(\tilde{x}')\right)^2\right] \\ &\le Ce^{D\tilde{t}'}\Big[\xi^2(\tilde{x}') + \left(\tilde{v}' \cdot \nabla\xi(\tilde{x}')\right)^2 \\ &\qquad - 2\left\{(E \cdot \nabla\xi)(\tilde{t}', \tilde{x}') + \tilde{v}' \cdot \nabla^2\xi(\tilde{x}') \cdot \tilde{v}'\right\}\xi(\tilde{x}')\Big] \\ &\le Ce^{D\tilde{t}}(\tilde{v} \cdot \nabla\xi(\tilde{x}))^2,\end{aligned}$$

where C depends on $\|E\|_\infty$, Ω, δ, and the bound for $|V(\tau)|$. Now we observe that the number $\#$ of such happenings of hitting the σ-ball through the whole cycle is uniformly bounded. Along the trajectory, we have

$$C\Delta t \ge \left|\int_{t'}^{t''} V(\tau)\, d\tau\right| = |\Delta x| \ge c_\sigma,$$

which implies that $\Delta t \ge C_\sigma > 0$. Then $C_\sigma \times \# \le \sum \Delta t \le T$ indicates that $\#$ is uniformly bounded. Now pick i and consider $\left|v^i \cdot \nabla\xi(x^i)\right|^2$. For the upper bound,

we have

$$\begin{aligned}
&e^{Dt^i}\left(v^i\cdot\nabla\xi\left(x^i\right)\right)^2\\
&\le C^{\#}e^{Dt}\left[\xi^2(x)+(v\cdot\nabla\xi(x))^2-2\left(E\cdot\nabla\xi(\bar{x})+v\cdot\nabla^2\xi(x)\cdot v\right)\xi(x)\right]\\
&\le C\times C^{\#}e^{Dt}\left[\xi(x)+(v\cdot\nabla\xi(x))^2\right].
\end{aligned}$$

On the other hand, the lower bound is achieved as

$$\begin{aligned}
&c^{\#}\left[2\delta\xi(x_0)+(v_0\cdot\nabla\xi(x_0))^2\right]\\
&\le c^{\#}\left[\xi^2(x_0)+(v_0\cdot\nabla\xi(x_0))^2-2\left(E\cdot\nabla\xi(\bar{x}_0)+v_0\cdot\nabla^2\xi(x_0)\cdot v_0\right)\xi(x_0)\right]\\
&\le e^{Dt^i}\left(v^i\cdot\nabla\xi\left(x^i\right)\right)^2.
\end{aligned}$$

Therefore, we get

$$C_1\left[\xi(x_0)+(v_0\cdot\nabla\xi(x_0))^2\right]\le\left(v^i\cdot\nabla\xi\left(x^i\right)\right)^2\le C_2\left[\xi(x)+(v\cdot\nabla\xi(x))^2\right],$$

where C_1 and C_2 are independent of l and dependent on $\|E\|_{C^1}$, δ, Ω, and the bound for $|V(\tau)|$ on the cycle. Similarly, we can get the second part of the lemma. □

We prove the following corollary by the same method as in [7], [8].

COROLLARY 2.8. *Suppose that $E\in C^{0,1}_{t,x}$ and $[E(t,x)\cdot n_x]|_\gamma\equiv E_0(x)>0$. Consider the back-time cycle of (t,x,v). Then there are C_1 and $C_2>0$ such that*

$$\begin{aligned}
C_1\left[\xi(x)+(v\cdot\nabla\xi(x))^2\right]&\le\left(v^i\cdot\nabla\xi\left(x^i\right)\right)^2\le C_2\left[\xi(x_0)+(v_0\cdot\nabla\xi(x_0))^2\right],\\
C_1\left[\xi(x_0)+(v_0\cdot\nabla\xi(x_0))^2\right]&\le\left(v^i\cdot\nabla\xi\left(x^i\right)\right)^2\le C_2\left[\xi(x)+(v\cdot\nabla\xi(x))^2\right],
\end{aligned}$$

where C_1 and C_2 are independent of the number of the bounces, depend on $\sup_{0\le\tau\le T}\|E\|_{C^{0,1/2}(\Omega)}(\tau)$, $\|E_0\|_{C^1}$, and the bound for $|V(\tau)|$ on the cycle.

We now see that $t_0(t,x,v)$, $x_0(t,x,v)$, $v_0(t,x,v)$ are C^1 functions of (t,x,v) locally when (t_0,x_0,v_0) connects with (t,x,v) through a trajectory:

$$x_0=x+\int_t^{t_0}\left[v+\int_t^s E(\tau)\,d\tau\right]ds.$$

Let ξ be the smooth function which defines the boundary in (2.8). Then we have

$$\begin{aligned}
0&=\xi(x_0)\\
&=\xi\left(x+\int_t^{t_0}\left[v+\int_t^s E(\tau)\,d\tau\right]ds\right):=\bar{\xi}(t_0;t,x,v)
\end{aligned}$$

with C^1 coefficients. By differentiating $\bar{\xi}$ with respect to t_0, we get, by Corollary 2.2,

$$\begin{aligned}
\frac{\partial\bar{\xi}}{\partial t_0}(t_0;t,x,v)&=\nabla\xi(x_0)\cdot\left[v+\int_t^{t_0}E(\tau)\,d\tau\right]\\
&=\nabla\xi(x_0)\cdot v_0=n_{x_0}\cdot v_0<0.
\end{aligned}$$

We thus have $t_0=t_0(t,x,v)\in C^1$ by the implicit function theorem and $v_0=v_0(t,x,v)$, $x_0=x_0(t,x,v)\in C^1$.

Now we consider trajectories without any bounces from a point to a boundary point, which are close to each other, and from any point to an initial point.

LEMMA 2.9. *Let (t,x,v) connect with $(0,x_0,v_0)$ through a trajectory. Then*

$$v_0 = v + \int_t^0 E\left(\tau, X\left(\tau;t,x,v\right)\right) d\tau,$$
$$x_0 = x - vt - \int_0^t \int_t^s E\left(\tau, X\left(\tau;t,x,v\right)\right) d\tau ds.$$

Considering x_0 and v_0 as functions of (t,x,v), we have

$$v_{0t} = -E\left(t,x\right) + \int_t^0 \nabla_x E \cdot X_t d\tau,$$
$$x_{0t} = -v + E\left(t,x\right) t - \int_0^t \int_t^s \nabla_x E \cdot X_t d\tau,$$
$$\nabla_x v_0 = \int_t^0 \nabla_x E \, \nabla_x X d\tau, \qquad \nabla_v v_0 = I + \int_t^0 \nabla_x E \, \nabla_v X d\tau,$$
$$\nabla_x x_0 = I - \int_0^t \int_t^s \nabla_x E \, \nabla_x X d\tau, \ \nabla_v x_0 = -tI - \int_0^t \int_t^s \nabla_x E \, \nabla_v X d\tau.$$

Here all the integrations are taken along the trajectory $\left(\tau, X\left(\tau;t,x,v\right), V\left(\tau;t,x,v\right)\right)$. $\nabla_x E \, \nabla_x X$ and $\nabla E_x \, \nabla_v X$ are matrix multiplications.

LEMMA 2.10. *Let (t,x,v) connect with (t_0,x_0,v_0) through a trajectory, where $x_0 \in \partial\Omega$. Then*

$$v_0 = v + \int_t^{t_0} E\left(\tau\right) d\tau, \quad x_0 = x + \int_t^{t_0} \left[v + \int_t^s E\left(\tau\right) d\tau \right] ds. \tag{2.11}$$

For v_0 with $n_{x_0} \cdot v_0 < 0$,

$$x_{0t} = t_{0t} v - v + t_{0t} \int_t^{t_0} E\left(\tau\right) d\tau - \int_t^{t_0} E\left(t,x\right) ds + \int_t^{t_0} \int_t^s \nabla_x E \cdot X_t d\tau,$$
$$\nabla_x x_0 = I + t_{0x} \otimes \left(v + \int_t^{t_0} E\left(\tau\right) d\tau \right) + \int_t^{t_0} \int_t^s \nabla_x E \, \nabla_x X d\tau ds,$$
$$\nabla_v x_0 = \left(t_0 - t\right) I + t_{0v} \otimes \left(v + \int_t^{t_0} E\left(\tau\right) d\tau \right) + \int_t^{t_0} \int_t^s \nabla_x E \, \nabla_v X d\tau ds,$$
$$v_{0t} = t_{0t} E\left(t_0,x_0\right) - E\left(t,x\right) + \int_t^{t_0} \nabla_x E \cdot X_t d\tau,$$
$$\nabla_x v_0 = E\left(t_0,x_0\right) \otimes t_{0x} + \int_t^{t_0} \nabla_x E \, \nabla_x X d\tau,$$
$$\nabla_v v_0 = I + E\left(t_0,x_0\right) \otimes t_{0v} + \int_t^{t_0} \nabla_x E \, \nabla_v X d\tau,$$
$$t_{0x} = \left(n_{x_0} \cdot v_0\right)^{-1} \left[n_{x_0} + \int_{t_0}^t \int_t^s n_{x_0} \cdot \left(\nabla_x E \, \nabla_x X \right) d\tau ds \right],$$
$$t_{0v} = \left(n_{x_0} \cdot v_0\right)^{-1} \left[\left(t - t_0\right) n_{x_0} + \int_{t_0}^t \int_t^s n_{x_0} \cdot \left(\nabla_x E \, \nabla_v X \right) d\tau ds \right],$$
$$t_{0t} = 1 + \left(n_{x_0} \cdot v_0\right)^{-1} n_{x_0} \cdot \left[\int_{t_0}^t E\left(\tau\right) d\tau + \left(t_0 - t\right) E\left(t,x\right) + \int_t^{t_0} \int_t^s \nabla_x E \cdot X_t d\tau ds \right].$$

Proof. Let ξ be the function defining the boundary as in (2.8); then we have

$$\xi\left(x_0\left(t,x,v\right)\right)=0. \tag{2.12}$$

We differentiate (2.12) with respect to x, v, and t to get

$$n_{x_0}\cdot\nabla_x x_0=0,\quad n_{x_0}\cdot\nabla_v x_0=0,\quad n_{x_0}\cdot x_{0t}=0. \tag{2.13}$$

We now differentiate the second equation of (2.11) to get

$$\nabla_x x_0=I+v_0\otimes t_{0x}+\int_t^{t_0}\int_t^s \nabla_x E\,\nabla_x X\,d\tau ds. \tag{2.14}$$

By multiplying (2.14) with n_{x_0} and by (2.13), we get

$$0=n_{0x}\cdot\nabla_x x_0=\left(n_{0x}\cdot v_0\right)t_{0x}+n_{0x}+\int_t^{t_0}\int_t^s n_{0x}\cdot\left(\nabla_x E\,\nabla_x X\right)d\tau ds.$$

We thus have if $n_{0x}\cdot v_0<0$,

$$t_{0x}=\left(n_{x_0}\cdot v_0\right)^{-1}\left[n_{x_0}+\int_{t_0}^t\int_t^s n_{x_0}\cdot\left(\nabla_x E\,\nabla_x X\right)d\tau ds\right]. \tag{2.15}$$

By differentiating the second equation of (2.11) with respect to v and t, we deduce the formulas for t_{0v} and t_{0t}. We differentiate the first equation of (2.11) and do the same thing to obtain the formulas for $\nabla_x v_0$, $\nabla_v v_0$, and v_{0t}. Thus our lemma follows. ∎

LEMMA 2.11. *Let (t,x,v) connect with (t_0,x_0,v_0) through a trajectory with t close to t_0, where $x_0\in\partial\Omega$. If $E\cdot n\geq\delta>0$ at the boundary for all time, then*

$$\left|t-t_0\right|\leq C\left|v_0\cdot n_{x_0}\right|,$$

where C depends on $\|E\|_{C^1}$ and δ.

Proof. We need only to consider the case when $|v_0\cdot n_{x_0}|$ is small. Notice that

$$x=x_0+v_0\left(t-t_0\right)+\int_{t_0}^t\int_{t_0}^s E\left(\tau\right)d\tau ds.$$

Setting

$$h\left(t\right)=\xi\left(x\right)=\xi\left(x_0+v_0\left(t-t_0\right)+\int_{t_0}^t\int_{t_0}^s E\left(\tau\right)d\tau ds\right),$$

we expand $h\left(t\right)$ around $t=t_0$ to get

$$\begin{aligned}\xi\left(x\right)&=\xi\left(x_0\right)+h^{'}\left(t_0\right)\left(t-t_0\right)+h^{''}\left(t_0\right)\left(t-t_0\right)^2+O\left(t-t_0\right)^3\\&=\left(v_0\cdot\nabla\xi\left(x_0\right)\right)\left(t-t_0\right)\\&\quad+\left[v_0\cdot\nabla^2\xi\left(x_0\right)\cdot v_0+E\left(t_0,x_0\right)\cdot\nabla\xi\left(x_0\right)\right]\left(t-t_0\right)^2+O\left(t-t_0\right)^3.\end{aligned}$$

Thus, we obtain

$$\begin{aligned}&t-t_0\\&=\frac{v_0\cdot\nabla\xi\left(x_0\right)\pm\sqrt{\left(v_0\cdot\nabla\xi\left(x_0\right)\right)^2+4\xi\left(x\right)\left[v_0\cdot\nabla^2\xi\left(x_0\right)\cdot v_0+E\left(t_0,x_0\right)\cdot\nabla\xi\left(x_0\right)\right]}}{-2\left[v_0\cdot\nabla^2\xi\left(x_0\right)\cdot v_0+E\left(t_0,x_0\right)\cdot\nabla\xi\left(x_0\right)\right]}\\&\quad+o\left(t-t_0\right).\end{aligned}$$

Since $-\left[v_0\cdot\nabla^2\xi\left(x_0\right)\cdot v_0+E\left(t_0,x_0\right)\cdot\nabla\xi\left(x_0\right)\right]\geq\delta\left|\nabla\xi\left(x_0\right)\right|>0$, we have

$$\left|t-t_0\right|\leq\frac{1}{2\delta\left|\nabla\xi\left(x_0\right)\right|}2\left|v_0\cdot\nabla\xi\left(x_0\right)\right|=\frac{1}{\delta}\left|v_0\cdot n_{x_0}\right|. \qquad ∎$$

3. Regularity for linear absorbing.

THEOREM 3.1. *Let Ω be a smooth bounded convex domain, $E(t,x) \in C^1\left([0,\infty)\times\bar{\Omega}\right)$, and $E(t,x)\cdot n_x \geq \delta > 0$ for all $x \in \partial\Omega$ for some fixed constant $\delta > 0$. Let an initial datum $f_0 \in C^1\left(\bar{\Pi}_0\right)$ and a boundary datum $g \in C^1\left(\bar{\gamma}^+\right)$ be compactly supported. Assume the following compatibility conditions hold for $x \in \partial\Omega$ and v with $n_x \cdot v < 0$ (2.1):*

$$f_0(x,v) = g(0,x,v), \tag{3.1}$$

$$g_t(0,x,v) + v\cdot\nabla_x f_0(x,v) + E(0,x)\cdot\nabla_v f_0(x,v) = 0. \tag{3.2}$$

(a) *Then there exists a solution $f \in C^1\left(\bar{\Pi}\backslash\gamma^0\right)$ to (2.4).*

(b) *Furthermore, assume the following vanishing conditions hold:*

$$|\nabla g(t,x,v)| \leq C|n_x\cdot v|^{1+\kappa}, \quad |\nabla f_0(x,v)| \leq C(|\xi(x)| + |n_x\cdot v|)^{\kappa}, \tag{3.3}$$

where ξ is the function defining the boundary $\partial\Omega$ in (2.8) and $\kappa > 0$. Then $f(t,x,v) \in C^1\left(\bar{\Pi}\right)$.

Proof. We define $f(t,x,v)$ as follows. For any $(t,x,v) \in [0,T]\times\bar{\Omega}\times\mathbb{R}^3\backslash\gamma^0$, let (t_0,x_0,v_0) be the first point on $\partial\Pi$ which connects with (t,x,v) through a back-time trajectory. By applying the velocity lemma, Lemma 2.1, it follows that $(t_0,x_0,v_0) \notin \gamma^0$; i.e., (t_0,x_0,v_0) is not in the singular set. If $t_0 = 0$, we define

$$f(t,x,v) = f_0(x_0,v_0).$$

The t-derivative of f is given by

$$\begin{aligned} f_t(t,x,v) &= \nabla_x f_0(x_0,v_0)\cdot x_{0t} + \nabla_v f_0(x_0,v_0)\cdot v_{0t} \\ &= \nabla_x f_0(x_0,v_0)\cdot\left[-v + E(t,x)\,t - \int_0^t\int_t^s \nabla_x E\cdot X_t d\tau ds\right] \\ &\quad + \nabla_v f_0(x_0,v_0)\cdot\left[-E(t,x) + \int_t^0 \nabla_x E\cdot X_t d\tau\right]. \end{aligned} \tag{3.4}$$

On the other hand, if $x_0 \in \partial\Omega$, we define

$$f(t,x,v) = g(t_0,x_0,v_0). \tag{3.5}$$

When $n_x\cdot v < 0$, by Lemma 2.10, the t-derivative of f is

$$f_t(t,x,v) = g_t(t_0,x_0,v_0)\,t_{0t} + \nabla_{\bar{x}} g(t_0,x_{0,}v_0)\cdot x_{0t} + \nabla_v g(t_0,x_{0,}v_0)\cdot v_{0t}. \tag{3.6}$$

It is clear to see that f is well defined when $t_0 = 0$ and $x_0 \in \partial\Omega$ by the assumption (3.1). We now show that the two different t-derivatives of f coincide in the case that $t_0 = 0$, $x_0 \in \partial\Omega$, and $n_{x_0}\cdot v_0 < 0$. Using the formulas

$$\begin{aligned} v &= v_0 + \int_0^t E(\tau)\,d\tau, \\ t_{0t} &= 1 + (n_{x_0}\cdot v_0)^{-1} n_{x_0}\cdot\left[\int_{t_0}^t E(\tau)\,d\tau + (t_0 - t)\,E(t,x) + \int_t^{t_0}\int_t^s \nabla_x E\cdot X_t d\tau ds\right] \end{aligned}$$

from Lemma 2.10 and the compatibility conditions (3.1), (3.2) into (3.6) yields

$$\begin{aligned} f_t(t,x,v) &= \left[-v_0\cdot\nabla_x f_0 - E(0,x_0)\cdot\nabla_v f_0\right] t_{0t} \\ &\quad + \nabla_x f_0\cdot\left[t_{0t}v_0 - v - E(t,x)\,t - \int_0^t\int_t^s \nabla_x E\cdot X_t d\tau ds\right] \\ &\quad + \nabla_v f_0\cdot\left[t_{0t}E(0,x_0) - E(t,x) + \int_t^0 \nabla_x E\cdot X_t d\tau\right] \\ &= \nabla_x f_0\cdot\left[-v - E(t,x)\,t - \int_0^t\int_t^s \nabla_x E\cdot X_t d\tau ds\right] \\ &\quad + \nabla_v f_0\cdot\left[-E(t,x) + \int_t^0 \nabla_x E\cdot X_t d\tau\right]. \end{aligned}$$

This is the same as (3.4). By similar computations, we see that f_x and f_v are continuous when $t_0 = 0$ and $x_0 \in \partial\Omega$. Thus part (a) follows.

For (b), we define $f(t,x,v) = 0$ for $(t,x,v) \in \gamma^0$. We show that $|\nabla f(t,x,v)| \to 0$ when (t,x,v) goes to a point in γ^0. If (t,x,v) connects with (t_0,x_0,v_0), then it follows from (3.5), (3.4), and from the velocity lemma, Lemma 2.1, and Lemma 2.10 that

$$\left|\nabla_{(t,x,v)} f(t,x,v)\right| \le C\frac{1}{|n_{x_0}\cdot v_0|}\left|\nabla g(t_0,x_0,v_0)\right| \le C\left|n_{x_0}\cdot v_0\right|^{\kappa}.$$

If (t,x,v) connects with $(0,x_0,v_0)$, then by (3.4)

$$|\nabla f(t,x,v)| \le C\left|\nabla f_0(x_0,v_0)\right| \le C(|\xi(x_0)| + |n_{x_0}\cdot v_0|)^{\kappa}.$$

By the velocity lemma, Lemma 2.1, as $\xi^2(x) + (n_x\cdot v)^2 \to 0$, $n_{x_0}\cdot v_0 \to 0$ and $|\xi(x_0)| + |n_{x_0}\cdot v_0| \to 0$. The theorem thus follows. □

We also deduce the following theorem.

THEOREM 3.2. *Let $E(t,x) \in C^1\left([0,\infty]\times\bar{\Omega}\right)$ with $E(t,x)\cdot n_x \ge \delta > 0$ on $\partial\Omega$. Let $F_0(x,v)$, $H(t,x,v)$, and $G(t,x,v)$ be $(n\times 1)$-vector-valued functions and $A(t,x,v)$ be an $(n\times n)$-matrix function such that $G(t,x,v) \in C^1(\bar{\gamma}^+)$, $F_0 \in C^1\left(\bar{\Pi}_0\right)$, $H(t,x,v)$ and $A(t,x,v) \in C^1\left(\bar{\Pi}\right)$, and H, G, and F_0 have compact support in v. Assume the compatibility conditions hold for $x \in \partial\Omega$, v with $v\cdot n_x < 0$:*

$$\begin{aligned} &F_0(x,v) = G(0,x,v), \\ &G_t(x,v) + v\cdot\nabla_x F_0 + E(0,x)\cdot\nabla_v F_0 = A(0,x,v)\,F_0(x,v) + H(0,x,v). \end{aligned} \tag{3.7}$$

(a) *Then there exists a unique $(n\times 1)$-vector-valued function $F(t,x,v) \in C^1\left(\bar{\Pi}\backslash\gamma_0\right)$ such that $F_t + v\cdot\nabla_x F + E\cdot\nabla_v F = AF + H$, $F|_{\gamma^+} = G$, $F|_{t=0} = F_0$ for $(t,x,v) \in \bar{\Pi}\backslash\gamma_0$.*

(b) *Furthermore, assume that the vanishing conditions hold:*

$$\begin{aligned} &|\nabla G(t,x,v)| \le C\left|v\cdot n_x\right|^{1+\kappa}, \quad |\nabla H(t,x,v)| \le C\left(|\xi(x)| + |v\cdot n_x|\right)^{1+\kappa}, \\ &|\nabla F_0(t,x,v)| \le C\left(|\xi(x)| + |v\cdot n_x|\right)^{\kappa}. \end{aligned}$$

Then $F(t,x,v) \in C^1\left(\bar{\Pi}\right)$.

From now on, ∇_x^T denotes the tangential derivative, $\nabla_x^{\perp}$ denotes the normal derivative, v^T denotes the tangential component of v, and $v^{\perp}$ denotes the normal component of v.

THEOREM 3.3. *Let $1 \le p \le \infty$, and let*

$$f_0 \in W^{1,p}(\Pi_0), \quad g \in W^{1,p}(\gamma^+),$$
$$|\nabla g| \le C\,|n_x \cdot v|. \tag{3.8}$$

Let f_0 and g have compact support and satisfy

$$f_0(x,v) = g(0,x,v) \text{ for all } x \in \partial\Omega \text{ and } v \text{ with } n_x \cdot v < 0.$$

Let $E(t,x) \in W^{1,\infty}\left([0,\infty) \times \bar{\Omega}\right)$ ($E \in W^{1,\infty} \cap C^1$ for $p = \infty$) and $E(t,x) \cdot n_x \ge \delta > 0$. Then there exists an $f(t,x,v) \in W^{1,p}(\Pi_s)$ for $0 \le s \le T$ such that

$$f_t + v \cdot \nabla_x f + E \cdot \nabla_v f = 0, \quad f|_{t=0} = f_0, \quad f|_{\gamma^+} = g \tag{3.9}$$

in the sense of distribution. The following estimates hold:

$$\begin{aligned}
\int_{\Pi_s} |f_t|^p + \int_{\gamma_s^-} (n_x \cdot v)\,|f_t|^p &\le \int_{\Pi_0} |f_{0t}|^p - \int_{\gamma_s^+} (n_x \cdot v)\,|g_t|^p \\
&\quad + C \int_0^s \int_{\Pi_\tau} |\nabla f|^p \, d\tau, \\
\int_{\Pi_s} \left|\nabla_x^T f\right|^p + \int_{\gamma_s^-} (n_x \cdot v)\left|\nabla_x^T f\right|^p &\le \int_{\Pi_0} \left|\nabla_x^T f_0\right|^p - \int_{\gamma_s^+} (n_x \cdot v)\left|\nabla_x^T g\right|^p \\
&\quad + C \int_0^s \int_{\Pi_\tau} |\nabla f|^p \, d\tau, \\
\int_{\Pi_s} \left|\nabla_x^\perp f\right|^p + \int_{\gamma_s^-} (n_x \cdot v)\left|\nabla_x^\perp f\right|^p &\le \int_{\Pi_0} \left|\nabla_x^\perp f_0\right|^p - \int_{\gamma_s^+} (n_x \cdot v)\left|\nabla_x^\perp g\right|^p \\
&\quad + C \int_0^s \int_{\Pi_\tau} |\nabla f|^p \, d\tau, \\
\int_{\Pi_s} |\nabla_v f|^p + \int_{\gamma_s^-} (n_x \cdot v)\,|\nabla_v f|^p &\le \int_{\Pi_0} |\nabla_v f_0|^p - \int_{\gamma_s^+} (n_x \cdot v)\,|\nabla_v g|^p \\
&\quad + C \int_0^s \int_{\Pi_\tau} |\nabla f|^p \, d\tau,
\end{aligned}$$

where

$$\begin{aligned}
\nabla_x^\perp f|_{\gamma^+} &= -(n_x \cdot v)^{-1}\left[g_t + v^T \cdot \nabla_x^T g + E \cdot \nabla_v g\right], \\
\nabla_x^\perp g &= -(n_x \cdot v)^{-1}\left[g_t + v^T \cdot \nabla_x^T g + E \cdot \nabla_v g\right], \\
\nabla_x^\perp f|_{t=0} &= f_{0x}, \qquad 0 \le s \le T.
\end{aligned}$$

Proof. Let N_σ be a σ-neighborhood of $\gamma^0 = \{x \in \partial\Omega,\ n_x \cdot v = 0\}$.
We construct $f_0^n \in C_c^\infty$, which is constant on N_σ, and $E^n \in C^\infty$ such that

$$f_0^n \to f_0 \text{ in } W^{1,p}(\Pi_0 \backslash N_\sigma), \quad E^n \to E \text{ in } W^{1,\infty}, \quad E^n \cdot n_x \ge \delta/2 > 0.$$

After choosing f_0^n and E^n, we construct g^n such that

$$\begin{aligned}
g^n(0,x,v) &= f_0^n(x,v) \text{ on } \gamma^+ \backslash N_\sigma, \\
g_t^n(0,x,v) &= -v \cdot \nabla_x f_0^n - E^n(0,x) \cdot \nabla_v f_0^n \text{ on } \gamma^+ \backslash N_\sigma, \\
g^n &\to g \text{ in } W^{1,p}\left(\gamma^+ \backslash N_\sigma\right).
\end{aligned}$$

We first choose $g^n \to g$ in $W^{1,p}(\gamma^+\backslash N_\sigma)$, where $g^n \in C_c^\infty$. Then modify it as $g^n(t,x,v)+[f_0^n(x,v)-g^n(0,x,v)]+t\chi(t)[-v\cdot\nabla_x f_0^n - E^n(0,x) - g_t^n(0,x,v)]$, where $\chi(t)\in C^\infty$, $\chi(0)=1$, and $\int|\chi(t)|^p\,dt$ is very small. We can see that this sequence satisfies all the above conditions. Clearly, from Theorem 3.1 there exists $f^n \in C^1\left(\bar{\Pi}\backslash\gamma^0\right)$ such that f^n satisfies (3.9) with initial and boundary data f_0^n and g^n, respectively. By applying the Gronwall inequality and since $|n_x\cdot v|\geq\sigma$, we deduce that

$$\|f^n\|_{W^{1,p}(\Pi\backslash N_\sigma)} \leq C_\sigma,$$
$$\|f^n\|_{W^{1,p}(\gamma^-\backslash N_\sigma)} \leq C_\sigma$$

for all $\sigma>0$, uniformly in n. By letting $n\to\infty$, we show that for $1<p<\infty$ there exists an f in $W^{1,p}$ such that

$$f^n \rightharpoonup f \text{ in } W^{1,p}(\Pi\backslash N_\sigma)\cap W^{1,p}(\gamma\backslash N_\sigma)$$

for all $\sigma>0$, and our theorem thus follows. For $p=1$, we show that $\{f^n\}$ is a Cauchy sequence in W^1 by considering f^n-f^m in (3.9). From the equation for f^n-f^m,

$$\partial_t(f^n-f^m)+v\cdot\nabla_x(f^n-f^m)+E^n\cdot\nabla_v(f^n-f^m)=(E^m-E^n)\cdot\nabla_v f^m,$$

we take derivatives (in the sense of distribution) and integrate with respect to x and v and then with respect to time. Using that $\{f_0^n\},\{g^n\}$ are Cauchy in $W^{1,1}$ and f^n and its derivatives are compactly supported uniformly in n (since $\|E^n\|_{W_{1,\infty}}$ are uniformly bounded in n and f_0^n, g^n are compactly supported), we can deduce that $\{f^n\}$ is a Cauchy sequence. For $p=\infty$, we use $E\in C^1$ itself instead of using approximate fields E^n in our construction of f_0^n, g^n to apply Theorem 3.1. Then we have, by taking derivatives,

$$\partial_t\partial(f^n-f^m)+v\cdot\nabla_x(\partial(f^n-f^m))+E\cdot\nabla_v(\partial(f^n-f^m))=\partial E\cdot\nabla_v(f^m-f^n).$$

By integrating along the corresponding trajectory, we get the Gronwall inequality for $\partial(f^n-f^m)$, which implies that $\{f^n\}$ is a Cauchy sequence in $W^{1,\infty}$. Here we note that by the vanishing assumption (3.8),

$$\|f\|_{W^{1,p}(\Pi)}\leq C\left(1+\|f_0\|_{W^{1,p}(\Pi_0)}+\|g\|_{W^{1,p}(\gamma^+)}\right),$$

where C depends on T, the support of f_0 and g, and the constant on the vanishing condition on g, $\|E\|_{W^{1,\infty}}$. Our theorem thus follows. □

We also deduce the following theorem.

THEOREM 3.4. *Let $1\leq p\leq\infty$, and let $F_0(x,v)$, $H(t,x,v)$, and $G(t,x,v)$ be $(n\times 1)$-vector-valued functions and $A(t,x,v)$ be an $(n\times n)$-matrix function such that $G(t,x,v)\in W^{1,p}(\gamma^+)$, $F_0(x,v)\in W^{1,p}(\Pi_0)$, $H(\cdot,x,v)\in W^{1,p}(\Pi_s)\cap W^{1,p}(\gamma^+)$ for $0\leq s\leq T$, and $A\in C^{0,1}(\Pi)$. Let the vanishing condition hold on $\partial\Omega$:*

$$|\nabla G|\leq C|n_x\cdot v|,\quad |\nabla H|\leq C|n_x\cdot v|.$$

Let F_0 and G have compact support in v, and let F_0 and G satisfy

$$F_0(x,v)=G(0,x,v) \text{ for all } x\in\partial\Omega \text{ and } v \text{ with } n_x\cdot v<0.$$

Let $E(t,x)\in W^{1,\infty}\left([0,\infty)\times\bar{\Omega}\right)$ ($E\in W^{1,\infty}\cap C^1$ for $p=\infty$) and $E(t,x)\cdot n_x\geq\delta>0$. Then there exists an $F(t,x,v)\in W^{1,p}(\Pi_s)\cap W^{1,p}(\gamma^+)$ for $0\leq s\leq T$ such that

$$F_t+v\cdot\nabla_x F+E\cdot\nabla_v F=A(F)+H,\quad F|_{t=0}=F_0,\quad F|_{\gamma^+}=G$$

in the sense of distribution. The following estimates hold:

$$\begin{aligned}
\int_{\Pi_s} |F_t|^p + \int_{\gamma_s^-} (n_x \cdot v) |F_t|^p &\le \int_{\Pi_0} |F_{0t}|^p - \int_{\gamma_s^+} (n_x \cdot v) |G_t|^p \\
&\quad + C \int_0^s \int_{\Pi_\tau} (|\nabla F|^p + |\nabla H|^p) \, d\tau, \\
\int_{\Pi_s} \left|\nabla_x^T F\right|^p + \int_{\gamma_s^-} (n_x \cdot v) \left|\nabla_x^T F\right|^p &\le \int_{\Pi_0} \left|\nabla_x^T F_0\right|^p - \int_{\gamma_s^+} (n_x \cdot v) \left|\nabla_x^T G\right|^p \\
&\quad + C \int_0^s \int_{\Pi_\tau} (|\nabla F|^p + |\nabla H|^p) \, d\tau, \\
\int_{\Pi_s} \left|\nabla_x^\perp F\right|^p + \int_{\gamma_s^-} (n_x \cdot v) \left|\nabla_x^\perp F\right|^p &\le \int_{\Pi_0} \left|\nabla_x^\perp F_0\right|^p - \int_{\gamma_s^+} (n_x \cdot v) \left|\nabla_x^\perp G\right|^p \\
&\quad + C \int_0^s \int_{\Pi_\tau} (|\nabla F|^p + |\nabla H|^p) \, d\tau, \\
\int_{\Pi_s} |\nabla_v F|^p + \int_{\gamma_s^-} (n_x \cdot v) |\nabla_v F|^p &\le \int_{\Pi_0} |\nabla_v F_0|^p - \int_{\gamma_s^+} (n_x \cdot v) |\nabla_v G|^p \\
&\quad + C \int_0^s \int_{\Pi_\tau} (|\nabla F|^p + |\nabla H|^p) \, d\tau,
\end{aligned}$$

where $\nabla_x^\perp F|_{\gamma^+} = -(n_x \cdot v)^{-1} \left[G_t + v^T \cdot \nabla_x^T G + E \cdot \nabla_v G\right]$, $\nabla_x^\perp F|_{t=0} = F_{0x}$, *and* $\nabla_x^\perp G = -(n_x \cdot v)^{-1} \left[G_t + v^T \cdot \nabla_x^T G + E \cdot \nabla_v G\right]$, $0 \le s \le T$.

DEFINITION 3.5 (boundary and initial operators). *Suppose that* $f \in C_c^\infty$ *satisfies*

$$f_t + v \cdot \nabla_x f + E \cdot \nabla_v f = 0, \quad f|_{t=0} = f_0, \quad f|_{\gamma^+} = g$$

in the classical sense. The unique boundary operator L_+ *and the unique initial operator* L_0 *are defined by*

$$\partial^\alpha f|_{\gamma^+} = L_+^\alpha \left(\nabla_x^T, \partial_v, \partial_t\right) f|_{\gamma^+}, \quad \partial^\alpha f|_{t=0} = L_0^\alpha \left(\partial_x, \partial_v\right) f|_{t=0},$$

where ∂ *is the usual differential operator of* t, x, v *with multi-index* α; $|\alpha|$ *is the order of* α.

For the higher regularity, we refer the reader to [7], [8].

THEOREM 3.6 (high regularity). *Suppose that* $E(t,x,v) \cdot n_x \ge \delta > 0$ *on* $\partial\Omega$, $E \in W^{k,\infty}$. *Let* $0 \le f_0 \in W^{k,p}(\Pi_0)$, f_0 *have compact support in* v, $0 \le g \in W^{k,p}(\gamma^+)$, *and* g *have compact support in* v. *Let* $\partial^\alpha g(t,x,v) = 0$ *for* $x \in \partial\Omega$, $v \cdot n_x = 0$, $|\alpha| = k$, *and let* $\left|\partial^{(k)} g\right| \le C |v \cdot n_x|^k$. *Assume that the following compatibility conditions are satisfied:*

$$\partial^\alpha f|_{\{t=0\} \cap \gamma^+} = \left\{L_+^\alpha \left(\nabla_x^T, \partial_v, \partial_t\right) g\right\}|_{t=0} = \left\{L_0^\alpha \left(\partial_x, \partial_v\right) f_0\right\}|_{\gamma^+}, \tag{3.10}$$

where $|\alpha| \le k-1$. *Then there is a unique* $W^{k,p}$ *solution* f *such that*

$$\int_{\Pi_s} \left|\partial^\theta f\right|^p + \int_{\gamma_s^-} (v \cdot n_x) \left|\partial^\theta f\right|^p \le C, \quad 0 \le s \le T,$$

where C *depends on* f_0, g, *and* E, $|\theta| \le k$. *For* $|\alpha| \le k$,

$$\partial^\alpha f|_{\gamma^+} = L_+^\alpha \left(\nabla_x^T, \partial_v, \partial_t\right) g, \quad \partial^\alpha f|_{t=0} = L_0^\alpha \left(\partial_x, \partial_v\right) f_0.$$

Sketch of proof. We use an induction on the order k. We omit the detailed proof; instead we shall prove our theorem with $|\alpha| = k = 2$:

$$\begin{aligned}
\partial_t\partial_t f + v\cdot\nabla_x\partial_t f + E\cdot\nabla_v\partial_t f &= -\partial_t E\cdot\nabla_v f,\\
\partial_t\partial_v f + v\cdot\nabla_x\partial_v f + E\cdot\nabla_v\partial_v f &= -\partial_x f,\\
\partial_t\partial_x f + v\cdot\nabla_x\partial_x f + E\cdot\nabla_v\partial_x f &= -\partial_x E\cdot\nabla_v f.
\end{aligned}$$

We think of $F = (\partial_t f, \partial_x f, \partial_v f)$ as an unknown vector-valued function. We already know that $F \in L^p$ by Theorem 3.3, and we want to prove here that F is actually in $W^{1,p}$. Theorem 3.4 applies to this case with $H = \vec{0}$, and

$$A = \begin{pmatrix} 0 & 0 & -\partial_t E \\ 0_3 & -I_3 & 0_3 \\ 0_3 & 0_3 & -\partial_x E \end{pmatrix},$$

which is in $W^{1,\infty}$ by assumption, where $F_0 = (-v\cdot\nabla_x f_0 - E\cdot\nabla_v f_0, \partial_x f_0, \partial_v f_0) \in W^{1,p}(\Pi_0)$, $G = \left(g_t, \partial_x^T g, -(v\cdot n_x)^{-1}\left[g_t + v^T\cdot\nabla_x^T g + E\cdot\nabla_v g\right], \partial_v g\right) \in W^{1,p}(\gamma^+)$. By our assumption that $\left|\partial^2 g\right| \le C\left|v\cdot n_x\right|^2$, $\|E\|_{W^{1,\infty}} \le M$, we have

$$|\nabla G| \le C\left|v\cdot n_x\right|.$$

Moreover, by our compatibility condition, we have

$$F_0(x,v) = G(0,x,v).$$

We then apply Theorem 3.4 to get $F \in W^{1,p}$. The theorem thus follows. ∎

4. Regularity for linear specular reflection. Now we study the purely specular problem:

$$\begin{aligned}
f_t + v\cdot\nabla_x f + E\cdot\nabla_v f = 0, \quad f|_{t=0} = f_0,\\
f(t,x,v) = f(t,x,v_*), \quad x\in\partial\Omega.
\end{aligned} \tag{4.1}$$

We seek the compatibility conditions. After the change of coordinates (flattening out the boundary), we transform the original Vlasov–Poisson system into an equivalent system. Using the same notation (t,x,v) and f, we have

$$f_t + v\cdot\nabla_x f + (E+J)\cdot\nabla_v f = 0,$$

where $J_1(x,v) = (v_2,v_3)\cdot\partial^2\phi(x)\cdot(v_2,v_3)$, $J_i(x,v) = 0$ for $j = 2,3$. From the specular reflection condition on f, we have $f(t,0,\bar{x},v_1,\bar{v}) = f(t,0,\bar{x},-v_1,\bar{v})$ for all $\bar{x}\in\mathbb{R}^2$ and $v\in\mathbb{R}^3$, which also implies that $f_0(0,\bar{x},v_1,\bar{v}) = f_0(0,\bar{x},-v_1,\bar{v})$. By taking the t-derivative and plugging in $t = 0$, we get

$$f_t(0,0,\bar{x},v_1,\bar{v}) = f_t(0,0,\bar{x},-v_1,\bar{v}),$$

where

$$\begin{aligned}
f_t(0,0,\bar{x},v_1,\bar{v}) = &-v_1 f_{0x_1}(0,\bar{x},v_1,\bar{v}) - v_2 f_{0x_2}(0,\bar{x},v_1,\bar{v}) - v_3 f_{0x_3}(0,\bar{x},v_1,\bar{v})\\
&-\sum_{i=1}^{3}(E_i+J_i) f_{0v_i}(0,\bar{x},v_1,\bar{v}),\\
f_t(0,0,\bar{x},-v_1,\bar{v}) = &\,v_1 f_{0x_1}(0,\bar{x},-v_1,\bar{v}) - v_2 f_{0x_2}(0,\bar{x},-v_1,\bar{v}) - v_3 f_{0x_3}(0,\bar{x},-v_1,\bar{v})\\
&-\sum_{i=1}^{3}(E_i+J_i) f_{0v_i}(0,\bar{x},-v_1,\bar{v}).
\end{aligned}$$

Since $f_{0v_1}(0,\bar{x},v_1,\bar{v}) = -f_{0v_1}(0,\bar{x},v_1,\bar{v})$, we get

$$v_1 f_{0x_1}(0,\bar{x},v_1,\bar{v}) + v_1 f_{0x_1}(0,\bar{x},-v_1,\bar{v}) + 2(E_1+J_1)(0,0,\bar{x},v) f_{0v_1}(0,\bar{x},v_1,\bar{v}) = 0. \tag{4.2}$$

Therefore, the corresponding compatibility conditions under the original coordinate system are

$$f_0(x,v) = f_0(x,v_*), \tag{4.3}$$

$$v_*^{\perp}\nabla_x^{\perp} f_0(x,v_*) + v^{\perp}\nabla_x^{\perp} f_0(x,v) + 2E^{\perp}(0,x)\nabla_v^{\perp} f_0(x,v) = 0 \tag{4.4}$$

for all $x \in \partial\Omega$. Assume that $E(t,x) \in C^1$ and $E(t,x)\cdot n_x \geq \delta > 0$ at the boundary. We also assume that $f_0 \in C^1$ and has compact support, and

$$f_0 \equiv \text{constant} \quad \text{when } \xi^2(x) + (v\cdot\nabla\xi(x))^2 \leq \omega_0$$

for some fixed $\omega_0 > 0$. We then define an iterating sequence as a family of the solutions of the following linear problems:

$$\begin{aligned} &f_t^{k+1} + v\cdot\nabla_x f^{k+1} + E\cdot\nabla_v f^{k+1} = 0, \quad f^{k+1}|_{t=0} = f_0, \\ &f^{k+1}(t,x,v) = f^k(t,x,v_*), \quad x\in\partial\Omega, \quad v\cdot n_x \leq 0 \end{aligned} \tag{4.5}$$

for $k = 0,1,2,\ldots$, where f_0 is a smooth extension of Π satisfying the compatibility conditions. Since $\|E\|_{L^\infty} \leq C$, it easily follows that f^k has a uniform bound for its support in x and v. The major result in this section is the following theorem.

THEOREM 4.1. *Let $E(t,x)\cdot n_x = E_0(x) > 0$ for all $x \in \partial\Omega$. Let f_0 have compact support, and assume that when $\xi^2(x) + (v\cdot\nabla\xi(x))^2 \leq \omega_0$ for some fixed $\omega_0 > 0$,*

$$f_0(x,v) \equiv \text{constant}.$$

(a) *Assume $f_0 \in C^1$, $E \in C^0([0,T]\times\Omega)$, and*

$$\sup_{0\leq t\leq T} \|\nabla_x E\|_{C^0(\Omega)}(t) < \infty.$$

Let f_0 satisfy (4.3). Then there exists a unique $W^{1,\infty}$ solution f of (4.1), and $\|f\|_{W^{1,\infty}}$ depends only on ω_0, $\|E\|_{C^0} + \sup_{0\leq t\leq T}\|\nabla_x E\|_{C^0(\Omega)}(t)$, $\|E_0\|_{C^1}$, and $\|f_0\|_{C^1}$.

(b) *Moreover, if $f_0 \in C^{1,\eta}$ for some $\eta > 0$, assume that $E \in C^{0,\eta}([0,T]\times\Omega)$ and*

$$\sup_{0\leq t\leq T} \|\nabla_x E\|_{C^{0,\eta}(\Omega)}(t) < \infty.$$

Let f_0 satisfy both (4.3) and (4.4). Then there exists a unique $C^{1,\mu}$ solution f of (4.1), for some $0 < \mu < \eta$ depending on ω_0, $\|E\|_{C^{0,\eta}} + \sup_{0\leq t\leq T}\|\nabla_x E\|_{C^{0,\eta}(\Omega)}(t)$, $\|E_0\|_{C^{1,\eta}}$, and $\|f_0\|_{C^{1,\eta}}$.

We first show a uniform C^1 bound for the iterating sequence f^k.

LEMMA 4.2 (C^1 bounds). *Suppose that $E(t,x)\cdot n_x = E_0(x) > 0$ for all $x \in \partial\Omega$, $E, E_0 \in C^1$, $f_0 \in C^{1,\beta}$, and*

$$f_0 \equiv \text{constant} \quad \text{when } \xi^2(x) + (v\cdot\nabla\xi(x))^2 \leq \omega_0,\ \omega_0 > 0.$$

Suppose that (4.5) has a solution f^k. Let $(0,x_0,v_0)$ be on the back-time cycle of (t,x,v). Then

$$\left|\nabla_{(t,x,v)} f(t,x,v)\right| \leq C\left|\nabla_{(x,v)} f_0(x_0,v_0)\right|,$$

where C is independent of k, and depends on f_0, ω_0, E_0, and E.

Proof. Let the back-time cycle from (t,x,v) be $\left(t^l,x^l,v^l\right)=(t,x,v)$, $\left(t^{l-1},x^{l-1},v^{l-1}\right),\ldots,\left(t^1,x^1,v^1\right),(0,x_0,v_0)$. We need only to consider when $\xi(x)+\left(v\cdot\nabla\xi(x)\right)^2$ is small. We distinguish, on the back-time cycle from (t,x,v) to (t_0,x_0,v_0), large-time intervals from small-time intervals in the following way. By Lemma 2.7, we may assume that $\left|v^j\cdot n_{x^j}\right|\geq c\,\omega_0^{1/2}$ for all j, since $\nabla f_0=0$ when $\xi^2(x)+|v\cdot\nabla\xi(x)|^2\leq\omega_0$. If $t^j-t^{j-1}\geq c\,\omega_0^{1/2}$, then it is called a large-time interval on the cycle, otherwise a small-time interval with $t^j-t^{j-1}\leq c\,\omega_0^{1/2}\leq\left|v^j\cdot n_{x^j}\right|$. We first treat the portion of our back-time cycle with small-time intervals, which is more complicated but crucial to estimate. Without loss of generality, we use the same back-time cycle as above for our convenience in dealing with small-time intervals. From our construction (4.5), we have

$$\begin{aligned}f^k(t,x,v)&=f^k\left(t^{l-1},x^{l-1},v^{l-1}\right)=f^{k-1}\left(t^{l-1},x^{l-1},v_*^{l-1}\right)\\&=f^{k-1}\left(t^{l-2},x^{l-2},v^{l-2}\right)=\cdots=f_0(x_0,v_0).\end{aligned}$$

From the first relation that $f^k(t,x,v)=f^k\left(t^{l-1},x^{l-1},v^{l-1}\right)$, with $x^{l-1}\in\partial\Omega$, $v^{l-1}\cdot n_{x^{l-1}}<0$, we have

$$\left|\nabla f^k(t,x,v)\right|=\left|I_{l-1}^l\nabla f^k\left(t^{l-1},x^{l-1},v^{l-1}\right)\right|,$$

where

$$I_{l-1}^l=\begin{pmatrix}\frac{\partial t^{l-1}}{\partial t}&\frac{\partial x^{l-1}}{\partial t}&\frac{\partial v^{l-1}}{\partial t}\\\frac{\partial t^{l-1}}{\partial x}&\frac{\partial x^{l-1}}{\partial x}&\frac{\partial v^{l-1}}{\partial x}\\\frac{\partial t^{l-1}}{\partial v}&\frac{\partial x^{l-1}}{\partial v}&\frac{\partial v^{l-1}}{\partial v}\end{pmatrix}.$$

By Lemmas 2.10 and 2.11, I_{l-1}^l takes the form

$$\begin{pmatrix}C&C&C\\C(\omega_0)&C(\omega_0)&C(\omega_0)\\C&C&C\end{pmatrix}.$$

Here we have used Lemma 2.7 and the assumption to get

$$\begin{aligned}\left|\frac{\partial t^{l-1}}{\partial x}\right|&\leq C+\frac{1}{\left|v^{l-1}\cdot n_{x^{l-1}}\right|}\leq C+\frac{1}{C\left[\xi(x_0)+\left(v_0\cdot\nabla\xi(x_0)\right)^2\right]^{1/2}}\\&\leq C+\frac{C}{C\omega_0}=C(\omega_0),\end{aligned}$$

where C depends on $\|E\|_{C^1}$. Next we consider

$$f^k\left(t^{l-1},x^{l-1},v^{l-1}\right)=f^{k-1}\left(t^{l-1},x^{l-1},v_*^{l-1}\right),$$

where $v_*^{l-1}=v^{l-1}-2\left(v^{l-1}\cdot n_{x^{l-1}}\right)n_{x^{l-1}}$. Clearly, we have

$$\left|\nabla f^k\left(t^{l-1},x^{l-1},v^{l-1}\right)\right|=\left|I_*^{l-1}\nabla f^{k-1}\left(t^{l-1},x^{l-1},v_*^{l-1}\right)\right|,$$

where I_*^{l-1} takes the form

$$\begin{pmatrix}1&0&0\\0&Id&C\\0&0&C\end{pmatrix}.$$

Now consider the map $\left(t^{j-1}, x^{j-1}, v^{j-1}\right) \longmapsto \left(t^j, x^j, v_*^j\right)$, where $x^{j-1}, x^j \in \partial\Omega$, $2 \leq j \leq l$. Let J_{j-1}^j be the Jacobian matrix of the map. We now estimate it. Since $\left(t^{j-1}, x^{j-1}, v^{j-1}\right)$ connects with $\left(t^j, x^j, v_*^j\right)$ through a trajectory,

$$\begin{aligned} v_*^j &= v^{j-1} + \int_{t^{j-1}}^{t^j} E(\tau)\, d\tau, \\ x^j &= x^{j-1} + v_*^j \left(t^j - t^{j-1}\right) + \int_{t^{j-1}}^{t^j} \int_{t^j}^{s} E(\tau)\, d\tau ds. \end{aligned}$$

Here we notice that

$$c\left(t^j - t^{j-1}\right) \leq \left|v^{j-1} \cdot n_{x^{j-1}}\right| \leq C\left(t^j - t^{j-1}\right) \tag{4.6}$$

by Lemma 2.11 and by expanding $\xi\left(x^j\right)$ around $\xi\left(x^{j-1}\right)$ as

$$\begin{aligned} 0 = \xi\left(x^j\right) &= \xi\left(x^{j-1} + v_*^{j-1}\left(t^j - t^{j-1}\right) + \int_{t^{j-1}}^{t^j} \int_{t^j}^{s} E(\tau)\, d\tau ds\right) \\ &= \left(v_*^{j-1} \cdot \nabla\xi\left(x^{j-1}\right)\right)\left(t^j - t^{j-1}\right) + O\left(t^j - t^{j-1}\right)^2. \end{aligned}$$

We then get a similar estimate

$$\left|\nabla f^{k-1}\left(t^j, x^j, v_*^j\right)\right| = \left|J_{j-1}^j \nabla f^{k-1}\left(t^{j-1}, x^{j-1}, v^{j-1}\right)\right|,$$

where

$$J_{j-1}^j = \begin{pmatrix} C & C & C \\ C(\omega_0) & C(\omega_0) & C(\omega_0) \\ C & C & C \end{pmatrix},$$

and C depends only on the C^1 norm of E. Since $\left(t^1, x^1, v^1\right)$ connects with $(0, x_0, v_0)$, with $x_1 \in \partial\Omega$, we have

$$\begin{aligned} \left|\nabla f^{k-(l-2)}\left(t^1, x^1, v^1\right)\right| &= \left|I_0^1 \nabla f_0(x_0, v_0)\right| \\ &\leq C\left|\nabla f_0(x_0, v_0)\right|. \end{aligned}$$

By Lemma 2.7 and (4.6), we have $\left|v^i \cdot n_{x^i}\right| \geq c(\omega_0)$ for all i and $l \times c(\omega_0) \leq \sum_i \left|v^i \cdot n_{x^i}\right| \leq \sum\left(t^{i+1} - t^i\right) \leq T$ to see that the number of bounces is uniformly bounded and dependent on ω_0, δ, and $\|E\|_{C^1}$. Therefore,

$$\begin{aligned} \left|\nabla f^k(t, x, v)\right| &= \left|I_{l-1}^l \prod_{j=1}^{l-1} I_*^j \prod_{j=2}^{l-1} J_{j-1}^j I_0^1 \nabla f_0(x_0, v_0)\right| \\ &\leq C(\omega_0)\left|\nabla f_0(x_0, v_0)\right|. \end{aligned}$$

In a similar manner, we can obtain estimates on the portion of the cycle with large-time intervals. Here we have the uniform bound on the number of bounces with large-time intervals in a rather trivial way since the size of each interval was chosen to be $\Delta t \geq c\,\omega_0^{1/2}$ by the construction. We thus deduce our lemma. □

LEMMA 4.3. *Suppose that* $E(t,x)\cdot n_x = E_0(x) > 0$ *at the boundary and that* (4.5) *has the solution* f^k. *Then*

$$\left\|f^k\right\|_{L^\infty(\Pi)} \le \|f_0\|_{L^\infty(\Pi_0)} \quad \textit{for all } k.$$

Proof. Let the cycle from (t,x,v) be $\left(t^i,x^i,v^i\right)$, $1\le i\le l$, and $(0,x_0,v_0)$. Clearly, on each trajectory f^k is a constant. ∎

We establish a uniform $C^{0,\mu}$ estimate for f^k in (4.5) by using Corollaries 2.2 and 2.3.

LEMMA 4.4. (a) *The sequence is well defined, and* $f^k \in C^1$.

(b) *If* $|x-y|$ *is small and the field satisfies*

$$\sup_{0\le t\le T} |E(t,x)-E(t,y)| \le -L|x-y|\log|x-y|,$$

then there is a $\omega>0$, *depending on* L *and* $\|E\|_\infty$, *such that if* $\xi^2(x)+(v\cdot\nabla\xi(x))^2 \le \omega$,

$$f^k(t,x,v) \equiv \text{constant} \quad \textit{for any } k.$$

(c) *Moreover, for constants* C *and* $\mu>0$ *depending on* L, ω, *and* α,

$$\left\|f^k\right\|_{C^{0,\mu}} \le C\|f_0\|_{C^{0,\alpha}}.$$

Proof. For (a), we apply Theorem 3.1. From the velocity lemma, Lemma 2.1, and Lemma 4.2, $\left|\nabla_{(t,x,v)} f^k(t,x,v)\right| \equiv 0$ when $|v\cdot n_x| \le C(\omega_0)$ to satisfy the vanishing condition (3.3). It suffices to check for any k that the compatibility condition in Theorem 3.1 is satisfied. First, it is trivial to see that $f_0(x,v) = f^k(0,x,v_*)$ by (4.2). We use an induction on k. Clearly, it is true for $k=0$ if we choose f_0 properly. Supposing that the condition for $k=n-1$ is true, we deduce from (4.5) and (4.2) that for $x\in\partial\Omega$,

$$\begin{aligned} f_t^n(0,x,v) &= f_t^{n-1}(t,x,v_*)\,|_{t=0} \\ &= -v_*\cdot\nabla_x f^{n-1}(0,x,v_*) - E(0,x)\cdot\nabla_v f^{n-1}(0,x,v_*) \\ &= -v_*\cdot\nabla_x f_0(x,v_*) - E(0,x)\cdot\nabla_v f_0(x,v_*) \\ &= -v\cdot\nabla_x f_0(x,v) - E(0,x)\cdot\nabla_v f_0(x,v). \end{aligned}$$

This is exactly (3.2) in Theorem 3.1.

For part (b), we use Corollary 2.3, since

$$\sup_{0\le t\le T} |E(t,x)-E(t,y)| \le C|x-y|^{1/2}.$$

For any $(t,x,v)\in\bar\Pi$ and for all k, let the back-time cycle of (t,x,v) be $\left(t^{l-1},x^{l-1},v^{l-1}\right)$, $\ldots$, $(0,x_0,v_0)$. From (4.5),

$$\begin{aligned} f^k(t,x,v) &= f^k\left(t^{l-1},x^{l-1},v^{l-1}\right) \\ &= f^{k-1}\left(t^{l-1},x^{l-1},v_*^{l-1}\right) \\ &= f^{k-1}\left(t^{l-2},x^{l-2},v^{l-2}\right) \\ &\vdots \\ &= f_0(x_0,v_0). \end{aligned}$$

By Corollary 2.8, we have

$$\begin{aligned} C_1\left[\xi^2(x)+\left(v\cdot\nabla\xi(x)\right)^2\right] &\le \left(v^i\cdot\nabla\xi\left(x^i\right)\right)^2 \le C_2\left[\xi^2(x_0)+\left(v_0\cdot\nabla\xi(x_0)\right)^2\right], \\ C_1\left[\xi^2(x_0)+\left(v_0\cdot\nabla\xi(x_0)\right)^2\right] &\le \left(v^i\cdot\nabla\xi\left(x^i\right)\right)^2 \le C_2\left[\xi^2(x)+\left(v\cdot\nabla\xi(x)\right)^2\right], \end{aligned}$$

for $1\le i\le l$, and C_1 and C_2 depend only on $L,\|E_0\|_{C^1}$. Let $\omega=C_1\omega_0$. Then we clearly have the conclusion of (b).

We omit the proof of part (c) and refer the reader to [8] for the proof in the case of a half space with a flat boundary. □

Now we are ready to prove Theorem 4.1.

Proof. From Lemma 4.2, $\left\|f^k\right\|_{C^1}$ is bounded uniformly in k, and it suffices to show that the iterated sequence f^k defined in (4.5) is indeed uniformly bounded in $C^{1,\alpha}$. Now we pick two points (t,x,v) and $\left(\tilde{t},\tilde{x},\tilde{v}\right)$. Consider the back-time cycles through the two points. Let $\varepsilon=\left(\left|t-\tilde{t}\right|+|x-\tilde{x}|+|v-\tilde{v}|\right)$. We keep track of the difference of these two points case by case.

Case 1. Both of the trajectories emanate from $\{t=0\}$.

This reduces to the Cauchy problem and the theory of ordinary differential equations.

Case 2. One trajectory emanates from $\{t=0\}$, and the other one emanates from the boundary $\partial\Omega$.

We first note that $\xi^2(x)+(v\cdot\nabla\xi(x))^2\ge\omega>0$ and $\xi^2(\tilde{x})+(\tilde{v}\cdot\nabla\xi(\tilde{x}))^2\ge\omega>0$ from the velocity lemma, Lemma 2.1, since otherwise $f^k\equiv$ constant. In this case, we have

$$\begin{aligned} v_0 &= v+\int_t^0 E(\tau)\,d\tau, \quad x=x_0+vt+\int_0^t\int_t^s E(\tau)\,d\tau ds, \\ \tilde{v}^1 &= \tilde{v}+\int_{\tilde{t}}^{\tilde{t}^1} E(\tau)\,d\tau, \quad \tilde{x}=\int_{\tilde{t}^1}^{\tilde{t}}\left[\tilde{v}+\int_{\tilde{t}}^s E(\tau)\,d\tau\right]ds. \end{aligned}$$

We choose a third point $\left(\hat{t},\hat{x},\hat{v}\right)$ such that $\left(\hat{t},\hat{x},\hat{v}\right)$ connects with $\left(0,\tilde{x}^1,v_0\right)$ through a trajectory and satisfies

$$\begin{aligned} &\left|t-\hat{t}\right|+\left|\hat{t}-\tilde{t}\right|\le 2\left|t-\tilde{t}\right|, \quad |x-\hat{x}|+|\hat{x}-\tilde{x}|\le 2|x-\tilde{x}|, \\ &|v-\hat{v}|+|\hat{v}-\tilde{v}|\le 2|v-\tilde{v}|. \end{aligned}$$

We can apply Lemmas 2.9 and 2.10 and the mean value theorem through the third point $\left(\hat{t},\hat{x},\hat{v}\right)$ to get

$$\left|\tilde{t}^1\right|+\left|x_0-\tilde{x}^1\right|+\left|v_0-\tilde{v}^1\right|\le C\left[\left|t-\tilde{t}\right|+|x-\tilde{x}|+|v-\tilde{v}|\right]=C\varepsilon,$$

where C depends on $\|E\|_{C^1},\delta,\omega$. From the fact that $\left|\nabla\xi\left(\tilde{x}^1\right)\cdot\tilde{v}_*^1\right|\ge\omega>0$ and $\nabla\xi\left(\tilde{x}^1\right)\cdot\tilde{v}_*^1<0$, we know that $\nabla\xi\left(\tilde{x}^1\right)\cdot\tilde{v}_*^1<-\omega<0$. Since

$$\left|\left[\nabla\xi\left(X\left(\tau\right)\right)\cdot V\left(\tau\right)\right]^{\bullet}\right| = \left|V\cdot\nabla^{2}\xi\cdot V+\nabla\xi\cdot E\right| \leq C,$$

we have

$$\begin{aligned}\nabla\xi\left(X\left(\tau\right)\right)\cdot V\left(\tau\right) &= \nabla\xi\left(\tilde{x}^{1}\right)\cdot\tilde{v}_{*}^{1}+\int_{\tilde{t}^{1}}^{\tau}\frac{d}{ds}\left[\nabla\xi\left(X\left(s\right)\right)\cdot V\left(s\right)\right]ds\\ &\leq \nabla\xi\left(\tilde{x}^{1}\right)\cdot\tilde{v}_{*}^{1}+O\left(\tilde{t}^{1}-\tau\right)\\ &\leq -c\left(\omega\right)+O\left(\varepsilon\right)<0\end{aligned}$$

for $0\leq\tau\leq\tilde{t}^{1}$. This means that the trajectory from $\left(\tilde{t}^{1},\tilde{x}^{1},\tilde{v}^{1}\right)$ hits $\{t=0\}$ directly and does not hit γ^{+}. Now we express $\nabla_{(t,x,v)}f^{k}\left(t,x,v\right)$ and $\nabla_{(t,x,v)}f^{k}\left(\tilde{t},\tilde{x},\tilde{v}\right)$ in terms of the initial value $\nabla_{(x,v)}f_{0}$. It will turn out that the compatibility condition (4.2) exactly guarantees our theorem in this case.

For computational simplicity, we flatten out the boundary near $\left(\tilde{t}^{1},\tilde{x}^{1},\tilde{v}^{1}\right)$ and $(0,x_{0},v_{0})$. We choose $\left(t^{'},x^{'},v^{'}\right)$ and $\left(\tilde{t}^{'},\tilde{x}^{'},\tilde{v}^{'}\right)$ near $\partial\Omega$ such that $\left(t^{'},x^{'},v^{'}\right)$ is on the trajectory from (t,x,v) to $(0,x_{0},v_{0})$, between (t,x,v) and $(0,x_{0},v_{0})$, and $\left(\tilde{t}^{'},\tilde{x}^{'},\tilde{v}^{'}\right)$ is on the trajectory from $\left(\tilde{t},\tilde{x},\tilde{v}\right)$ to $\left(\tilde{t}^{1},\tilde{x}^{1},\tilde{v}^{1}\right)$, between $\left(\tilde{t},\tilde{x},\tilde{v}\right)$ and $\left(\tilde{t}^{1},\tilde{x}^{1},\tilde{v}^{1}\right)$, respectively. Hence we have

$$\left|t^{'}-\tilde{t}^{'}\right|+\left|x^{'}-\tilde{x}^{'}\right|+\left|v^{'}-\tilde{v}^{'}\right|\leq C\varepsilon.$$

Since

$$\begin{aligned}\left|t_{t}^{'}-\tilde{t}_{\tilde{t}}^{'}\right|+\left|x_{t}^{'}-\tilde{x}_{\tilde{t}}^{'}\right|+\left|v_{t}^{'}-\tilde{v}_{\tilde{t}}^{'}\right|&\leq C\varepsilon,\\ \left|t_{t}^{'}\right|+\left|\tilde{t}_{\tilde{t}}^{'}\right|+\left|x_{t}^{'}\right|+\left|\tilde{x}_{\tilde{t}}^{'}\right|+\left|v_{t}^{'}\right|+\left|\tilde{v}_{\tilde{t}}^{'}\right|&\leq C,\end{aligned}$$

and $\left|\nabla f^{k}\right|\leq C$, it reduces to the case when (t,x,v) and $\left(\tilde{t},\tilde{x},\tilde{v}\right)$ are all near the boundary. Recall that in the flat coordinates, the Vlasov equation is transformed to $f_{t}^{k}+v\cdot\nabla_{x}f^{k}+(E+J)\cdot\nabla_{v}f^{k}=0$, where $J_{1}=v\cdot\partial^{2}\Phi\cdot v\leq 0$, $J_{2}=J_{3}=0$.

We first consider $f_{t}^{k}\left(t,x,v\right)$. Since the back-time trajectory emanates from $t=0$ directly, we have

$$\begin{aligned}&f_{t}^{k}\left(t,x,v\right)\\ &=\nabla_{x}f_{0}\left(x_{0},v_{0}\right)\cdot\left[-v+t\left(E+J\right)\left(t,x,v\right)-\int_{0}^{t}\int_{t}^{s}\left\{\nabla_{x}\left(E+J\right)\cdot X_{t}+\nabla_{v}J\cdot V_{t}\right\}d\tau ds\right]\\ &\quad+\nabla_{v}f_{0}\left(x_{0},v_{0}\right)\cdot\left[-\left(E+J\right)\left(t,x,v\right)+\int_{t}^{0}\left\{\nabla_{x}\left(E+J\right)\cdot X_{t}+\nabla_{v}J\cdot V_{t}\right\}d\tau\right]\end{aligned}$$

$$
\begin{aligned}
&= \nabla_x f_0\left(\tilde{x}^1, v_0\right)\cdot\left[-\tilde{v}+\left(\tilde{t}-\tilde{t}^1\right)(E+J)\left(\tilde{t},\tilde{x},\tilde{v}\right)\right.\\
&\qquad\left.-\int_{\tilde{t}^1}^{\tilde{t}}\int_{\tilde{t}}^{s}\left\{\nabla_x(E+J)\cdot X_t+\nabla_v J\cdot V_t\right\}d\tau ds\right]\\
&\quad+\nabla_v f_0\left(\tilde{x}^1, v_0\right)\cdot\left[-(E+J)\left(\tilde{t},\tilde{x},\tilde{v}\right)+\int_{\tilde{t}}^{\tilde{t}^1}\left\{\nabla_x(E+J)\cdot X_t+\nabla_v J\cdot V_t\right\}d\tau\right]\\
&\quad+O\left(\varepsilon^{\eta}\right),
\end{aligned}
$$

where we have used $\left(\tilde{t},\tilde{x},\tilde{v}\right)=(t,x,v)+O(\varepsilon)$, $\tilde{x}^1=x_0+O(\varepsilon)$, $E\in C^{1,\eta}$, and $f_0\in C^{1,\eta}$. Notice that

$$
\begin{aligned}
&-\tilde{v}_1+\left(\tilde{t}-\tilde{t}^1\right)(E_1+J_1)-\int_{\tilde{t}^1}^{\tilde{t}}\int_{\tilde{t}}^{s}\left\{\nabla_x(E_1+J_1)\cdot X_t+\nabla_v J_1\cdot V_t\right\}d\tau ds\\
&=-\tilde{t}_{\tilde{t}}^1\tilde{v}_1^1=-\tilde{t}_{\tilde{t}}^1 v_{01},
\end{aligned}
$$

since $\tilde{v}^1=v_0+O(\varepsilon)$. Therefore, we get

(4.7)

$$
\begin{aligned}
&f_t^k(t,x,v)\\
&=\left(-\tilde{t}_{\tilde{t}}^1 v_{01}\right)f_{0x_1}\left(\tilde{x}^1,v_0\right)\\
&\quad+\sum_{j=2}^{3}f_{0x_j}\left(\tilde{x}^1,v_0\right)\left[-\tilde{v}_j+\left(\tilde{t}-\tilde{t}^1\right)E_j\left(\tilde{t},\tilde{x}\right)-\int_{\tilde{t}^1}^{\tilde{t}}\int_{\tilde{t}}^{s}\nabla_x E_j\cdot X_t d\tau ds\right]\\
&\quad+f_{0v_1}\left(\tilde{x}^1,v_0\right)\left[-(E_1+J_1)\left(\tilde{t},\tilde{x},\tilde{v}\right)+\int_{\tilde{t}}^{\tilde{t}^1}\left\{\nabla_x(E_1+J_1)\cdot X_t+\nabla_v J_1\cdot V_t\right\}d\tau\right]\\
&\quad+\sum_{j=2}^{3}f_{0v_j}\left(\tilde{x}^1,v_0\right)\left[-E_j\left(\tilde{t},\tilde{x}\right)+\int_{\tilde{t}}^{\tilde{t}^1}\nabla_x E_j\cdot X_t d\tau\right]+O\left(\varepsilon^{\eta}\right),
\end{aligned}
$$

where we have used that $\tilde{t}_{\tilde{t}}^1\le C(\omega)$ and $\left|\nabla f^k\right|\le C$.

Now we treat $f_t^k\left(\tilde{t},\tilde{x},\tilde{v}\right)$. The trajectory first hits $\left(\tilde{t}^1,\tilde{x}^1,\tilde{v}^1\right)$, reflects $\left(\tilde{t}^1,\tilde{x}^1,P\tilde{v}^1\right)$ with $P\tilde{v}^1=\left(-\tilde{v}_1^1,\tilde{v}_2^1,\tilde{v}_3^1\right)$, and then hits $(0,\tilde{x}_0,\tilde{v}_0)$. We have

$$
f_t^k\left(\tilde{t},\tilde{x},\tilde{v}\right)=f_{\tilde{t}^1}^k\left(\tilde{t}_{\tilde{t}}^1\right)+f_{\tilde{x}_2^1}^k\partial_t\tilde{x}_2^1+f_{\tilde{x}_3^1}^k\partial_t\tilde{x}_3^1+\sum_{j=1}^{3}f_{\tilde{v}_j^1}^k\partial_t\tilde{v}_j^1
$$

at $\left(\tilde{t}^1,\tilde{x}^1,\tilde{v}^1\right)$. From the specular reflection condition on f, we have

$$
\begin{aligned}
&f_{\tilde{t}^1}^k\left(\tilde{t}^1,\tilde{x}^1,\tilde{v}^1\right)=f_{\tilde{t}^1}^{k-1}\left(\tilde{t}^1,\tilde{x}^1,P\tilde{v}^1\right),\quad f_{\tilde{x}_j^1}^k\left(\tilde{t}^1,\tilde{x}^1,\tilde{v}^1\right)=f_{\tilde{x}_j^1}^{k-1}\left(\tilde{t}^1,\tilde{x}^1,P\tilde{v}^1\right),\quad j=2,3,\\
&f_{\tilde{v}_1^1}^k\left(\tilde{t}^1,\tilde{x}^1,\tilde{v}^1\right)=-f_{\tilde{v}_1^1}^{k-1}\left(\tilde{t}^1,\tilde{x}^1,P\tilde{v}^1\right),\quad f_{\tilde{v}_j^1}^k\left(\tilde{t}^1,\tilde{x}^1,\tilde{v}^1\right)=f_{\tilde{v}_j^1}^{k-1}\left(\tilde{t}^1,\tilde{x}^1,\tilde{v}^1\right),\quad j=2,3.
\end{aligned}
$$

Since the trajectory finally hits $(0,\tilde{x}_0,\tilde{v}_0)$ directly, we have that

$$
\begin{aligned}
&f_t^k\left(\tilde{t},\tilde{x},\tilde{v}\right)\\
&=\left\{\nabla_x f_0\cdot\left[-P\tilde{v}^1+\tilde{t}^1(E+J)\left(\tilde{t}^1,\tilde{x}^1,\tilde{v}^1\right)\right.\right.\\
&\qquad\left.\left.-\int_0^{\tilde{t}^1}\int_{\tilde{t}^1}^{s}\left\{\nabla_x(E+J)\cdot X_t+\nabla_v J\cdot V_t\right\}d\tau ds\right]\right.
\end{aligned}
$$

$$
\begin{aligned}
&+\nabla_v f_0\cdot\left[-\left(E+J\right)\left(\tilde{t}^1,\tilde{x}^1,\tilde{v}^1\right)+\int_{\tilde{t}^1}^0\left\{\nabla_x\left(E+J\right)\cdot X_t+\nabla_v J\cdot V_t\right\}d\tau\right]\Bigg\}\tilde{t}_{\tilde{t}}^1\\
&+\Bigg\{\nabla_x f_0\cdot\left[\delta_{ij}+\int_{\tilde{t}^1}^0\int_{\tilde{t}^1}^s\left\{\nabla_x\left(E+J\right)\,\nabla_x X+\nabla_v J\,\nabla_x V\right\}d\tau ds\right]\\
&\quad+\nabla_v f_0\cdot\left[\int_{\tilde{t}^1}^0\left\{\nabla_x\left(E+J\right)\,\nabla_x X+\nabla_v J\,\nabla_x V\right\}d\tau\right]\Bigg\}\cdot\partial_{\tilde{t}}\tilde{x}^1\\
&+\Bigg\{\nabla_x f_0\cdot\left[-\tilde{t}^1\delta_{ij}-\int_0^{\tilde{t}^1}\int_{\tilde{t}^1}^s\left\{\nabla_x\left(E+J\right)\,\nabla_x X+\nabla_v J\,\nabla_x V\right\}d\tau ds\right]\\
&\quad+\nabla_v f_0\cdot\left[\delta_{ij}+\int_{\tilde{t}^1}^0\left\{\nabla_x\left(E+J\right)\,\nabla_x X+\nabla_v J\,\nabla_x V\right\}d\tau\right]\Bigg\}\cdot\partial_{\tilde{t}}P\tilde{v}^1
\end{aligned}
$$

at $(0,\tilde{x}_0,\tilde{v}_0)$. Notice that $\tilde{t}^1=O\left(\varepsilon\right)$ to deduce

$$
\begin{aligned}
&f_t\left(\tilde{t},\tilde{x},\tilde{v}\right)\\
&=\tilde{t}_{\tilde{t}}^1\left[-\nabla_x f_0\cdot P\tilde{v}^1-\nabla_v f_0\cdot\left(E+J\right)\left(\tilde{t}^1,\tilde{x}^1,\tilde{v}^1\right)\right]\\
&\quad+\sum_{j=2}^3\partial_{\tilde{t}}\tilde{x}_j^1 f_{0x_j}-f_{0v_1}\partial_{\tilde{t}}\tilde{v}_1^1+\sum_{j=2}^3 f_{0v_j}\partial_{\tilde{t}}\tilde{v}_j^1+O\left(\varepsilon\right),
\end{aligned}
$$

evaluated at $(0,x_0,v_0)$. By using $(0,\tilde{x}_0,\tilde{v}_0)=\left(0,\tilde{x}^1,P\tilde{v}^1\right)+O\left(\varepsilon\right)=\left(0,\tilde{x}^1,Pv_0\right)+O\left(\varepsilon\right)$, we get, at $\left(0,\tilde{x}^1,Pv_0\right)$,

$$
\begin{aligned}
&f_t^k\left(\tilde{t},\tilde{x},\tilde{v}\right)\\
&=\tilde{t}_{\tilde{t}}^1\left[-Pv_0\cdot\nabla_x f_0\left(\tilde{x}^1,Pv_0\right)-\left(E+J\right)\left(0,\tilde{x}^1,v_0\right)\cdot\nabla_v f_0\right]\\
&\quad+\sum_{j=2}^3\partial_{\tilde{t}}\tilde{x}_j^1 f_{0x_j}-\partial_{\tilde{t}}\tilde{v}_1^1 f_{0v_1}+\sum_{j=2}^3\partial_{\tilde{t}}\tilde{v}_j^1 f_{0v_j}+O\left(\varepsilon^\eta\right),
\end{aligned}
$$

where we used $f_0\in C^{1,\eta}$. Hence, we have, by Lemma 2.10,

(4.8)

$$
\begin{aligned}
&f_t^k\left(\tilde{t},\tilde{x},\tilde{v}\right)\\
&=\tilde{t}_{\tilde{t}}^1 v_{01}f_{0x_1}\left(\tilde{x}^1,Pv_0\right)-\sum_{j=2}^3\tilde{t}_{\tilde{t}}^1 v_{0j}f_{0x_j}\left(\tilde{x}^1,v_0\right)+\tilde{t}_{\tilde{t}}^1\left(E_1+J_1\right)\left(0,\tilde{x}^1,v_0\right)f_{0v_1}\left(\tilde{x}^1,v_0\right)\\
&\quad-\sum_{j=2}^3\tilde{t}_{\tilde{t}}^1 E_j\left(0,\tilde{x}^1\right)f_{0v_j}\left(\tilde{x}^1,v_0\right)+\sum_{j=2}^3\tilde{t}_{\tilde{t}}^1 v_{0j}f_{0x_j}\left(\tilde{x}^1,v_0\right)\\
&\quad+\sum_{j=2}^3\left[-\tilde{v}_j+\left(\tilde{t}-\tilde{t}^1\right)E_j\left(\tilde{t},\tilde{x}\right)-\int_{\tilde{t}^1}^{\tilde{t}}\int_{\tilde{t}}^s\nabla_x E_j\cdot X_t d\tau ds\right]f_{0x_j}\left(\tilde{x}^1,v_0\right)\\
&\quad+\tilde{t}_{\tilde{t}}^1\left(E_1+J_1\right)\left(0,\tilde{x}^1,v_0\right)f_{0v_1}\left(\tilde{x}^1,v_0\right)\\
&\quad+\left[-\left(E_1+J_1\right)\left(\tilde{t},\tilde{x}\right)+\int_{\tilde{t}}^{\tilde{t}^1}\left\{\nabla_x\left(E_1+J_1\right)\cdot X_t+\nabla_v J\cdot V_t\right\}d\tau\right]f_{0v_1}\left(\tilde{x}^1,v_0\right)\\
&\quad+\sum_{j=2}^3\tilde{t}_{\tilde{t}}^1 E_j\left(0,\tilde{x}^1\right)f_{0v_j}\left(\tilde{x}^1,v_0\right)+\sum_{j=2}^3\left[-E_j\left(\tilde{t},\tilde{x}\right)+\int_{\tilde{t}}^{\tilde{t}^1}\nabla_x E_j\cdot X_t d\tau\right]f_{0v_j}\left(\tilde{x}^1,v_0\right)\\
&\quad+O\left(\varepsilon^\eta\right).
\end{aligned}
$$

Now we estimate the difference of (4.8) and (4.7) as

$$
\begin{aligned}
(4.9)\qquad & \left|f_t^k(t,x,v)-f_t^k\left(\tilde{t},\tilde{x},\tilde{v}\right)\right| \\
&= \left|\tilde{t}_t^1\left[v_{01}f_{0x_1}\left(\tilde{x}^1,-v_{01},v_{02},v_{03}\right)+v_{01}f_{0x_1}\left(\tilde{x}^1,v_{01},v_{02},v_{03}\right)\right.\right. \\
&\qquad \left.\left.+2\left(E_1+J_1\right)\left(0,\tilde{x}^1,v_0\right)f_{0v_1}\left(\tilde{x}^1,v_0\right)\right]\right|+O\left(\varepsilon^\eta\right).
\end{aligned}
$$

By our compatibility condition (4.2), the first term exactly vanishes. Similar computations hold for x- and v-derivatives. Therefore, in Case 2, we obtain from (4.9) that

$$
\left|\nabla_{(t,x,v)}f^k(t,x,v)-\nabla_{(t,x,v)}f^k\left(\tilde{t},\tilde{x},\tilde{v}\right)\right|\le C\left[\left|t-\tilde{t}\right|+|x-\tilde{x}|+|v-\tilde{v}|\right]^\eta.
$$

Case 3. The trajectories emanate from $\left(t^{l-1},x^{l-1},v^{l-1}\right)$ and $\left(\tilde{t}^{l-1},\tilde{x}^{l-1},\tilde{v}^{l-1}\right)$. In this case, we also have

$$
(4.10)\qquad \left|t^{l-1}-\tilde{t}^{l-1}\right|+\left|x^{l-1}-\tilde{x}^{l-1}\right|+\left|v^{l-1}-\tilde{v}^{l-1}\right|\le C\varepsilon,\quad C=C(\omega).
$$

Consider the back-time trajectories from $\left(t^{l-1},x^{l-1},v^{l-1}\right)$ and $\left(\tilde{t}^{l-1},\tilde{x}^{l-1},\tilde{v}^{l-1}\right)$. Assume that they are $\left(t^{i-1},x^{i-1},v^{i-1}\right)$ and $\left(\tilde{t}^{i-1},\tilde{x}^{i-1},\tilde{v}^{i-1}\right)$. Without loss of generality, we may assume that the first trajectory hits $\{t=0\}$ after l bounces. We have then

$$
\begin{aligned}
\nabla f^{k-j}\left(t^j,x^j,v^j\right)&=J_{j-1}^j\nabla f^{k-j-1}\left(t^{j-1},x^{j-1},v^{j-1}\right),\\
\nabla f^{k-j}\left(\tilde{t}^j,\tilde{x}^j,\tilde{v}^j\right)&=J_{j-1}^j\nabla f^{k-j-1}\left(\tilde{t}^{j-1},\tilde{x}^{j-1},\tilde{v}^{j-1}\right).
\end{aligned}
$$

Taking the difference, we get

$$
\begin{aligned}
&\nabla f^{k-j}\left(t^j,x^j,v^j\right)-\nabla f^{k-j}\left(\tilde{t}^j,\tilde{x}^j,\tilde{v}^j\right)\\
&=J_{j-1}^j\Delta\nabla f^{k-j-1}+\Delta J_{j-1}^j\nabla f^{k-j-1}\left(t^{j-1},x^{j-1},v^{j-1}\right).
\end{aligned}
$$

By induction on j,

$$
\begin{aligned}
&\left|\nabla f^k\left(t^{l-1},x^{l-1},v^{l-1}\right)-\nabla f^k\left(\tilde{t}^{l-1},\tilde{x}^{l-1},\tilde{v}^{l-1}\right)\right|\\
&\le\left|\prod_{j=1}^l J_{j-1}^j\left[\nabla f^{k-l-1}\left(t^1,x^1,v^1\right)-\nabla f^{k-l-1}\left(\tilde{t}^1,\tilde{x}^1,\tilde{v}^1\right)\right]\right|\\
&\quad+\sum_{j=1}^l\left|\Delta J_{j-1}^j\left[\nabla f^{k-j-1}\left(t^{l-j+1},x^{l-j+1},v^{l-j+1}\right)\right.\right.\\
&\qquad\qquad\left.\left.-\nabla f^{k-j-1}\left(\tilde{t}^{l-j+1},\tilde{x}^{l-j+1},\tilde{v}^{l-j+1}\right)\right]\right|.
\end{aligned}
$$

Notice that the number of bounces is uniformly bounded as in Lemma 4.2, and thus $\left|\prod_{j=1}^l J_{j-1}^j\right|\le C$. We thus get

$$
\begin{aligned}
(4.11)\qquad &\left|\nabla f^k\left(t^{l-1},x^{l-1},v^{l-1}\right)-\nabla f^k\left(\tilde{t}^{l-1},\tilde{x}^{l-1},\tilde{v}^{l-1}\right)\right|\\
&\le C\left|\nabla f^{k-l-1}\left(t^1,x^1,v^1\right)-\nabla f^{k-l-1}\left(\tilde{t}^1,\tilde{x}^1,\tilde{v}^1\right)\right|+C\sum_{j=1}^l\left|\Delta J_{j-1}^j\right|.
\end{aligned}
$$

Since $E\in C^{1,\eta}$ and by Lemma 2.10, (4.10), and applying the mean value theorem, we obtain

$$
(4.12)\qquad \left|\Delta J_{j-1}^j\right|\le C\varepsilon^\eta.
$$

Similarly, by using $f_0 \in C^{1,\eta}$, we get

$$\left|\nabla f^{k-l-1}\left(t^1, x^1, v^1\right) - \nabla f^{k-l-1}\left(\tilde{t}^1, \tilde{x}^1, \tilde{v}^1\right)\right| \le C\varepsilon^\eta. \tag{4.13}$$

Hence by plugging (4.12), (4.13) into (4.11), we finally obtain

$$\left|\nabla f^{k}\left(t^{l-1}, x^{l-1}, v^{l-1}\right) - \nabla f^{k}\left(\tilde{t}^{l-1}, \tilde{x}^{l-1}, \tilde{v}^{l-1}\right)\right| \le C\varepsilon^\eta.$$

Since $\left|t^1 - \tilde{t}^1\right| + \left|x^1 - \tilde{x}^1\right| + \left|v^1 - \tilde{v}^1\right| \le C\varepsilon$, the second cycle hits $t = 0$ after at most one bounce, as in Case 2. Applying Case 2 yields

$$\left|\nabla_{(t,x,v)} f^k(t,x,v) - \nabla_{(t,x,v)} f^k\left(\tilde{t}, \tilde{x}, \tilde{v}\right)\right| \le C\varepsilon^\eta.$$

This completes part (b) of the theorem.

For part (a), we construct f_0^n smooth such that $f_0^n \to f_0$ a.e., $\|f_0^n\|_{C^1}$ is uniformly bounded (depending on $\|f_0\|_{C^1}$), and f_0^n satisfies (4.3) and (4.4). By the result of part (b), there is a unique solution f^n of (4.1) with data f_0^n such that $\|f^n\|_{C^1} \le C\|f_0^n\|_{C^1} \le C$. Hence part (a) follows by letting $n \to \infty$. For part (c), we refer the reader to [8]. ∎

5. Regularity for the Vlasov–Poisson system with the absorbing boundary condition. In this section, we consider the fully nonlinear Vlasov–Poisson system with the absorbing boundary condition for the Vlasov and the Dirichlet boundary condition for the Poisson equation:

$$\begin{gathered} f_t + v\cdot\nabla_x f + \nabla\varphi\cdot\nabla_v f = 0, \\ f|_{t=0} = f_0, \quad f|_{\gamma^+} = g, \\ \Delta\varphi = \rho = 4\pi\int f dv, \\ \varphi|_{\partial\Omega} = 0. \end{gathered}$$

THEOREM 5.1. *Let $k \ge 1$, $3 < p \le \infty$. Let $f_0 \in W^{k,p}(\Pi_0)$ and $g \in W^{k,p}(\gamma^+)$ have compact support and $f_0 \ge 0$, $g \ge 0$. Assume the compatibility condition* (3.10) *holds for $x \in \partial\Omega$ and v with $n_x \cdot v < 0$ and for $|\alpha| \le k-1$. Moreover, assume the vanishing condition:*

$$\begin{gathered} g(t,x,v) \equiv 0 \ \text{ on } \gamma^0, \\ \left|\partial^\alpha g(t,x,v)\right| \le C\left|n_x\cdot v\right|^{|\alpha|} \ \text{ on } \gamma^+, \quad |\alpha| = k, \end{gathered}$$

where α is a multi-index. Then there exists a unique solution $f \in W^{k,p}(\Pi)$ and $\varphi \in W^{k+2,p}$, where f has compact support in v.

We shall construct approximate solutions by establishing an iterating system. Let f^0 be a suitable smooth extension of f_0 to Π and satisfy the corresponding compatibility condition (3.10). Let the iterating sequence be

$$\begin{gathered} \partial_t f^{n+1} + v\cdot\nabla_x f^{n+1} + \nabla\varphi^n\cdot\nabla_v f^{n+1} = 0, \\ f^{n+1}|_{t=0} = f_0, \quad f^{n+1}|_{\gamma^+} = g, \end{gathered} \tag{5.1}$$

$$\Delta\varphi^n = \rho^n = 4\pi\int f^n dv, \quad \varphi^n|_{\partial\Omega} = 0. \tag{5.2}$$

Since $f_0 \geq 0$, $g \geq 0$, and $\varphi^n|_{\partial\Omega} = 0$, by the Vlasov equation (5.1), we have

$$\Delta\varphi^n = 4\pi\rho^n = 4\pi \int f^n dv \geq 0.$$

By the strong maximal principle and since f_0 is not identically zero,

$$\varphi^n < 0 \text{ on } \Omega.$$

We then apply the Hopf boundary principle to get

$$E^n(t,x) \cdot n_x = \frac{\partial \varphi^n}{\partial n}(t,x) \geq \delta_n > 0$$

on $\partial\Omega \cap \{\text{support of } f^n\}$. From Theorem 3.6, f^{n+1} is well defined in $W^{k,p}$ for every fixed n.

We shall use the idea which was in [18] for the Cauchy problem without boundary. The key step is to represent the macrocharge density ρ^n in the presence of the complex particle path, along the straight-line trajectory

$$\frac{dX}{dt} = V, \quad \frac{dV}{dt} = 0. \tag{5.3}$$

We consider the back-time trajectory of $dX/dt = V$, $dV/dt = 0$ from a generic point (t,x,v). We denote by $B(t,x,v) = (t_0, x_0, v)$ the possible boundary point when the trajectory hits $\partial\Omega$. We first note that for $v \neq 0$, there exists a unique $x_0 = x_0(x,v) \in \partial\Omega$ along the straight-line trajectory from (t,x,v) since Ω is convex. Let t_0 be the time when the trajectory from (t,x,v) hits the boundary. Then $x_0 = x + v(t_0 - t)$ and $t_0 - t = [(x_0 - x) \cdot v] / |v|^2$. We define

$$a(x,v) = -[(x_0 - x) \cdot v] / |v|^2$$

to see that the function $a(x,v)$ is locally differentiable as follows: Let ξ be the function which defines the boundary (2.8). Then we have

$$0 = \xi(x_0) = \xi(x + v(t_0 - t)).$$

Set $s = t_0 - t$ to get that $0 = \xi(x + sv) = \xi(s; x, v)$ and $\partial\xi/\partial s = \nabla\xi(x_0) \cdot v = n_{x_0} \cdot v < 0$ since Ω is convex. By the implicit function theory, $s = t_0 - t = -a(x,v)$ is a locally differentiable function of x and v. Before giving the representation formula for the macrocharge density, we present two preliminary lemmas.

LEMMA 5.2. *$a \cdot \nabla_x a + \nabla_v a = 0$ for $v \neq 0$, $x \in \Omega$.*

Proof. Let ξ be the function which defines the boundary (2.8). Since $x_0 = x - va(x,v)$, we have

$$0 = \xi(x_0) = \xi(x - va(x,v)). \tag{5.4}$$

Differentiate (5.4) with respect to x to get

$$\nabla\xi - (\nabla\xi \cdot v)\nabla_x a = 0. \tag{5.5}$$

We now differentiate (5.4) with respect to v to get

$$-a\nabla\xi - (\nabla\xi \cdot v)\nabla_v a = 0. \tag{5.6}$$

By multiplying (5.5) with a and adding it to (5.6), we get

$$(\nabla\xi(x_0)\cdot v)\left[a\nabla_x a+\nabla_v a\right]=0.$$

Since $\nabla\xi(x_0)\cdot v\neq 0$ from the convexity of Ω, the lemma follows. □

First we note that for fixed x and t, $a(x,v)=t$ defines a smooth surface except for the origin and $a(x,v)<t$ defines the three-dimensional unbounded set outside of the surface while $a(x,v)>t$ defines the object inside of the surface. We shall use the spherical coordinates (r,ϕ,θ) instead of the usual rectangular coordinates (v_1,v_2,v_3) for the moment. We denote by $\nu=\left(\nu^r,\nu^\phi,\nu^\theta\right)$ the outward normal in the spherical coordinates to the surface defined by $a(x,r,\phi,\theta)=t$ for fixed x. We also denote by $dS(r,\phi,\theta)$ the surface infinitesimal increment for the surface $a(x,r,\phi,\theta)=t$.

LEMMA 5.3. *Let $\Gamma(x,v)$ be the C^1-vector-valued function in x,v. Then for fixed t, we have*

$$\begin{aligned}\operatorname{div}_x\int_{a(x,v)\le t}\Gamma(x,v)\,dv&=\int_{a(x,v)\le t}\operatorname{div}_x\Gamma(x,v)\,dv\\&\quad-\int_{a(x,r,\phi,\theta)=t}\Gamma(x,r,\phi,\theta)\,r^2\sin\phi\cdot\nabla_x a\frac{\nu^r}{a_r}dS(r,\phi,\theta),\end{aligned}$$

$$\begin{aligned}\operatorname{div}_x\int_{a(x,v)\ge t}\Gamma(x,v)\,dv&=\int_{a(x,v)\ge t}\operatorname{div}_x\Gamma(x,v)\,dv\\&\quad+\int_{a(x,r,\phi,\theta)=t}\Gamma(x,r,\phi,\theta)\,r^2\sin\phi\cdot\nabla_x a\frac{\nu^r}{a_r}dS(r,\phi,\theta).\end{aligned}$$

Proof. Note, by multiplying (5.6) with v, that $v\cdot\nabla_v a\neq 0$ if $x\notin\partial\Omega$ and $v\neq 0$ since $a=t_0-t\neq 0$, $\nabla\xi\neq 0$, and $\nabla\xi\cdot v\neq 0$. By using the spherical coordinates, we get $r\partial_r a\neq 0$ and so $\partial_r a\neq 0$. We first consider a C^1-scalar function $h(r,\phi,\theta)$ without x-variables. We change variables as follows:

$$(r,\phi,\theta)\longmapsto(\eta_1,\eta_2,\eta_3),$$

where for fixed x,

$$\eta_1=a(x,r,\phi,\theta),\quad \eta_2=\phi,\quad \eta_3=\theta. \tag{5.7}$$

We find the Jacobian of (η_1,η_2,η_3) with respect to (r,ϕ,θ):

$$J\left(\frac{\eta_1,\eta_2,\eta_3}{r\ ,\phi\ ,\theta}\right)=\det\begin{bmatrix}\frac{\partial a}{\partial r}&\frac{\partial a}{\partial\phi}&\frac{\partial a}{\partial\theta}\\0&1&0\\0&0&1\end{bmatrix}=\frac{\partial a}{\partial r}.$$

For fixed η_1, we differentiate with respect to x the equation $\eta_1=a(x,r,\phi,\theta)$ to get

$$0=\nabla_x\eta_1=\nabla_x a+a_r\nabla_x r \tag{5.8}$$

or

$$\nabla_x r = -\frac{\nabla_x a}{a_r}.$$

We now consider, using the change of variables (5.7) back and forth and by changing the orders of the integration,

$$\begin{aligned}
&\frac{\partial}{\partial x_i}\int_{a(x,r,\phi,\theta)\le t} h(r,\phi,\theta)\,dr d\phi d\theta \\
&= \frac{\partial}{\partial x_i}\int_{\eta_1\le t} h(r(\eta_1,\eta_2,\eta_3),\eta_2,\eta_3)\frac{1}{a_r}d\eta_1 d\eta_2 d\eta_3 \\
&= \int_{\eta_1\le t}\frac{\partial h}{\partial r}\frac{\partial r}{\partial x_i}\frac{1}{a_r}d\eta_1 d\eta_2 d\eta_3 - \int_{\eta_1\le t} h\frac{a_{rx_i}+a_{rr}r_{x_i}}{a_r^2}d\eta_1 d\eta_2 d\eta_3 \\
&= \int_{-\infty}^{t}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left[-\frac{h_r a_{x_i}}{a_r^2} - \frac{h a_{rx_i}}{a_r^2} + \frac{h a_{x_i} a_{rr}}{a_r^3}\right] d\eta_2 d\eta_3 d\eta_1 \qquad (5.9)\\
&= \int_{\eta_1\le t}\left[-\frac{(ha_{x_i})_r}{a_r^2} + \frac{h a_{x_i} a_{rr}}{a_r^3}\right] d\eta \\
&= \int_{a(x,r,\phi,\theta)\le t}\left[-\frac{(ha_{x_i})_r}{a_r^2} + \frac{h a_{x_i} a_{rr}}{a_r^3}\right] a_r dr d\phi d\theta \\
&= \int_{a(x,r,\phi,\theta)\le t}\left[-\frac{(ha_{x_i})_r}{a_r} + \frac{h a_{x_i} a_{rr}}{a_r^2}\right] dr d\phi d\theta \\
&= -\int_{a(x,r,\phi,\theta)\le t}\left(\frac{h a_{x_i}}{a_r}\right)_r dr d\phi d\theta \\
&= -\int_{a(x,r,\phi,\theta)=t}\frac{h a_{x_i}}{a_r}\nu^r dS(r,\phi,\theta), \qquad (5.10)
\end{aligned}$$

where we have used (5.8) in (5.9) and the Gauss theorem in (5.10). We now consider a C^1-scalar function $h(v_1,v_2,v_3)$. We then take the x_i-derivative ($i=1,2,3$) of the integration of h with respect to v over $a(x,v)\le t$ for fixed t. By changing variables from (v_1,v_2,v_3) to (r,ϕ,θ), we have, by (5.10),

$$\begin{aligned}
&\frac{\partial}{\partial x_i}\int_{a(x,v)\le t} h(v_1,v_2,v_3)\,dv_1 dv_2 dv_3 \qquad (5.11)\\
&= \frac{\partial}{\partial x_i}\int_{a(s,r,\phi,\theta)\le t} h(r,\phi,\theta)\,r^2\sin\phi\, dr d\phi d\theta \\
&= -\int_{a(x,r,\phi,\theta)=t}\frac{h r^2\sin\phi a_{x_i}}{a_r}\nu^r dS(r,\phi,\theta).
\end{aligned}$$

Now we consider the integration in our lemma

$$\begin{aligned}
&\mathrm{div}_x\int_{a(x,v)\le t}\Gamma(x,v)\,dv \\
&= \sum_{i=1}^{3}\frac{\partial}{\partial x_i}\int_{a(x,v)\le t}\Gamma^i(x,v)\,dv \\
&= \sum_{i=1}^{3}\left[\int_{a(x,v)\le t}\partial_{x_i}\Gamma^i(x,v)\,dv - \int_{a(x,r,\phi,\theta)=t}\Gamma^i r^2\sin\phi\frac{a_{x_i}}{a_r}\nu^r dS\right],
\end{aligned}$$

where we have used the product rule of the differentiation and (5.11). This leads to the conclusion of the first part of the lemma. Similarly, we get the second part, and the lemma thus follows. ∎

We present the representation formula for the macrocharge density in the following lemma.

LEMMA 5.4 (charge density). *Let f^{n+1} and φ^n be defined in* (5.1) *and* (5.2). *Let*

$$B(t,x,v) = (t - a(x,v), x - a(x,v)\, v, v) \in \gamma.$$

Then

(5.12)

$$\begin{aligned}\rho^{n+1}(t,x) = &\int_{a(x,v)\geq t} f_0(x - tv, v)\, dv + \int_{a(x,v)\leq t} g \circ B(t,x,v)\, dv \\ &- \operatorname{div}_x \int_{a(x,v)\geq t} \int_0^t (t-\tau) \left(\nabla_x \varphi^n f^{n+1}\right)(\tau, x - (t-\tau)\, v, v)\, d\tau dv \\ &- \operatorname{div}_x \int_{a(x,v)\leq t} \int_{t-a(x,v)}^t (t-\tau) \left(\nabla_x \varphi^n f^{n+1}\right)(\tau, x - (t-\tau)\, v, v)\, d\tau dv.\end{aligned}$$

Proof. We fix x and t and consider the back-time straight-line trajectory from a generic point (t,x,v). If $t \leq a(x,v)$, then the back-time trajectory of (5.1) hits $\{t = 0\}$ directly. From the transport equation (5.1), we have

$$\begin{aligned}f^{n+1}(t,x,v) &= f_0(x - tv, v) + \int_0^t \frac{d}{d\tau} f^{n+1}(\tau, x - (t-\tau)\, v, v)\, d\tau \\ &= f_0(x - tv, v) + \int_0^t \big[\partial_t f^{n+1}(\tau, x - (t-\tau)\, v, v) \\ &\qquad + v \cdot \nabla_x f^{n+1}(\tau, x - (t-\tau)\, v, v)\big]\, d\tau \\ &= f_0(x - tv, v) - \int_0^t \left[\operatorname{div}_v \left(\nabla_x \varphi^n f^{n+1}\right)\right](\tau, x - (t-\tau)\, v, v)\, d\tau \\ &= f_0(x - tv, v) - \int_0^t \operatorname{div}_v \left[\left(\nabla_x \varphi^n f^{n+1}\right)(\tau, x - (t-\tau)\, v, v)\right] d\tau \\ &\quad - \int_0^t (t-\tau) \operatorname{div}_x \left(\nabla_x \varphi^n f^{n+1}\right)(\tau, x - (t-\tau)\, v, v)\, d\tau \\ &= f_0(x - tv, v) - \operatorname{div}_v \int_0^t \left(\nabla_x \varphi^n f^{n+1}\right)(\tau, x - (t-\tau)\, v, v)\, d\tau \\ &\quad - \operatorname{div}_x \int_0^t (t-\tau) \left(\nabla_x \varphi^n f^{n+1}\right)(\tau, x - (t-\tau)\, v, v)\, d\tau.\end{aligned} \tag{5.13}$$

On the other hand, if $t \geq a(x,v)$, then the backward trajectory hits the boundary

$\partial\Omega$. Hence we have

(5.14)

$$\begin{aligned}
f^{n+1}(t,x,v) &= g\circ B(t,x,v)+\int_{t-a(x,v)}^{t}\frac{d}{d\tau}f^{n+1}(\tau,x-(t-\tau)v,v)\,d\tau\\
&= g\circ B(t,x,v)+\int_{t-a(x,v)}^{t}\big[\partial_t f^{n+1}(\tau,x-(t-\tau)v,v)\\
&\qquad\qquad +v\cdot\nabla_x f^{n+1}(\tau,x-(t-\tau)v,v)\big]\,d\tau\\
&= g\circ B-\int_{t-a(x,v)}^{t}\left[\operatorname{div}_v\left(\nabla_x\varphi^n f^{n+1}\right)\right](\tau,x-(t-\tau)v,v)\,d\tau\\
&= g\circ B-\int_{t-a(x,v)}^{t}\operatorname{div}_v\left[\left(\nabla_x\varphi^n f^{n+1}\right)(\tau,x-(t-\tau)v,v)\right]d\tau\\
&\quad -\int_{t-a(x,v)}^{t}(t-\tau)\operatorname{div}_x\left(\nabla_x\varphi^n f^{n+1}\right)(\tau,x-(t-\tau)v,v)\,d\tau\\
&= g\circ B-\operatorname{div}_v\int_{t-a(x,v)}^{t}\left(\nabla_x\varphi^n f^{n+1}\right)(\tau,x-(t-\tau)v,v)\,d\tau\\
&\quad -\operatorname{div}_x\int_{t-a(x,v)}^{t}(t-\tau)\left(\nabla_x\varphi^n f^{n+1}\right)(\tau,x-(t-\tau)v,v)\,d\tau\\
&\quad +\left(\nabla_x\varphi^n f^{n+1}\right)(t-a(x,v),x-a(x,v)v,v)\cdot\nabla_v a\\
&\quad +\left(\nabla_x\varphi^n f^{n+1}\right)(t-a(x,v),x-a(x,v)v,v)\cdot a\nabla_x a\\
&= g\circ B-\operatorname{div}_v\int_{t-a(x,v)}^{t}\left(\nabla_x\varphi^n f^{n+1}\right)(\tau,x-(t-\tau)v,v)\,d\tau\\
&\quad -\operatorname{div}_x\int_{t-a(x,v)}^{t}(t-\tau)\left(\nabla_x\varphi^n f^{n+1}\right)(\tau,x-(t-\tau)v,v)\,d\tau,
\end{aligned}$$

since we apply Lemma 5.2 to get, for $v\neq 0$, $x\notin\partial\Omega$,

$$\begin{aligned}
&\left(\nabla_x\varphi^n f^{n+1}\right)(t-a(x,v),x-a(x,v)v,v)\cdot\nabla_v a\\
&\quad+\left(\nabla_x\varphi^n f^{n+1}\right)(t-a(x,v),x-a(x,v)v,v)\cdot a\nabla_x a\\
&=\left(\nabla_x\varphi^n f^{n+1}\right)(t_0,x_0,v)\cdot[\nabla_v a+a\nabla_x a]=0.
\end{aligned}$$

For fixed t and x, we now integrate v over $\mathbb{R}^3$. By dividing v by the region $\{a(x,v)\geq t\}$ and the region $\{a(x,v)\leq t\}$, we get

$$\begin{aligned}
\rho^{n+1}(t,x) &= \int_{a(x,v)\geq t}f^{n+1}(t,x,v)\,dv+\int_{a(x,v)\leq t}f^{n+1}(t,x,v)\,dv\\
&= I_1+I_2.
\end{aligned}$$

For I_1, we use (5.13), Lemma 5.3, and the Gauss theorem to get

$$\begin{aligned}
&\int_{a(x,v)\geq t} f^{n+1}(t,x,v)\,dv \\
&= \int_{a(x,v)\geq t} f_0(x-tv,v)\,dv \\
&\quad + \int_{a(x,v)=t}\int_0^t (\nabla_x\varphi^n\cdot n_v) f^{n+1}(\tau,x-(t-\tau)v,v)\,d\tau dS(v) \\
&\quad - \int_{a(x,v)\geq t} \operatorname{div}_x \int_0^t (t-\tau)\left(\nabla_x\varphi^n f^{n+1}\right)(\tau,x-(t-\tau)v,v)\,d\tau dv \\
&= \int_{a(x,v)\geq t} f_0(x-tv,v)\,dv \\
&\quad + \int_{a(x,v)=t}\int_0^t (\nabla_x\varphi^n\cdot n_v) f^{n+1}(\tau,x-(t-\tau)v,v)\,d\tau dS(v) \\
&\quad - \operatorname{div}_x \int_{a(x,v)\geq t}\int_0^t (t-\tau)\left(\nabla_x\varphi^n f^{n+1}\right)(\tau,x-(t-\tau)v,v)\,d\tau dv \\
&\quad + \int_{a(x,r,\phi,\theta)=t}\int_0^t (t-\tau)\left(\nabla_x\varphi^n f^{n+1}\right) r^2\sin\phi\cdot\nabla_x a\frac{\nu^r}{a_r} dS(r,\phi,\theta),
\end{aligned}$$

where n_v is the outward normal to the surface $\{a(x,v)=t\}$ which contains inside the region $\{a(x,v)\leq t\}$. For I_2, using (5.15), Lemma 5.3, and the Gauss theorem, we have

$$\begin{aligned}
&\int_{a(x,v)\leq t} f^{n+1}(t,x,v)\,dv \\
&= \int_{a(x,v)\leq t} g\circ B(t,x,v)\,dv \\
&\quad - \int_{a(x,v)=t}\int_0^t (\nabla_x\varphi^n\cdot n_v) f^{n+1}(\tau,x-(t-\tau)v,v)\,d\tau dS(v) \\
&\quad - \int_{a(x,v)\leq t} \operatorname{div}_x \int_0^t (t-\tau)\left(\nabla_x\varphi^n f^{n+1}\right)(\tau,x-(t-\tau)v,v)\,d\tau dv \\
&= \int_{a(x,v)\leq t} g\circ B(t,x,v)\,dv \\
&\quad - \int_{a(x,v)=t}\int_0^t (\nabla_x\varphi^n\cdot n_v) f^{n+1}(\tau,x-(t-\tau)v,v)\,d\tau dS(v) \\
&\quad - \operatorname{div}_x \int_{a(x,v)\leq t}\int_0^t (t-\tau)\left(\nabla_x\varphi^n f^{n+1}\right)(\tau,x-(t-\tau)v,v)\,d\tau dv \\
&\quad - \int_{a(x,r,\phi,\theta)=t}\int_0^t (t-\tau)\left(\nabla_x\varphi^n f^{n+1}\right) r^2\sin\phi\cdot\nabla_x a\frac{\nu^r}{a_r} dS(r,\phi,\theta).
\end{aligned}$$

Therefore, by all the cancellations out of I_1 and I_2, we obtain our lemma. □

In the following, we shall give some estimates on the sequences of f^n and φ^n, uniformly in n.

LEMMA 5.5. *We have* $\|f^n t\|_{L^p} \le C$ *for* $1 \le p \le \infty$ *and*

$$\begin{aligned}
&\left\|\nabla_x \varphi^{n+1}\right\|_{L^p(\Omega)}(t) \\
&\le \left\| \int_{\mathbb{R}^3} \int_0^t \mathbf{1}_{\{a(x,v)\ge t\}}(v)(t-\tau)\left|\left(\nabla\varphi^n f^{n+1}\right)(\tau, x-(t-\tau)v, v)\right| d\tau dv \right\|_{L^p(\Omega)} \\
&\quad + \left\| \int_{\mathbb{R}^3} \int_0^t \mathbf{1}_{\{a\le t\}}(v)\mathbf{1}_{(t-a,t)}(\tau)(t-\tau)\left|\left(\nabla\varphi^n f^{n+1}\right)(\tau, x-(t-\tau)v, v)\right| d\tau dv \right\|_{L^p(\Omega)} \\
&\quad + C,
\end{aligned}$$

where $1 \le p < \infty$*, and* C *is a constant independent of* n*, depending only on the data* f_0 *and* g*.*

Proof. The first estimate on f^n easily follows from standard estimates for the transport equation (5.1). For the second estimate, we employ the elliptic equation $\Delta\varphi^{n+1} = \rho^{n+1}$ and the representation formula for φ^{n+1} (5.12). We note that since Ω is bounded and the v-support of f_0 is compact, we get

$$\frac{1}{|x|^2} * \left[\int f_0(x-tv, v)\mathbf{1}_{\{a(x,v)\ge t\}}(v)\, dv\right] \le C. \tag{5.15}$$

Similarly, we have

$$\frac{1}{|x|^2} * \left[\int f(B(t,x,v))\mathbf{1}_{\{a(x,v)\le t\}}(v)\, dv\right] \le C. \tag{5.16}$$

By standard elliptic estimates for a bounded domain [15] and by (5.15), (5.16), our lemma follows. □

Now we give the lemma which is a major step for the global bound on the velocity.

LEMMA 5.6 (high moments bound). *Let* f^n *and* φ^n *be defined in* (5.1), (5.2)*. Then for a fixed* $m > 3$*, we have*

$$\sup_n \int_{\Omega\times\mathbb{R}^3} |v|^m f^n(t,x,v)\, dxdv < \infty$$

for all $0 \le t \le T$*. In particular, there is a uniform bound (independent of* n*) for the support of* f^n*.*

Proof. We shall closely follow the method given in [18]. We first define

$$M_m(f^n)(s) = \sup_{0\le t\le s} \int_{\Omega\times\mathbb{R}^3} |v|^m f^n(t,x,v)\, dxdv.$$

Then note that

$$\begin{aligned}
\frac{d}{dt}\int_{\Omega\times\mathbb{R}^3} |v|^m f^n(t,x,v)\, dxdv &= -\int \nabla_x\cdot\left(|v|^m v f^n\right) dxdv \\
&\quad - \int |v|^m \nabla_v\cdot\left(\nabla_x\varphi^{n-1} f^n\right) dxdv.
\end{aligned} \tag{5.17}$$

By the Gauss theorem, the first integral on the RHS of (5.17) becomes

$$\begin{aligned}
-\int_{\partial\Omega\times\mathbb{R}^3} |v|^m f^n v\cdot n_x dS(x)\, dv &= -\int_{v\cdot n_x\le 0} - \int_{v\cdot n_x\ge 0} \\
&\le -\int_{\gamma_t^+} |v|^m g(t,x,v)\, v\cdot n_x dSdv \\
&\le C
\end{aligned}$$

since g has compact support. For the second integral on the RHS of (5.17), we first use the integration by parts to get

$$\begin{aligned}
&-\int \nabla_v \cdot \left(|v|^m \nabla_x \varphi^{n-1} f^n\right) dxdv + m\int |v|^m f^n \nabla_x \varphi^{n-1} \cdot \nabla_v \left(|v|\right) dxdv \\
&= m\int |v|^m f^n \nabla_x \varphi^{n-1} \cdot \nabla_v \left(|v|\right) dxdv,
\end{aligned}$$

where the first integral vanishes. Now, by using the interpolation method, we get

$$\begin{aligned}
&\left|-\int |v|^m \nabla_v \cdot \left(\nabla_x \varphi^{n-1} f^n\right) dxdv\right| \\
&\le \left|m\int |v|^m f^n \nabla_x \varphi^{n-1} \cdot \nabla_v \left(|v|\right) dxdv\right| \\
&\le C\int_\Omega \left|\nabla_x \varphi^{n-1}(t,x)\right| \left(\int_{|v|\ge R} |v|^{m-1} f^n dv + \int_{|v|\le R} |v|^{m-1} f^n dv\right) dx \\
&\le C\int_\Omega \left|\nabla_x \varphi^{n-1}(t,x)\right| \left(R^{-1}\int |v|^m f^n dv + CR^{m+2}\right) dx \\
&\le C\int_\Omega \left|\nabla_x \varphi^{n-1}(t,x)\right| \left(\int |v|^m f^n dv\right)^{\frac{m+2}{m+3}} dx \\
&\le C\left\|\nabla_x \varphi^{n-1}(t)\right\|_{L^{m+3}(\Omega)} M_m \left(f^n\right)^{\frac{m+2}{m+3}},
\end{aligned}$$

where $R = \left(\int |v|^m f^n dv\right)^{1/(m+3)}$ was used for the optimal inequality, and we used the Hölder inequality. Thus we get

(5.18)
$$\frac{d}{ds} M_m\left(f^n\right)(s) \le C_1 + C\sup_{0\le t\le s} \left\|\nabla_x \varphi^{n-1}(t)\right\|_{L^{m+3}(\Omega)} M_m\left(f^n\right)(s)^{(m+2)/(m+3)},$$

where C_1 depends on f_0 and g. We shall estimate $\left\|\nabla_x \varphi^{n-1}(t)\right\|_{L^{m+3}(\Omega)}$. From Lemma 5.4, we have

(5.19)
$$\begin{aligned}
&\left\|\nabla_x \varphi^{n-1}\right\|_{m+3}(t) \\
&\le \left\|\int_{\mathbb{R}^3}\int_0^t \mathbf{1}_{\{a(x,v)\ge t\}}(v)(t-\tau)\left|\left(\nabla\varphi^{n-2} f^{n-1}\right)(\tau, x-(t-\tau)v, v)\right| d\tau dv\right\|_{m+3} \\
&\quad + \left\|\int_{\mathbb{R}^3}\int_0^t \mathbf{1}_{\{a(x,v)\le t\}}(v)\mathbf{1}_{(t-a,t)}(\tau)(t-\tau) \right. \\
&\qquad\qquad \left. \times \left|\left(\nabla\varphi^{n-2} f^{n-1}\right)(\tau, x-(t-\tau)v, v)\right| d\tau dv\right\|_{m+3} \\
&\quad + C.
\end{aligned}$$

We shall estimate the two terms in the RHS by the sum of the long-time integral and the short-time integral:

$$\left\|\int_{t_0}^t \cdots\right\|_{m+3} + \left\|\int_0^{t_0} \cdots\right\|_{m+3}, \tag{5.20}$$

where t_0 is some small time to be chosen later. We first do the long-time estimate which is the first term in (5.20). Choose $r^{'}=3$, $r=3/2$. By the Hölder inequality,

$$\begin{aligned}\left\|\int_{t_0}^{t}\cdots\right\|_{m+3} &\le C\int_{t_0}^{t}\frac{1}{\tau}d\tau\left[\sup_{\tau\in(0,T)}\left\|E\left(\tau\right)\right\|_{3/2}M_m\left(f^n\right)^{1/(m+3)}\right]\left(\tau\right)\\ &\le C\log t_0\sup_{\tau\in(0,T)}\left\|E\left(\tau\right)\right\|_{3/2}M_m\left(f^n\right)^{1/(m+3)}\left(t\right)\\ &\le C\log t_0 M_m\left(f^n\right)^{1/(m+3)}\left(t\right).\end{aligned}\tag{5.21}$$

For the short-time integral over $(0,t_0)$, we use the standard interpolation estimate for ρ^{n-2},

$$\begin{aligned}\rho^{n-2}=\int_{\mathbb{R}^3}f^{n-2}dv &= \int_{|v|\le R}+\int_{|v|\ge R}\\ &\le CR^3\left\|f^{n-2}\right\|_{\infty}+CR^{-m}\int|v|^m f^{n-2}dv\\ &\le C\left(\int|v|^m f^{n-2}dv\right)^{3/(m+3)}\quad\text{for any } R,\end{aligned}\tag{5.22}$$

where we have chosen R so as to optimize the last inequality. Since $\left|\nabla_x\varphi^{n-2}\right|\le\frac{1}{|x|^2}*\rho^{n-2}\left(x\right)$, we apply the Hardy–Littlewood–Sobolev inequality to get

$$\begin{aligned}\sup_{\tau\in(0,t)}\left\|\nabla_x\varphi^{n-2}\right\|_r &\le C\sup_{\tau\in(0,t)}\left\|\rho^{n-2}\right\|_{\frac{3r}{3+r}}\\ &\le C\sup_{\tau\in(0,t)}\left\|\rho^{n-2}\right\|_{\frac{m+3}{3}}\\ &\le CM_m^{\alpha}\left(f^{n-2}\right)\left(t\right),\end{aligned}\tag{5.23}$$

where $\alpha=3/\left(m+3\right)<1$, and we have used that $\frac{3r}{3+r}<\frac{m+3}{3}$ and Ω is bounded. By the same argument as in (5.22), we have

$$\begin{aligned}&\sup_{\tau\in(0,t)}\left\|\left[\int_{\mathbb{R}^3}\mathbf{1}_{\{a(x,v)\ge t\}}f^{n-1}\left(t-\tau,x-\tau v,v\right)dv\right]^{1/r^{'}}\right\|_{m+3}\\ &=\sup_{\tau\in(0,t)}\left\|\int_{\mathbb{R}^3}\mathbf{1}_{\{a(x,v)\ge t\}}f^{n-1}\left(t-\tau,x-\tau v,v\right)dv\right\|_{\frac{m+3}{r^{'}}}^{\frac{1}{r^{'}}}\\ &\le CM_k^{1/(m+3)}\left(f^{n-1}\right)\left(t\right)\\ &\le C+C\sup_{\tau\in(0,t)}\left\|\nabla_x\varphi^{n-1}\right\|_{k+3}^{(k+3)/(m+3)}\\ &\le C+C\sup_{\tau\in(0,t)}\left\|\nabla_x\varphi^{n-1}\right\|_{q}^{(k+3)/(m+3)}\\ &\le C+C\sup_{\tau\in(0,t)}\left\|\rho^{n-1}\left(\tau\right)\right\|_{(m+3)/3}^{(k+3)/(m+3)}\\ &\le C+CM_m^{\beta}\left(f^{n-1}\right)\left(t\right),\end{aligned}\tag{5.24}$$

where $\frac{m+3}{r^{'}}=\frac{k+3}{3}$, $\beta=\frac{3(k+3)}{(m+3)^2}$, we have chosen q,k such that $\frac{1}{q}=\frac{3}{m+3}-\frac{1}{3}$, $m+3\lesssim k+3\le q$, and we have used $M_k\left(t\right)\le C\left\{M_k\left(0\right)+\sup_{\tau\in(0,t)}\left\|\nabla_x\varphi^{n-1}\left(\tau\right)\right\|_{k+3}^{k+3}\right\}$.

Therefore, by (5.23), (5.24),

(5.25)

$$\begin{aligned}
&\left\|\int_0^{t_0}\cdots\right\|_{m+3}\\
&\le Ct_0^{2-3/r}\sup_{\tau\in(0,t)}\left\|\nabla_x\varphi^{n-2}\right\|_r\\
&\qquad\times\sup_{\tau\in(0,t)}\left\|\left[\int_{\mathbb{R}^3}\mathbf{1}_{\{a(x,v)\ge t\}}f^{n-1}\left(t-\tau,x-\tau v,v\right)dv\right]^{1/r'}\right\|_{m+3}\\
&\quad+Ct_0^{2-3/r}\sup_{\tau\in(0,t)}\left\|\nabla_x\varphi^{n-2}\right\|_r\\
&\qquad\times\sup_{\tau\in(0,t)}\left\|\left[\int_{\mathbb{R}^3}\mathbf{1}_{\{a(x,v)\le t\}}\mathbf{1}_{(t-a(x,v),t)}f^{n-1}\left(t-\tau,x-\tau v,v\right)dv\right]^{1/r'}\right\|_{m+3}\\
&\le Ct_0^{2-3/r}M_m^{\alpha}\left(f^{n-1}\right)(t)\left[C+CM_m^{\beta}\left(f^{n-1}\right)(t)\right]\\
&\le Ct_0^{2-3/r}M_m^{\alpha+\beta}\left(f^{n-1}\right)(t).
\end{aligned}$$

Hence we have from (5.19), (5.21), and (5.25)

$$\sup_{0\le t\le s}\left\|\nabla_x\varphi^{n-1}\right\|_{m+3}(t)\le C+C\log t_0M_m\left(f^n\right)^{1/(m+3)}(t)+Ct_0^{2-3/r}M_m^{\alpha+\beta}\left(f^{n-1}\right)(t).$$

Setting $\bar{M}_{n,m}=\max_{1\le i\le n}M_m\left(f^i\right)$, we get, by (5.18),

$$\frac{d}{dt}\bar{M}_{n,m}(t)\le C+C\log t_0\bar{M}_{n,m}(t)+Ct_0^{2-3/r}\bar{M}_{n,m}^{\alpha+\beta+(m+2)/(m+3)}(t).$$

Choosing $t_0^{2-3/r}=\bar{M}_{n,m}^{1-\alpha-\beta-(m+2)/(m+3)}$, we deduce our lemma. By choosing $m>6$, we have $\|\nabla_x\varphi^n\|_{L^\infty}$ uniformly bounded and get a uniform upper bound for the v-support of f^n. □

Now we shall prove the main theorem of this section, Theorem 5.1.

Proof of Theorem 5.1. We first show that

$$\left\|f^{n+1}\right\|_{W^{1,p}}<\infty,$$

uniformly in n, for fixed $3<p\le\infty$. It suffices to show it in $W^{1,\infty}$. In doing so, we shall prove that $\|\nabla\varphi^n\|_{W^{1,\infty}}$ does not grow faster than $\log\|f^n\|_{W^{1,\infty}}$ in time. Since $\Delta\varphi^n=\rho^n$, $\varphi^n|_{\partial\Omega}=0$, we have

$$\varphi^n(t,x)=\int_\Omega\rho^n(t,y)\,G(x,y)\,dy,$$

where $G(x,y)$ is Green's function for the Laplacian, associated with the domain Ω. Then we find in [6] or [17] that

$$|\nabla_xG(x,y)|\le\frac{C}{|x-y|^2},\quad\left|\nabla_x^2G(x,y)\right|\le\frac{C}{|x-y|^3}.$$

Assuming that the v-support of f^n is uniformly bounded in n, we have

$$\begin{aligned}
&\partial_{x_ix_j}\varphi^n(t,x)\\
&=\partial_{x_ix_j}\int_\Omega G(x,y)\,\rho^n(t,y)\,dy\\
&=\partial_{x_i}\int_\Omega \partial_{x_j}G(x,y)\,\rho^n(t,y)\,dy\\
&=\partial_{x_i}\int_\Omega \partial_{x_j}G(x,y)\,[\rho^n(t,y)-\rho^n(t,x)]\,dy+\partial_{x_i}\int_\Omega \partial_{x_j}G(x,y)\,\rho^n(t,x)\,dy\\
&=\int_\Omega \partial_{x_ix_j}G(x,y)\,[\rho^n(t,y)-\rho^n(t,x)]\,dy-\int_\Omega \partial_{x_j}G(x,y)\,\partial_{x_i}\rho^n(t,x)\,dy\\
&\quad+\partial_{x_i}\rho^n(t,x)\int_\Omega \partial_{x_j}G(x,y)\,dy+\rho^n(t,x)\,\partial_{x_ix_j}\int_\Omega G(x,y)\,dy\\
&=\rho^n(t,x)\,\partial_{x_ix_j}\int_\Omega G(x,y)\,dy+\int_{|x-y|\ge a}\partial_{x_ix_j}G(x,y)\,[\rho^n(t,y)-\rho^n(t,x)]\,dy\\
&\quad+\int_{|x-y|\le a}\partial_{x_ix_j}G(x,y)\,[\rho^n(t,y)-\rho^n(t,x)]\,dy.
\end{aligned}$$

Hence we get

$$\begin{aligned}
&\left|\partial_{x_ix_j}\varphi^n(t,x)\right|\\
&\le C+C\int_{|x-y|\ge a}\frac{1}{|x-y|^3}dy+C\int_{|x-y|\le a}\frac{1}{|x-y|^3}\,\|\rho^n\|_{W^{1,\infty}(\Omega)}\,|x-y|\,dy\\
&\le C+C\,|\log a|+Ca\,\|f^n\|_{W^{1,\infty}(\Omega)}\\
&\le C\left[1+\log\left(1+\|f^n\|_{W^{1,\infty}(\Omega)}\right)\right].
\end{aligned}\tag{5.26}$$

Here we have chosen a with $a=\|f^n\|^{-1}_{W^{1,\infty}(\Omega)}$. For $\partial_t\nabla\varphi^n$, we employ from (5.1)

$$\rho^n_t+\operatorname{div}_x j^n=0,$$

where $j^n(t,x)=\int vf^n(t,x,v)\,dv$, to get

$$\begin{aligned}
&\partial_{tx_i}\varphi^n(t,x)\\
&=\partial_{x_i}\int_\Omega \operatorname{div}_y j^n(t,y)\,G(x,y)\,dy\\
&=\int_{|x-y|\le a}\operatorname{div}_y j^n(t,y)\,\partial_{x_i}G(x,y)\,dy+\int_{|x-y|\ge a}\operatorname{div}_y j^n(t,y)\,\partial_{x_i}G(x,y)\,dy\\
&=\int_{|x-y|\le a}\operatorname{div}_y j^n(t,y)\,\partial_{x_i}G(x,y)\,dy+\int_{|x-y|\ge a}\operatorname{div}_y\left(j^n(t,y)\,\partial_{x_i}G(x,y)\right)dy\\
&\quad-\int_{|x-y|\ge a}j^n(t,y)\,\nabla_y\partial_{x_i}G(x,y)\,dy.
\end{aligned}$$

Therefore, by the same choice of a, we have

$$\begin{aligned}
&|\partial_{tx_i}\varphi^n(t,x)| \\
&\le Ca\|f^n\|_{W^{1,\infty}(\Omega)} + \left|\int_{|x-y|=a} \frac{x-y}{a}\cdot j^n(t,y)\,\partial_{x_i}G(x,y)\,dS_a(y)\right| \\
&\quad + \int_{|x-y|\ge a} \frac{|j^n(t,y)|}{|x-y|^3}dy \\
&\le Ca\|f^n\|_{W^{1,\infty}(\Omega)} + C + C|\log a| \\
&\le C\left[1+\log\left(1+\|f^n\|_{W^{1,\infty}(\Omega)}\right)\right].
\end{aligned}$$

Now we start with the first derivatives of f^n. By taking v-derivatives of (5.1), we get

$$\partial_t(\partial_v f^n) + v\cdot\nabla_x(\partial_v f^n) + \nabla_x\varphi^{n-1}\cdot\nabla_v(\partial_v f^n) = -\partial_x f^n.$$

Along the trajectory given by $\frac{d}{ds}X^{n-1} = V^{n-1}$, $\frac{d}{ds}V^{n-1} = \nabla_x\varphi^{n-1}$, we have

$$\partial_s\left[\partial_v f^n\left(s, X^{n-1}(s), V^{n-1}(s)\right)\right] = -\partial_x f^n,$$

and thus

$$|\partial_v f^n(t,x,v)| \le C + \int_0^t |\partial_x f^n(s)|\,ds.$$

For $\partial_x f^n$, we have

$$\partial_s\left[\partial_x f^n\left(s, X^{n-1}(s), V^{n-1}(s)\right)\right] = -\left(\nabla_x\partial_x\varphi^{n-1}\right)\cdot\nabla_v f^n\left(s, X^{n-1}(s), V^{n-1}(s)\right),$$

which implies, upon integrating over time,

$$\partial_x f^n(t,x,v) = \begin{cases} \partial_x f_0(x_0,v_0) - \int_0^t \left(\partial_x^2\varphi^{n-1}\right)\cdot\left(\nabla_v f^n\right)(s)\,ds, \\ \lim_{s\to t_0(t,x,v)}\partial_x f^n(s) - \int_{t_0}^t \left(\partial_x^2\varphi^{n-1}\right)\cdot\left(\nabla_v f^n\right)(s)\,ds, \end{cases}$$

depending on the back trajectory from (t,x,v) to either $(0,x_0,v_0)$ or (t_0,x_0,v_0) with $x_0\in\partial\Omega$. To compute $\lim_{s\to t_0(t,x,v)}\partial_x f^n(s)$, we look at $\nabla_x f^n(t_0,x_0,v_0) = \nabla^T g(t_0,x_0,v_0) + \lim_{s\to t_0}\nabla^\perp f^n\left(s, X^{n-1}(s), V^{n-1}(s)\right)$. From the transport equation

$$f_t^n + v^T\cdot\nabla^T f^n + (v\cdot n_x)\nabla^\perp f^n + \nabla_x\varphi^{n-1}\cdot\nabla_v f^n = 0,$$

we get, by the assumption $|\nabla g(t_0,x_0,v_0)| \le C|v_0\cdot n_{x_0}|$,

$$\begin{aligned}
\left|\lim_{s\to t_0(t,x,v)}\partial_x f^n(s)\right| &= \left|-(v\cdot n_x)^{-1}\left[g_t + v_0^T\cdot\nabla^T g + \nabla_x\varphi^{n-1}\cdot\nabla_v g\right]|_{(t_0,x_0,v_0)}\right| \\
&\le C,
\end{aligned}$$

since the support of g is bounded and $\left\|\nabla_x\varphi^{n-1}\right\|_\infty$ is uniformly bounded. Therefore, we have, by (5.26),

$$\begin{aligned}
\left\|\partial_{(x,v)}f^n(t)\right\|_\infty &\le C + \int_0^t \left(1+\left\|\nabla_x\varphi^{n-1}(s)\right\|_{W^{1,\infty}}\right)\left\|\partial_{(x,v)}f^n(s)\right\|_\infty ds \\
&\le C + \int_0^t \log\left(1+\left\|f^{n-1}(s)\right\|_{W^{1,\infty}}\right)\left\|\partial_{(x,v)}f^n(s)\right\|_\infty ds,
\end{aligned}$$

where C depends on $\|f_0\|_{W^{1,\infty}}$, $\|g\|_{W^{1,\infty}}$, the support of f_0 and g, $\left\|\nabla\varphi^{n-1}\right\|_\infty$, and the constant C which appears in the vanishing condition on g. We thus get a uniform bound on $\|f^n(t)\|_{W^{1,\infty}}$ by the Gronwall inequality. For the t-derivative, we employ

$$f_t^n + v\cdot\nabla_x f^n + \nabla_x\varphi^{n-1}\cdot\nabla_v f^n = 0$$

to get a uniform bound on $\|\partial_t f^n\|_\infty$. It follows that $\|f^n\|_{W^{1,p}}$ is uniformly bounded in n. For higher derivatives, for general $i \le k$, we can take the derivatives repeatedly in (5.1). The only term involving $\left\|\nabla_x\varphi^{n-1}\right\|_{W^{1,p}}$ is that $\partial^\alpha\nabla\varphi^{n-1}\partial_v f^n$, where $|\alpha| = i$. Since $\|\partial_v f^n\|_\infty \le C$, we then have

$$\left\|\partial^\alpha\nabla\varphi^{n-1}\partial_v f^n\right\|_p \le \left\|\partial^\alpha\nabla\varphi^{n-1}\right\|_p \|\partial_v f^n\|_\infty \le \left\|\partial^\alpha\nabla\varphi^{n-1}\right\|_p.$$

We know, by elliptic theory, that

$$\|\nabla\varphi^n\|_{W^{i,p}} \le C\|f^n\|_{W^{i-1,p}}.$$

Thus, $\left\|f^{n+1}\right\|_{W^{k,p}}$ is unformly bounded in n, and the uniform boundedness of $\|\varphi^n\|_{W^{k+2,p}}$ is also established from elliptic theory again. Once we have shown that $\left\|f^{n+1}\right\|_{W^{k,p}}$ and $\|\varphi^n\|_{W^{k+2,p}}$ are uniformly bounded in n, we obtain their weak limits f and φ, respectively, in $W^{k,p}$ and $W^{k+2,p}$ by a standard compactness argument. Last, since $\nabla\varphi^n$ converges strongly to $\nabla\varphi$ by compact embedding ($p > 3$), we can pass to the limits in (5.1) and (5.2). The uniqueness of solution for (5.1) and (5.2) can be attained by considering the difference between two solutions with the same initial and boundary conditions. It reduces to looking at the solution for the same (Vlasov–Poisson) system (5.1) and (5.2) with vanishing initial and boundary data. Integrating (5.1) over all x, v leads to the decrease of the L^1 norm in time. Together with the positivity of solution, we obtain the uniqueness. Hence, our theorem follows. ∎

6. Regularity for the Vlasov–Poisson system with the purely specular boundary condition. In this section, we assume the purely specular boundary condition for the Vlasov equation with the Dirichlet boundary condition on the electric potential. The Vlasov–Poisson system takes the form

$$\begin{gathered} f_t + v\cdot\nabla_x f + \nabla\varphi\cdot\nabla_v f = 0, \quad f|_{t=0} = f_0(x,v), \\ f(t,x,v) = f(t,x,v_*), \quad x\in\partial\Omega, \\ \Delta\varphi = \rho = 4\pi\int f dv, \\ \varphi|_{\partial\Omega} = 0, \end{gathered} \tag{6.1}$$

where f_0 is a given initial datum. We now restrict ourselves to the case when $\Omega = B$ and f_0 is spherically symmetric, where B is the unit ball in $\mathbb{R}^3$. Then we look for a spherically symmetric solution of (6.1) with datum f_0. Our main result in this section is the following.

THEOREM 6.1. *Assume that there is an $\omega_0 > 0$ such that $f_0(x,v)$ is constant for $\left(1-|x|^2\right)^2 + (2v\cdot x)^2 \le \omega_0$.*

(a) *Assume $f_0 \in C^1$. Let f_0 have compact support and satisfy the compatibility conditions* (4.3)*. Let f_0 be spherically symmetric. Then there exists a unique spherically symmetric solution (f,φ) of* (6.1) *such that $f \in W^{1,\infty}$ with compact support.*

(b) *Assume $f_0 \in C^{1,\eta}$ for some $\eta > 0$. Let f_0 have compact support and satisfy the compatibility conditions* (4.3) *and* (4.4)*. Let f_0 be spherically symmetric. Then*

there exists a unique spherically symmetric solution (f,φ) *of* (6.1) *such that* $f \in C^{1,\mu}$, $\varphi \in C^{3,\mu}$ *for some* $0 < \mu < \eta$, *with compact support.*

We first give a preliminary lemma on some conserved quantities in a bounded domain without the spherically symmetric assumption.

LEMMA 6.2. *Let* f *be a classical solution of* (6.1) *on some time interval* $(0,T]$ *with a nonnegative compactly supported datum* $f_0 \in C^1\left(\Omega \times \mathbb{R}^3\right)$. *Then we have the following:*

(a) *The total mass is conserved, i.e.,*

$$\int_\Omega \int_{\mathbb{R}^3} f dv dx \equiv \text{constant} = M_0.$$

(b) *The total energy is conserved, i.e.,*

$$\int_\Omega \left[\int_{\mathbb{R}^3} |v|^2 f dv + |E|^2 \right] dx \equiv \text{constant} = \epsilon_0.$$

(c) $\|\rho(t)\|_{L^{5/3}} \le C$ *for* $0 \le t \le T$, $C = C\left(\|f_0\|_\infty, \epsilon_0\right)$.

Proof. For (a), notice that from the specular boundary condition $f|_\gamma$ is an even function of the normal component $x \cdot v$ of v. Hence by integrating the Vlasov equation over x and v, we get

$$0 = -\int_{\partial B} \int_{\mathbb{R}^3} x \cdot v f(t,x,v)\, dv dS_x = \int_B \rho(t,x)\, dx - \int \rho_0(x)\, dx.$$

Therefore, the total mass is conserved. For (b), by multiplying the Vlasov equation (6.1) with $|v|^2$ and then integrating over x and v, we get

$$\begin{aligned} 0 &= \partial_t \int_\Omega \int_{\mathbb{R}^3} |v|^2 f dv dx + \int_{\mathbb{R}^3} \int_\Omega \nabla_x \cdot \left(|v|^2 v f \right) dx dv \\ &\quad - 2 \iint v f \cdot E dv dx \\ &= \partial_t \int_\Omega \int_{\mathbb{R}^3} |v|^2 f dv dx + \int_{\mathbb{R}^3} \int_{\partial\Omega} v \cdot n_x |v|^2 f dS_x dv \\ &\quad - 2 \iint v f \cdot E dv dx \\ &= \partial_t \int_\Omega \int_{\mathbb{R}^3} |v|^2 f dv dx - 2 \int_\Omega j \cdot E dx. \end{aligned}$$

Here we have used the specular reflection condition at the boundary. Now we integrate (6.1) over x and v to get

$$\rho_t + \nabla_x \cdot j = 0,$$

where $j(t,x)=\int vfdv$. Next we observe that

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\int_\Omega |E|^2\,dx &= \int_\Omega E\cdot E_t dx = \int_\Omega \nabla\varphi\cdot\nabla\varphi_t dx\\ &= \int_{\partial\Omega}\varphi n_x\cdot\nabla\varphi_t dS_x - \int_\Omega \varphi\Delta\varphi_t dx\\ &= -\int_\Omega \varphi\rho_t dx\\ &= \int_\Omega \varphi\nabla_x\cdot j dx\\ &= \int_{\partial\Omega}\varphi\left[\int n_x\cdot vfdv\right]dS_x - \int_\Omega\nabla\varphi\cdot j dx\\ &= -\int_\Omega j\cdot E dx,\end{aligned}$$

since $\varphi|_{\partial\Omega}=0$. Hence we obtain part (b). Since $|E(t,x)|\le Cr^{-2}*\rho(t,x)$, we have, by the Hardy–Littlewood–Sobolev lemma,

$$\begin{aligned}\|E(t)\|_{L^2} &\le C\left\|r^{-2}*\rho(t,\cdot)\right\|_{L^2}\le C\|\rho(t)\|_{L^{6/5}}\\ &\le C\|\rho\|_{L^1}^{7/12}\|\rho(t)\|_{L^{5/3}}^{5/12}\le C\|\rho(t)\|_{L^{5/3}}^{5/12}.\end{aligned}$$

By a standard interpolation method, we get

$$\int_\Omega \rho^{5/3}dx\le C\int|v|^2 fdv\le C. \qquad \square$$

Now we construct approximate solutions for (6.1) through an iterating sequence. Let f^0 be a suitable smooth extension of f_0 to Π, which satisfies the compatibility conditions (4.3), (4.4). Consider the following iterating sequences:

$$\begin{aligned}&f_t^{n+1}+v\cdot\nabla_x f^{n+1}+\nabla\varphi^n\cdot\nabla_v f^{n+1}=0,\quad f^{n+1}|_{t=0}=f_0,\\ &f^{n+1}(t,x,v)=f^{n+1}(t,x,v_*),\quad x\in\partial B,\\ &\Delta\varphi^n=\rho^n=4\pi\int f^n dv,\quad \varphi|_{\partial B}=0.\end{aligned}\tag{6.2}$$

In contrast to the absorbing boundary case, we shall adopt the idea in [14] in order to get a global bound for the velocity in the spherically symmetric case. The key point is to employ the invariance of the angular momentum in order to control the particles with small tangential angles near the boundary.

We assume that f_0 is spherically symmetric; i.e., $f_0(\Lambda x,\Lambda v)=f_0(x,v)$ for every proper rotation Λ on $\mathbb{R}^3$. It is known that the solution $f(t,x,v)$ satisfies the same property in x and v, and therefore depends only on $r\equiv|x|$, $u\equiv|v|$, α, and t, where α is the angle between x and v. The density ρ depends then only on r and t and has the following representation:

$$\rho(t,r)=2\pi\int_0^\infty\int_0^\pi f(t,r,u,\alpha)\,u^2\sin\alpha d\alpha du.$$

Thus φ is also radial and has the following relation with ρ:

$$\begin{aligned}\varphi(t,r)&=-\frac{1}{r}\int_0^r\lambda^2\rho(t,\lambda)\,d\lambda-\int_r^1\lambda\rho(t,\lambda)\,d\lambda+M_0/4\pi,\\ &\lim_{r\to0}r^2\varphi_r(t,r)=0.\end{aligned}$$

To see this relation, we use the harmonic operator Δ_x in the spherical coordinates, $\partial_{rr} + \frac{2}{r}\partial_r$. Hence the corresponding Poisson equation in these spherical coordinates is

$$\varphi_{rr} + \frac{2}{r}\varphi_r = \rho. \tag{6.3}$$

By multiplying (6.3) with r and r^2, respectively, we get

$$(r\varphi_r + \varphi)_r = r\rho, \tag{6.4}$$

$$\left(r^2\varphi_r\right)_r = r^2\rho. \tag{6.5}$$

We integrate (6.4) from r to 1 and (6.5) from 0 to r to get

$$\varphi_r(t,1) - r\varphi_r - \varphi = \int_r^1 \lambda\rho(t,\lambda)\,d\lambda, \tag{6.6}$$

$$r\varphi_r = \frac{1}{r}\int_0^r \lambda^2\rho(t,\lambda)\,d\lambda, \tag{6.7}$$

where we used the Dirichlet boundary condition. Plugging (6.7) into (6.6) yields

$$\varphi(t,r) = -\frac{1}{r}\int_0^r \lambda^2\rho(t,\lambda)\,d\lambda - \int_r^1 \lambda\rho(t,\lambda)\,d\lambda + \varphi_r(t,1),$$
$$E(t,x) = \nabla_x\varphi(t,r) = \frac{x}{r^3}\int_0^r \lambda^2\rho(t,\lambda)\,d\lambda = r^{-2}M(t,r)\frac{x}{r},$$

where $M(t,r) = \int_0^r \lambda^2\rho(t,\lambda)\,d\lambda$. Note that $|E| = r^{-2}M(t,r)$ and $M(t,1) = M_0/4\pi$ for all t. Note also that for $x \in \partial B$, $n_x \cdot E(t,x) = x \cdot E(t,x) = M(t,1) = M_0/4\pi \equiv$ constant. Hence, $\varphi_r(t,1) = \frac{x}{r}\cdot\nabla_x\varphi(t,x)|_{r=1} = M(t,1) = M_0/4\pi$. Moreover, the normal component of the electric field at the boundary is unchanged over time. This satisfies the condition for Corollaries 2.3 and 2.8.

Spherical symmetry also leads to a simplification of the trajectory equations:

$$\begin{aligned}
\frac{dR}{d\tau} &= U\cos A,\\
\frac{dU}{d\tau} &= \frac{\cos A}{R^2}M(\tau,R),\\
\frac{dA}{d\tau} &= -\left(\frac{M(\tau,R)}{R^2U} + \frac{U}{R}\right)\sin A,
\end{aligned} \tag{6.8}$$

where $R(t;t,r,u,\alpha) = r$, $U(t;t,r,u,\alpha) = u$, $A(t;t,r,u,\alpha) = \alpha$. Notice from (6.8) that we have the invariance of angular momentum, i.e.,

$$RU\sin A = ru\sin\alpha \text{ for all } \tau. \tag{6.9}$$

This is a crucial fact which will be used to treat such trajectories with small tangential angles near the boundary and lead to the velocity bounds. We now define, for $t \geq 0$,

$$P(t) = \sup\left\{U(s;0,r,u,\alpha)\,\middle|\,0 \leq s \leq t,\ (r,u,\alpha) \in \{\text{support of } f_0\}\right\}$$

and note that P is nondecreasing.

LEMMA 6.3. *There exists a constant C_1 such that for $r \geq 0$ and $0 \leq t \leq T$,*

$$|E(t,x)| = \frac{M(t,r)}{r^2} \leq \min\left(M_0 r^{-2}, C_1 P^{4/3}(t)\right),$$

where C_1 depends only on $\|f_0\|_\infty$ and $\|\rho(t)\|_{L^{5/3}}$ and M_0 is the total mass.

Proof. Clearly $|E(t,x)| \leq M_0 r^{-2}$. We employ the Poisson equation $\Delta\varphi = \rho$ with the Dirichlet boundary condition for φ and $E = \nabla_x \varphi$. Let $0 < R_0 < 2$, and note that

$$\begin{aligned}
|E(t,x)| &\leq \int \frac{\rho(t,y)}{|x-y|^2} dy \\
&= \int_{|x-y|<R_0} \frac{\rho(t,y)}{|x-y|^2} dy + \int_{|x-y|>R_0} \frac{\rho(t,y)}{|x-y|^2} dy \\
&\leq \|\rho(t)\|_\infty \int_{|x-y|<R_0} |x-y|^{-2} dy \\
&\quad + \|\rho(t)\|_{L^{5/3}} \left[\int_{|x-y|>R_0} \left(|x-y|^{-2}\right)^{5/2} dy\right]^{2/5} \\
&= \|\rho(t)\|_\infty 4\pi R_0 + \|\rho(t)\|_{L^{5/3}} \left[2\pi\left(R_0^{-2} - 2^{-2}\right)\right]^{2/5} \\
&\leq \|\rho(t)\|_\infty 4\pi R_0 + \|\rho(t)\|_{L^{5/3}} (2\pi)^{2/5} R_0^{-4/5},
\end{aligned}$$

where we have used that $|\nabla G(x,y)| \leq C/|x-y|^2$ for Green's function $G(x,y)$ for the unit ball. Now we choose $R_0 > 0$ such that $R_0 \|\rho(t)\|_\infty = \|\rho(t)\|_{L^{5/3}} R_0^{-4/5}$ or $R_0 = \left(\|\rho(t)\|_{L^{5/3}} / \|\rho(t)\|_\infty\right)^{5/9}$. Here we may assume that $R_0 < 2$ because otherwise we would have $\|\rho(t)\|_\infty \leq C$, and so $\|E(t,x)\|_\infty \leq 8\pi \|\rho(t)\|_\infty \leq C$. Then we have

$$\begin{aligned}
|E(t,x)| &\leq C \|\rho(t)\|_{L^{5/3}}^{5/9} \|\rho(t)\|_\infty^{4/9} \\
&\leq C \|\rho(t)\|_{L^{5/3}}^{5/9} \|f_0\|_\infty^{4/9} P^{4/3}(t) \\
&\leq C_1 P^{4/3}(t).
\end{aligned}$$

This completes the proof of the lemma. □

Now consider a trajectory through some point (r_0, u_0, α_0) with a positive angular momentum $r_0 u_0 \sin\alpha_0 > 0$. Then $L = R(s) U(s) \sin A(s)$ is a positive invariant along the trajectory. We define

$$K(t,r) = -\int_r^1 \min\left(M_0 \lambda^{-2}, C_1 P^{4/3}(t)\right) d\lambda$$

for $0 \leq r \leq 1$ and $t \geq 0$. Note that K is continuously differentiable in r and increasing in r. We also let $R_0 = M_0^{1/2} \left(C_1 P^{4/3}(t)\right)^{-1/2}$. If $0 \leq \lambda \leq R_0$, then $C_1 P^{4/3}(t) \leq M_0 \lambda^{-2}$, and if $R_0 \leq \lambda \leq 1$, then $M_0 \lambda^{-2} \leq C_1 P^{4/3}(t)$. Here again we may assume without loss of generality that $R_0 < 1$, since otherwise $M_0^{1/2} \left(C_1 P^{4/3}(t)\right)^{-1/2} \geq 1$

would imply the bound for $P(t)$. We compute

$$\begin{aligned}
K(t,0) &= -\int_0^1 \min\left(M_0\lambda^{-2}, C_1P^{4/3}(t)\right)d\lambda \\
&= -\int_0^{R_0} C_1P^{4/3}(t)\,d\lambda - \int_{R_0}^1 M_0\lambda^{-2}d\lambda \\
&= -C_1P^{4/3}(t)\,R_0 + M_0 - M_0^{1/2}\left(C_1P^{4/3}(t)\right)^{1/2} \\
&= -M_0^{1/2}\left(C_1P^{4/3}(t)\right)^{1/2} + M_0 - M_0^{1/2}\left(C_1P^{4/3}(t)\right)^{1/2} \\
&\geq -2M_0^{1/2}\left(C_1P^{4/3}(t)\right)^{1/2}.
\end{aligned}$$

Therefore

$$\begin{aligned}
|K(t,r_1) - K(t,r_2)| &\leq |K(t,0)| \leq 2M_0^{1/2}\left(C_1P^{4/3}(t)\right)^{1/2} \\
&= C_2P^{4/6}(t), \quad C_2 = 2M_0^{1/2}C_1^{1/2}.
\end{aligned}$$

LEMMA 6.4. *Assume that either $\dot{R} \geq 0$ on $[t_1,t_2]$ or $\dot{R} \leq 0$ on $[t_1,t_2]$. Then*

$$\left|\frac{1}{2}U^2(t_2) - \frac{1}{2}U^2(t_1)\right| \leq |K(R(t_2),t_2) - K(R(t_1),t_2)|.$$

Proof. Note that

$$\begin{aligned}
|K(R(t_2),t_2) - K(R(t_1),t_2)| &= \left|\int_{t_1}^{t_2} \frac{\partial K}{\partial r}(R(s),t_2)\,\dot{R}(s)\,ds\right| \\
&= \int_{t_1}^{t_2}\left|\frac{\partial K}{\partial r}(R(s),t_2)\,\dot{R}(s)\right|ds,
\end{aligned}$$

since $\dot{R}$ is of one sign on $[t_1,t_2]$ and $\partial K/\partial r \geq 0$. We also note that for $t_2 \geq s$,

$$\begin{aligned}
\frac{\partial K}{\partial r}(R(s),t_2) &= \min\left(M_0\lambda^{-2}, C_1P^{4/3}(t_2)\right) \\
&\geq \min\left(M_0\lambda^{-2}, C_1P^{4/3}(s)\right) \\
&= \frac{\partial K}{\partial r}(R(s),s).
\end{aligned}$$

Since $|E(t,x)| = r^{-2}M(t,r) \leq \partial K/\partial r\,(t,r)$ and by the trajectory equations (6.8), we have

$$\begin{aligned}
\left|\frac{\partial K}{\partial r}(R(s),t_2)\,\dot{R}(s)\right| &\geq \left|\frac{\partial K}{\partial r}(R(s),s)\,\dot{R}(s)\right| \\
&\geq \left|\frac{M(R(s),s)}{R^2(s)}U(s)\cos A(s)\right| = \left|U(s)\,\dot{U}(s)\right| \\
&= \left|\frac{d}{ds}\frac{1}{2}U^2(s)\right|.
\end{aligned}$$

Therefore

$$\begin{aligned}|K(R(t_2),t_2)-K(R(t_1),t_2)| &\geq \int_{t_1}^{t_2}\left|\frac{d}{ds}\frac{1}{2}U^2(s)\right|ds\\ &\geq \left|\int_{t_1}^{t_2}\frac{d}{ds}\frac{1}{2}U^2(s)\,ds\right|\\ &= \left|\frac{1}{2}U^2(t_2)-\frac{1}{2}U^2(t_1)\right|.\end{aligned}$$

Thus the proof of the lemma is complete. □

LEMMA 6.5. *On each interval where the trajectory is smooth, $\dot{R}$ can be zero at most one value of s. If $\dot{R}(t_1)=0$, then R has an absolute minimum at t_1 on the interval.*

Proof. Recall that $R(s)U(s)\sin A(s) = r_0u_0\sin\alpha_0 \neq 0$ by hypothesis. So $R(s)\neq 0$, $U(s)\neq 0$, and $\sin A(s)\neq 0$ for all s. From (6.8),

$$\dot{R}(s) = U(s)\cos A(s);$$

thus $\dot{R}=0$ only if $A(s)=\pi/2$. However, also from (6.8),

$$\dot{A}(s) = -\left(\frac{M(R(s),s)}{R^2(s)U(s)}+\frac{U(s)}{R(s)}\right)\sin A(s) < 0$$

for all s. $A(s)$ is thus strictly decreasing for s on the interval and hence can attain the value of $\pi/2$ at most once. So $\dot{R}$ can be zero at most once as long as the trajectory is smooth on the interval. Now suppose that $\dot{R}(t_1)=0$; then $A(s)>\pi/2$ for $s<t_1$, $A(t_1)=\pi/2$, and $A(s)<\pi/2$ for $s>t_1$. From (6.8), $\dot{R}<0$ for $s<t_1$, $\dot{R}(t_1)=0$, and $\dot{R}>0$ for $s>t_1$. Therefore R has an absolute minimum at t_1 on the smooth interval. □

Now we consider the trajectory from a generic point (t,x,v), and we compute the lower bound on the time spent travelling from one boundary point to another along the trajectory.

LEMMA 6.6. *Let $\left(t^0,x^0,v^0\right)$ and $\left(t^1,x^1,v^1\right)$ be two points on the trajectory, where $x^0,x^1\in\partial B$, $t^0<t^1$. Suppose that the trajectory stays inside the unit ball B on the interval $\left(t^0,t^1\right)$. Then*

$$t^1-t^0 \geq \min\left(-\frac{\left|v^0\right|\cos\alpha^0}{3\sup_{0\leq s\leq t}\|E(s)\|_\infty}, -\frac{\cos\alpha^0}{3\left|v^0\right|}, \frac{1}{\left[\sup_{0\leq s\leq t}\|E(s)\|_\infty\right]^{1/2}}\right),$$

where α^0 is the angle between x^0 and v^0.

Proof. Since

$$x^1 = x^0+\int_{t^0}^{t^1}V(\tau)\,d\tau, \quad V(\tau) = v^0+\int_{t^0}^{\tau}E(s)\,ds,$$

we have

$$\begin{aligned}1=\left|x^1\right|^2 &= \left(x^0+\int_{t^0}^{t^1}V(\tau)\,d\tau\right)\cdot\left(x^0+\int_{t^0}^{t^1}V(\tau)\,d\tau\right)\\ &= 1+2x^0\cdot\int_{t^0}^{t^1}\left[v^0+\int_{t^0}^{\tau}E(s)\,ds\right]d\tau+\left|\int_{t^0}^{t^1}V(\tau)\,d\tau\right|^2.\end{aligned}$$

So, if $t^1 - t^0 \le \frac{1}{[\sup_{0\le s\le t}\|E(s)\|_\infty]^{1/2}}$, then

$$\begin{aligned}
0 &\le 2x^0 \cdot v^0 \left(t^1 - t^0\right) + 2 \sup_{0\le s\le t} \|E(s)\|_\infty \left(t^1 - t^0\right)^2 \\
&\quad + 2\left|v^0\right|^2 \left(t^1 - t^0\right)^2 + 2\left(\sup_{0\le s\le t} \|E(s)\|_\infty\right)^2 \left(t^1 - t^0\right)^4, \\
0 &\le x^0 \cdot v^0 + \sup_{0\le s\le t} \|E(s)\|_\infty \left(t^1 - t^0\right) + \left|v^0\right|^2 \left(t^1 - t^0\right) \\
&\quad + \sup_{0\le s\le t} \|E(s)\|_\infty \left(t^1 - t^0\right).
\end{aligned}$$

Thus we get

$$t^1 - t^0 \ge \frac{-x^0 \cdot v^0}{\left|v^0\right|^2 + 2\sup_{0\le s\le t}\|E(s)\|_\infty}.$$

Note that $-x^0 \cdot v^0 > 0$. If $\sup_{0\le s\le t}\|E(s)\|_\infty \le \left|v^0\right|^2$,

$$t^1 - t^0 \ge -\frac{\left|v^0\right|\cos\alpha^0}{3\sup_{0\le s\le t}\|E(s)\|_\infty},$$

and if $\sup_{0\le s\le t}\|E(s)\|_\infty \ge \left|v^0\right|^2$,

$$t^1 - t^0 \ge -\frac{\cos\alpha^0}{3\left|v^0\right|}.$$

We thus obtain the lemma. □

In the presence of the boundary, the central obstacle comes from the particles near the boundary with so many bounces or with small tangential angles, in addition to the difficulty of controlling the particles with high velocity. However, the invariance of the angular momentum enables us to overcome this main barrier. The angular momentum of the particles near the boundary with small tangential angle amounts approximately to the full velocity. This observation suggests that the initial control on the invariant angular momentum would reduce to the concern only on the high velocities and thus lead to resolving our difficulty.

Now fix $M_1 > 0$. Suppose that $f_0(x,v) \equiv 0$ when the angular momentum

$$\digamma = ru\sin\alpha = |x|\,|v|\sin\alpha \ge M_1.$$

Note that if (t,x,v) connects with an initial point in the support of f_0, $x \in \partial B$, and $|\alpha - \pi/2| \le \pi/6$, then

$$\frac{1}{2}|v| \le |v|\sin\alpha = \digamma \le M_1,$$

and therefore

$$|v| \le 2M_1. \tag{6.10}$$

Let $\left(t^0, x^0, v^0\right)$ be the first point at the boundary on the back-time cycle from (t,x,v) such that $|v_0| \le 2M_1$, and, if it does not happen through the whole cycle, then let

$t_0 = 0$. Let the back-time cycle from the time t to the time t^0 be $(t, x, v) = (t^n, x^n, v^n)$, $\left(t^{n-1}, x^{n-1}, v^{n-1}\right)$, $\ldots$, $\left(t^0, x^0, v^0\right)$. We compute $\Delta t^i = t^{i+1} - t^i$ for $0 \le i \le n-2$. By the definition of the time t^0, we have, for all i, $\left|\alpha^i - \pi/2\right| > \pi/6$ and $\left|v^i\right| \ge 2M_1$ from (6.10). We then apply Lemma 6.6 to compute the time spent for each bounce, and we have

$$
\begin{aligned}
-\frac{\left|v^i\right| \cos\alpha^i}{3 \sup_{0\le s\le t} \|E(s)\|_\infty} &\ge \frac{2M_1 \frac{\sqrt{3}}{2}}{3C_1 P^{4/3}(t)} = \frac{M_1}{\sqrt{3}C_1} P^{-4/3}(t), \\
-\frac{\cos\alpha^i}{3\left|v^i\right|} &\ge \frac{\sqrt{3}/2}{3P(t)} = \frac{1}{2\sqrt{3}} P^{-1}(t), \\
\frac{1}{\left[\sup_{0\le s\le t} \|E(s)\|_\infty\right]^{1/2}} &\ge C_1^{-1/2} P^{-4/6}(t).
\end{aligned} \tag{6.11}
$$

Since we are concerned only with large $P(t)$, we may assume that the minimum among (6.11) is $\frac{M_1}{\sqrt{3}C_1} P^{-4/3}(t)$, and thus $\Delta t^i \ge \frac{M_1}{\sqrt{3}C_1} P^{-4/3}(t)$. Therefore the number of bounces on the cycle through (t, x, v) until the time t^0 is at most

$$
\frac{\sqrt{3}TC_1}{M_1} P^{4/3}(t) + 2. \tag{6.12}
$$

Now we assert the control on the increase in velocity.

LEMMA 6.7. *Let f be a classical solution of* (6.1) *on $[0,T)$ with a smooth, non-negative, spherically symmetric data f_0 which has compact support and vanishes for $(r,u,\alpha) \notin (0,1] \times (0,\infty) \times (0,\pi)$, $ru\sin\alpha \ge M_1$, where M_1 is a fixed constant such that $M_1 > 8\sqrt{3}M_0^{1/2}C_1^{3/2}T$, M_0 is the total mass, and $C_1 = C\|\rho(t)\|_{L^{5/3}}^{5/9}\|f_0\|_\infty^{4/9}$. Then $P(t)$ is uniformly bounded on $[0,T)$.*

Proof. Let $t \in [0,T)$. Consider the back-time cycle from a generic point (t,x,v) with $0 < ru\sin\alpha \le M_1$. Suppose $0 \le t_1 \le t_2 \le t$ and the trajectory remains smooth on $[t_1,t_2]$, i.e., $[t_1,t_2]$ is between the two jump times, and suppose that $\dot{R} \ge 0$ on $[t_1,t_2]$ or $\dot{R} \le 0$ on $[t_1,t_2]$. Then, by Lemma 6.4,

$$
\begin{aligned}
\frac{1}{2}U^2(t_2) &\le \frac{1}{2}U^2(t_1) + \left|K(R(t_2),t_2) - K(R(t_1),t_2)\right| \\
&\le \frac{1}{2}U^2(t_1) + C_2 P^{4/6}(t_2) \\
&\le \frac{1}{2}U^2(t_1) + C_2 P^{4/6}(t),
\end{aligned} \tag{6.13}
$$

where $C_2 = 2M_0^{1/2}C_1^{1/2}$. Now we consider the back-time cycle from (t,x,v) until the point $\left(t^0,x^0,v^0\right)$ with $\left|v^0\right| \le 2M_1$. Suppose that $\dot{R}$ vanishes somewhere on $\left[t^i, t^{i+1}\right]$. By Lemma 6.5, there is only one such point where $\dot{R}$ vanishes; call it $\hat{t}^i \in \left[t^i, t^{i+1}\right]$. Then $\dot{R}$ cannot change sign on $\left[t^i, \hat{t}^i\right]$ or on $\left[\hat{t}^i, t^{i+1}\right]$. Hence applying (6.13) twice yields

$$
\begin{aligned}
\frac{1}{2}U^2\left(t^{i+1}\right) &\le \frac{1}{2}U^2\left(\hat{t}^i\right) + C_2 P^{4/6}(t) \\
&\le \frac{1}{2}U^2\left(t^i\right) + 2C_2 P^{4/6}(t).
\end{aligned}
$$

Thus by (6.12), we have through the back-time cycle until t^0,

$$\begin{aligned}\frac{1}{2}U^2(t) &\le \frac{1}{2}U^2\left(t^0\right) + 2NC_2P^{4/6}(t)\\ &\le \frac{1}{2}U^2\left(t^0\right) + 2C_2\left(\frac{\sqrt{3}TC_1}{M_1}P^{4/3}(t) + 2\right)P^{4/6}(t)\\ &\le \frac{1}{2}(2M_1)^2 + \frac{2\sqrt{3}C_1C_2T}{M_1}P^2(t) + 4C_2P^{4/6}(t),\end{aligned}$$

where N is the number of bounces through the cycle. Applying this argument to all possible (t,x,v), we deduce

$$P^2(t) \le (2M_1)^2 + 8C_2P^{4/6}(t) + \frac{4\sqrt{3}C_1C_2T}{M_1}P^2(t).$$

Since $4\sqrt{3}C_1C_2T = 8\sqrt{3}M_0^{1/2}C_1^{3/2}T < M_1$, we have

$$P^2(t) \le (2M_1)^2 + 8C_2P^{4/6}(t) + C_3P^2(t),$$

where $C_3 = \frac{4\sqrt{3}C_1C_2T}{M_1} < 1$. This implies that $P(t)$ is bounded by a constant depending on C_1, C_2, T, and M_1 or a constant depending only on the total mass M_0, the total energy ε_0, T, $\|f_0\|_\infty$, and M_1. Thus this completes the proof of the lemma. ∎

We consider our iterated scheme (6.2). Lemma 6.7 yields the uniform and global bound on the support of f^n.

Now we are ready to establish our main theorem in this section, Theorem 6.1.

Proof of Theorem 6.1. Uniqueness can be proven by using a standard Gronwall-type argument, since now the solutions are regular. We consider only the existence. By Lemma 6.7, ρ^n is uniformly bounded in L^∞. Therefore, φ^n is uniformly bounded in $W^{2,p}$ for $1 \le p < \infty$. Hence φ^n is uniformly bounded in $C^{1,\eta}$. We claim that

$$|\nabla\varphi^n(t,x) - \nabla\varphi^n(t,y)| \le -L|x-y|\log|x-y|,$$

where L is independent of n, and $|x-y|$ is small.

Proof of the claim. By the representation formula for the Poisson equation with the Dirichlet boundary condition for φ, we have

$$\nabla\varphi^n(t,x) = \int_B \rho^n(t,z)\left[\frac{x-z}{|x-z|^3} - \frac{x-\bar{z}}{|z|\,|x-\bar{z}|^3}\right]dz,$$

where $|\cdot|$ is the Euclidean distance in $\mathbb{R}^3$ and $\bar{z} = z/|z|^2$. Therefore,

$$\begin{aligned}|\nabla[\varphi^n(t,x) - \varphi^n(t,y)]| \le &\left|\int_B \rho^n(t,z)\left[\frac{x-z}{|x-z|^3} - \frac{y-z}{|y-z|^3}\right]dz\right|\\ &+ \left|\int_B \rho^n(t,z)\left[\frac{x-\bar{z}}{|z|\,|x-\bar{z}|^3} - \frac{y-\bar{z}}{|z|\,|y-\bar{z}|^3}\right]dz\right|.\end{aligned} \tag{6.14}$$

We first estimate the first term of (6.14) by splitting it as

$$\begin{aligned}&\left|\int_B \rho^n(t,z)\left[\frac{x-z}{|x-z|^3}-\frac{y-z}{|y-z|^3}\right]dz\right|\\&=\int_{\{z:\min[|x-z|,|y-z|]\geq|x-y|/2\}}+\int_{\{z:\min[|x-z|,|y-z|]\leq|x-y|/2\}}\\&=I_1+I_2.\end{aligned}$$

For I_1, we apply the mean value theorem to get

$$\left|\frac{x-z}{|x-z|^3}-\frac{y-z}{|y-z|^3}\right|\leq C|x-y|\left[\frac{1}{|x-z|^3}+\frac{1}{|y-z|^3}\right].$$

Thus

$$I_1\leq C|x-y|\int_{|x-y|/2}^{C}\frac{1}{r}dr\leq -L|x-y|\log|x-y|,$$

where L is a constant, independent of n. For I_2, without loss of generality, we may assume that $|x-z|\leq|x-y|/2$ and $|y-z|\geq|x-y|/2$. Then we have

$$\begin{aligned}\left|\frac{x-z}{|x-z|^3}-\frac{y-z}{|y-z|^3}\right|&=\left|\frac{x-z}{|x-z|^3}-\frac{x-z}{|y-z|^3}+\frac{x-z}{|y-z|^3}-\frac{y-z}{|y-z|^3}\right|\\&\leq\frac{2}{|x-z|^2}+\frac{|x-y|}{|y-z|^3}.\end{aligned}$$

Hence,

$$I_2\leq C\int_0^{|x-y|/2}1dr+C|x-y|\int_{|x-y|/2}^{C}\frac{1}{r}dr\leq -L|x-y|\log|x-y|.$$

Now we estimate the second term of (6.14). By the change of coordinates $z\longmapsto\bar{z}=z/|z|^2$,

$$\begin{aligned}\left|\int_B\rho^n(t,z)\left[\frac{x-\bar{z}}{|z||x-\bar{z}|^3}-\frac{y-\bar{z}}{|z||y-\bar{z}|^3}\right]dz\right|&\leq C\int_{|\bar{z}|\geq1}\left|\frac{x-\bar{z}}{|x-\bar{z}|^3}-\frac{y-\bar{z}}{|y-\bar{z}|^3}\right||z|\,d\bar{z}\\&\leq C\int\left|\frac{x-\bar{z}}{|x-\bar{z}|^3}-\frac{y-\bar{z}}{|y-\bar{z}|^3}\right|d\bar{z},\end{aligned}$$

which reduces to the first case. Thus our claim holds.

By Lemma 4.4, f^n is uniformly bounded in $C^{0,\eta}$ for some $\eta>0$. Hence, from the Poisson equation $\Delta\varphi^n=\rho^n$, $\sup_{1\leq\tau\leq t}\|\nabla_x\varphi^n\|_{C^{1,\eta}}$ and $\|\nabla_x\varphi^n\|_{C^{0,\eta}}$ are uniformly bounded. Applying Theorem 4.1 to f^n yields that f^n is uniformly bounded in $C^{1,\mu}$, $\mu>0$. Let f and φ be the limits of f^n and φ^n, respectively, such that

$$\sup_{0\leq\tau\leq t}\|\nabla_x\varphi\|_{C^{1,\mu}}+\|\nabla_x\varphi\|_{C^{0,\mu}}+\|f\|_{C^{1,\mu}}<\infty.$$

Therefore, our theorem follows by letting $n\to\infty$ in (6.2). □

Acknowledgments. The author truly thanks her thesis advisor, Yan Guo, for his encouragement and discussions. She also thanks Walter Strauss, Constantine Defermos, and Gerhard Rein for their constant interest and comments on this work. Finally, she would like to give her thanks to the referees for their helpful comments.

REFERENCES

[1] J. Batt, *Global symmetric solutions of the initial-value problem of stellar dynamics*, J. Differential Equations, 25 (1977), pp. 342–364.
[2] C. Bardos and D. Degond, *Global existence for the Vlasov-Poisson equation in* 3 *space variables with small initial data*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 2 (1985), pp. 101–118.
[3] R. Beals and V. Protopopescu, *Abstract time-dependent transport equations*, J. Math. Anal. Appl., 121 (1987), pp. 370–405.
[4] J. Batt and G. Rein, *Global classical solutions of the periodic Vlasov-Poisson system in three dimensions*, C. R. Acad. Sci. Paris Sér. I Math., 313 (1991), pp. 411–416.
[5] J. Cooper and A. Klimas, *Boundary value problems for the Vlasov-Maxwell equation in one dimension*, J. Math. Anal. Appl., 75 (1980), pp. 306–329.
[6] D. M. Eidus, *Inequalities for Green's function*, Mat. Sb., 87 (1958), pp. 455–470 (in Russian).
[7] Y. Guo, *Singular solutions of Vlasov-Maxwell boundary problems in one dimension*, Arch. Rational Mech. Anal., 131 (1995), pp. 241–304.
[8] Y. Guo, *Regularity for the Vlasov equations in a half space*, Indiana Univ. Math. J., 43 (1994), pp. 255–320.
[9] R. T. Glassey, *The Cauchy Problem in Kinetic Theory*, SIAM, Philadelphia, 1996.
[10] W. Greenberg, C. Van De Mee, and V. Protopopescu, *Boundary Value Problems in Abstract Kinetic Theory*, Oper. Theory Adv. Appl. 23, Birkhäuser Verlag, Basel, 1987.
[11] C. Greengard and P.-A. Raviart, *A boundary-value problem for the stationary Vlasov-Poisson equations: The plane diode*, Comm. Pure Appl. Math., 43 (1990), pp. 473–507.
[12] R. Glassey and W. Strauss, *Singularity formation in a collisionless plasma could occur only at high velocities*, Arch. Rational Mech. Anal., 92 (1986), pp. 59–90.
[13] R. Glassey and W. Strauss, *Absence of shocks in an initially dilute collisionless plasma*, Comm. Math. Phys., 113 (1987), pp. 191–208.
[14] R. Glassey and J. Schaeffer, *On symmetric solutions of the relativistic Vlasov-Poisson system*, Comm. Math. Phys., 101 (1985), pp. 459–473.
[15] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Grundlehren Math. Wiss. 224, Springer-Verlag, New York, 1977.
[16] E. Horst, *On the asymptotic growth of the solutions of the Vlasov-Poisson system*, Math. Methods Appl. Sci., 16 (1993), pp. 75–86.
[17] T. Kato, *On the classical solutions of the two-dimensional non-stationary Euler equation*, Arch. Rational Mech. Anal., 25 (1967), pp. 188–200.
[18] P. L. Lions and B. Perthame, *Régularité des solutions du système de Vlasov-Poisson en dimension* 3, C. R. Acad. Sci. Paris Sér. I Math., 311 (1990), pp. 205–210.
[19] S. Mischler, *On the initial boundary value problem for the Vlasov-Poisson-Boltzmann system*, Comm. Math. Phys., 210 (2000), pp. 447–466.
[20] N. Abdallah, *Weak solutions of the initial-boundary value problem for the Vlasov-Poisson system*, Math. Methods Appl. Sci., 17 (1994), pp. 451–476.
[21] F. Poupaud, *Boundary value problems for the stationary Vlasov-Maxwell system*, Forum Math., 4 (1992), pp. 499–527.
[22] K. Pfaffelmoser, *Global classical solutions of the Vlasov-Poisson system in three dimensions for general initial data*, J. Differential Equations, 95 (1992), pp. 281–303.
[23] J. Schaeffer, *Global existence of smooth solutions to the Vlasov-Poisson system in three dimensions*, Comm. Partial Differential Equations, 16 (1991), pp. 1313–1335.
[24] J. Schaeffer, *Global existence for the Poisson-Vlasov system with nearly symmetric data*, J. Differential Equations, 69 (1987), pp. 111–148.
[25] J. Weckler, *On the initial-boundary-value problem for the Vlasov-Poisson system: Existence of weak solutions and stability*, Arch. Rational Mech. Anal., 130 (1995), pp. 145–161.

SIAM J. MATH. ANAL.
Vol. 36, No. 1, pp. 172–185

UNIQUENESS OF LIMIT SOLUTIONS TO A FREE BOUNDARY PROBLEM FROM COMBUSTION*

J. FERNÁNDEZ BONDER† AND N. WOLANSKI†

Abstract. We investigate the uniqueness of limit solutions for a free boundary problem in heat propagation that appears as a limit of a parabolic system that arises in flame propagation.

Key words. free boundary problem, combustion, heat equation, uniqueness, classical solution, limit solution

AMS subject classifications. 35K05, 35K60, 80A25

DOI. 10.1137/S0036141002412938

1. Introduction.

In this paper we consider the following problem arising in combustion theory:

$$\begin{cases} \Delta u^\varepsilon - u_t^\varepsilon &= Y^\varepsilon f_\varepsilon(u^\varepsilon) \quad \text{in } \mathcal{D}, \\ \Delta Y^\varepsilon - Y_t^\varepsilon &= Y^\varepsilon f_\varepsilon(u^\varepsilon) \quad \text{in } \mathcal{D}, \end{cases} \tag{1.1}$$

where $\mathcal{D} \subset \mathbb{R}^{N+1}$.

This model appears in combustion theory in the analysis of the propagation of curved flames. It is derived in the framework of the theory of equidiffusional premixed flames analyzed in the relevant limit of high activation energy for Lewis number 1. In this application, Y^ε represents the fraction of some reactant (and hence it is assumed to be nonnegative), and u^ε is minus the temperature (more precisely, $u^\varepsilon = \lambda(T_f - T^\varepsilon)$, where T_f is the flame temperature and λ is a normalization factor). Observe that the term $Y^\varepsilon f_\varepsilon(u^\varepsilon)$ acts as an absorption term in (1.1). Since $T^\varepsilon = T_f - (u^\varepsilon/\lambda)$, it is in fact a reaction term for the temperature. In the flame model, such a term represents the effect of the exothermic chemical reaction and f has, accordingly, a number of properties: it is a nonnegative Lipschitz continuous function which is positive in an interval $(-\infty, \varepsilon)$ and vanishes otherwise (i.e., reaction occurs only when $T > T_f - \frac{\varepsilon}{\lambda}$). The parameter ε is essentially the inverse of the activation energy of the chemical reaction. For the sake of simplicity we will assume that $f_\varepsilon(s) = \frac{1}{\varepsilon^2} f(\frac{s}{\varepsilon})$, where f is a Lipschitz continuous function with $f(s) > 0$ if $s < 1$ and $f(s) = 0$ if $s \geq 1$.

For the derivation of the model, we cite [1].

Here we are interested in high activation energy limits (i.e., $\varepsilon \to 0$). These limits are currently the subject of active investigation, especially in the case $u^\varepsilon = Y^\varepsilon$. This is a natural assumption in the case of traveling waves.

In a previous paper [5] we have studied this problem in the case in which the initial values for u^ε and Y^ε, both converging to the same function u_0, satisfy the

*Received by the editors August 13, 2002; accepted for publication (in revised form) May 23, 2003; published electronically June 22, 2004. This work was partially supported by grant BID1201/OC-AR PICT 03-00000-05009 and CONICET grant PIP0660/98.

http://www.siam.org/journals/sima/36-1/41293.html

†Departamento de Matemática, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, (1428), Buenos Aires, Argentina (jfbonder@dm.uba.ar, wolanski@dm.uba.ar). The second author is a member of CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas of Argentina).

condition

$$\frac{Y_0^\varepsilon(x) - u_0^\varepsilon(x)}{\varepsilon} \to w_0(x) \quad \text{uniformly in } \mathbb{R}^N \tag{1.2}$$

with $w_0 > -1$.

Problem (1.1) reduces to a single equation, namely,

$$\Delta u^\varepsilon - u_t^\varepsilon = (u^\varepsilon + w^\varepsilon) f_\varepsilon(u^\varepsilon), \tag{P_ε}$$

where the function $w^\varepsilon(x,t)$ is the solution of the heat equation with initial datum $Y_0^\varepsilon(x) - u_0^\varepsilon(x)$. Observe that $u^\varepsilon + w^\varepsilon = Y^\varepsilon \geq 0$.

By (1.2) there exists the limit

$$\lim_{\varepsilon \to 0} \frac{w^\varepsilon(x,t)}{\varepsilon} = w_0(x,t), \tag{1.3}$$

and $w_0(x,t)$ is the solution of the heat equation with initial datum $w_0(x)$.

In this way, at least formally, the reaction term converges to a delta function, and a free boundary problem appears. In fact, we have proved in [5] that every sequence of uniformly bounded solutions to (1.1), $\{u^{\varepsilon_n}\}$, with $\varepsilon_n \to 0$ has a subsequence $\{u^{\varepsilon_{n_k}}\}$ converging to a limit function $u \geq 0$ which is a solution of the following free boundary problem:

$$\text{(P)} \qquad \begin{cases} \Delta u - u_t = 0 & \text{in } \{u > 0\}, \\ |\nabla u^+| = \sqrt{2M(x,t)} & \text{on } \partial\{u > 0\}, \end{cases}$$

where $M(x,t) = \int_{-w_0(x,t)}^1 \big(s + w_0(x,t)\big) f(s) ds$.

We see that the free boundary condition strongly depends on the approximation u_0^ε, Y_0^ε of the initial datum u_0. In particular, the limit function u is different for different approximations of the initial datum u_0.

It is therefore natural to wonder whether the only condition that determines the limit function u is condition (1.2).

The purpose of this paper is to prove that this is indeed the case, at least under some monotonicity assumption on the initial value u_0. This monotonicity assumption is similar to that used to prove uniqueness of the limit for the case $u^\varepsilon = Y^\varepsilon$ in [9].

In fact, we follow here some of the ideas of [9] which are based on the fact that any limit function is a supersolution to (P). This is still true in our case. Unfortunately, the simple construction in [9] of supersolutions of (P_ε) that approximate a strict classical supersolution of (P), when $w^\varepsilon = 0$, does not work in the general case unless one asks for a lot of complementary conditions on the reaction function f.

Therefore, we follow here the construction done in [7]. The proof that this construction works is based on blow up of the constructed functions. This technique was already seen to work very well for (P_ε), under condition (1.2), in [5].

Our result can be summarized as saying that *under suitable assumptions on the domain and on the initial datum u_0, there exists at most one limit solution to the free boundary problem* (P) *with nonvanishing gradient near its free boundary, as long as the approximate initial data, converging uniformly to u_0 with supports that converge to the support of u_0, satisfy* (1.2).

Moreover, under the same geometric assumptions, *if there exists a classical solution to* (P)*, this is the only limit of solutions to* (P_ε) *with initial data satisfying the conditions above. In particular, it is the only classical solution to* (P).

As already stated, in the case $u^\varepsilon = Y^\varepsilon$, uniqueness results for limit solutions under geometric hypotheses similar to the ones made here can be found in [9]. Also, in [7] the authors study the uniqueness and agreement between different concepts of solutions of problem (P) (again in the case $u^\varepsilon = Y^\varepsilon$) under the assumption of the existence of a classical solution and under different geometric assumptions. See also [8] for a similar result in the two-phase case.

Notation. Throughout the paper N will denote the spatial dimension. In addition, the following notation will be used:

For any $x_0 \in \mathbb{R}^N$, $t_0 \in \mathbb{R}$, and $\tau > 0$, $B_\tau(x_0) := \{x \in \mathbb{R}^N / |x - x_0| < \tau\}$ and $B_\tau(x_0, t_0) := \{(x,t) \in \mathbb{R}^{N+1} / |x - x_0|^2 + |t - t_0|^2 < \tau^2\}$.

When necessary, we will denote points in $\mathbb{R}^N$ by $x = (x_1, x')$, with $x' \in \mathbb{R}^{N-1}$. Given a function v, we will denote $v^+ = \max(v, 0)$.

The symbols Δ and ∇ will denote the corresponding operators in the space variables; the symbol ∂_p applied to a domain will denote parabolic boundary.

Finally, we will say that u is supercaloric if $\Delta u - u_t \le 0$, and u is subcaloric if $\Delta u - u_t \ge 0$.

Outline of the paper. An outline of the contents is as follows. In section 2 we give precise definitions of classical sub- and supersolutions and prove a comparison result for problem (P) (Lemma 2.1). In section 3 we state some auxiliary results. In section 4 we prove that a strict classical supersolution to problem (P) is the uniform limit of a family of supersolutions to problem (P_ε) (Theorem 4.1), and as a consequence we obtain the boundedness of the support for limit solutions in the geometry under consideration (Proposition 4.1). Finally, in section 5 we prove our main result (Theorem 5.1). We discuss in a final section (section 6) the results proved in the paper as well as other possible geometries that can be considered.

2. Preliminaries. Following [9] we will define what we will understand by a classical supersolution of problem (P). Note that the meaning of *classical* here differs from the usual one since we are not assuming that the function is C^1 up to the free boundary or that the free boundary is C^1.

DEFINITION 2.1. *A continuous nonnegative function u in $\overline{Q}_T = \mathbb{R}^N \times [0,T]$, $T > 0$, is called a classical supersolution of* (P) *if $u \in C^1(\{u > 0\})$ and*

(i) $\Delta u - u_t \le 0$ *in* $\Omega = \{u > 0\}$;

(ii) $\limsup_{\Omega \ni (y,s) \to (x,t)} |\nabla u(y,s)| \le \sqrt{2M(x,t)}$ *for every* $(x,t) \in \partial\Omega \cap Q_T$;

(iii) $u(\cdot, 0) \ge u_0$.

Respectively, u is a classical subsolution of (P) *if conditions* (i), (ii), *and* (iii) *are satisfied with reversed inequalities and* $\liminf$ *instead of* $\limsup$ *in* (ii).

A function u is a classical solution of (P) *if it is both a classical subsolution and a classical supersolution of* (P).

Next, a classical supersolution u of (P) *is a strict classical supersolution of* (P) *if there is a $\delta > 0$ such that the stronger inequalities*

(ii') $\limsup_{\Omega \ni (y,s) \to (x,t)} |\nabla u(y,s)| \le \sqrt{2M(x,t) - \delta}$ *for every* $(x,t) \in \partial\Omega \cap Q_T$,

(iii') $u(\cdot, 0) \ge u_0 + \delta$ *on* $\Omega_0 = \{u_0 > 0\}$

hold. Analogously, a strict classical subsolution is defined.

As a consequence of the results in [5], one can check that every limit solution $u = \lim_{j \to \infty} u^{\varepsilon_j}$ of (P) is a classical supersolution in the sense of Definition 2.1.

PROPOSITION 2.1. *Let u^{ε_j} be solutions to* (P_{ε_j}), *with w^{ε_j} satisfying* (1.3) *and $w_0 > -1$, such that $u^{\varepsilon_j} \to u$ uniformly on compact sets and $\varepsilon_j \to 0$. Assume that the initial datum u_0 is Lipschitz continuous and that the approximations of the initial*

datum verify $|u_0^\varepsilon(x)|, |\nabla u_0^\varepsilon(x)| \le C$, *and* $u_0^\varepsilon \in C^1(\overline{\{u_0^\varepsilon > 0\}})$. *Then* u *is a classical supersolution of* (P).

Proof. We have to verify conditions (i)–(iii) of Definition 2.1.

From our assumptions on the initial datum u_0, by Proposition 5.2.1 of [6], we have that $u^\varepsilon \to u$ uniformly on compact sets of $\overline{Q_T}$ so that u is continuous up to $t = 0$ and (iii) holds.

Now (i) is proved in [5].

Finally, (ii) is a straightforward modification of Theorem 6.1 of [2] using Lemmas 2.1, 2.2, and 2.3 of [5] instead of Lemma 3.2 and Propositions 5.2 and 5.3 of [2], respectively. □

Let us suppose that the initial datum u_0 of problem (P) is starshaped with respect to a point x_0, which we always assume to be 0, in the following sense: For every $\lambda \in (0,1)$ and $x \in \mathbb{R}^N$,

$$u_0(\lambda x) \ge u_0(x), \qquad \lambda\Omega_0 \subset\subset \Omega_0, \tag{2.1}$$

where $\Omega_0 = \{u_0 > 0\}$.

Also, assume that

$$w_0(\lambda x, 0) \le w_0(x, 0) \quad \text{if } x \in \mathbb{R}^N,\ 0 < \lambda < 1. \tag{2.2}$$

Let u be a classical supersolution of (P). Let λ and λ' be two real numbers with $0 < \lambda < \lambda' < 1$. Define

$$u_\lambda(x,t) = \frac{1}{\lambda'} u(\lambda x, \lambda^2 t) \tag{2.3}$$

in Q_{T/λ^2}. The rescaling is taken so that u_λ is a supersolution of the heat equation in

$$\Omega_\lambda = \{(x,t) :\ (\lambda x, \lambda^2 t) \in \Omega\}. \tag{2.4}$$

Moreover, the fact that $0 < \lambda < \lambda' < 1$ makes u_λ a strict classical supersolution of (P).

In fact, let us first see that

$$M(\lambda x, \lambda^2 t) \le M(x,t) \qquad \text{if } 0 < \lambda < 1.$$

This is a consequence of the fact that the function

$$a \longrightarrow \int_{-a}^{1} (s + a) f(s)\, ds$$

is nondecreasing and

$$w_0(\lambda x, \lambda^2 t) \le w_0(x,t) \qquad \text{if } 0 < \lambda < 1. \tag{2.5}$$

In fact, the function $w_\lambda(x,t) = w_0(\lambda x, \lambda^2 t)$ is caloric, and $w_\lambda(x,0) \le w_0(x,0)$ if $0 < \lambda < 1$ by hypothesis. Thus, by the comparison principle, $w_\lambda(x,t) \le w_0(x,t)$ in $\mathbb{R}^N \times (0,T)$.

Now let $(x_0, t_0) \in \partial\{u_\lambda > 0\}$. Then

$$\begin{aligned}\limsup_{\Omega_\lambda \ni (x,t) \to (x_0,t_0)} |\nabla u_\lambda(x,t)| &= \limsup_{\Omega \ni (\lambda x, \lambda^2 t) \to (\lambda x_0, \lambda^2 t_0)} \left| \frac{\lambda}{\lambda'} \nabla u(\lambda x, \lambda^2 t) \right| \\ &\le \frac{\lambda}{\lambda'} \sqrt{2M(\lambda x_0, \lambda^2 t_0)} \le \sqrt{2M(x_0,t_0)} - \left(1 - \frac{\lambda}{\lambda'}\right)\sqrt{2M_0},\end{aligned}$$

where $0 < M_0 < M(x,t)$ in $\mathbb{R}^N \times (0,T)$.

On the other hand, since $\lambda\Omega_0 \subset\subset \Omega_0$, there holds that

$$u_0(\lambda x) \geq \gamma > 0 \qquad \text{if } x \in \Omega_0.$$

Thus, for $x \in \Omega_0$,

$$\begin{aligned} u_\lambda(x,0) = \frac{1}{\lambda'} u_0(\lambda x) &= u_0(\lambda x) + \Big(\frac{1}{\lambda'} - 1\Big) u_0(\lambda x) \\ &\geq u_0(x) + \Big(\frac{1}{\lambda'} - 1\Big)\gamma. \end{aligned}$$

The following comparison lemma for problem (P) can be proved as Lemma 2.4 in [9]. We omit the proof.

LEMMA 2.1. *Let u_0 satisfy* (2.1) *and w_0 satisfy* (2.2). *Then every classical subsolution of* (P) *with bounded support is smaller than every classical supersolution of* (P)*; i.e., if u' is a classical subsolution such that Ω' is bounded and u is a classical supersolution, then*

$$\Omega' \subset \Omega \quad \textit{and} \quad u' \leq u,$$

where $\Omega' = \{u' > 0\}$ and $\Omega = \{u > 0\}$.

3. Auxiliary results. This section contains results on the following problem:

$$\Delta u - u_t = (u + \omega_0) f(u), \tag{P_0}$$

where the function f is as in section 1 and ω_0 is a constant, $\omega_0 > -1$. The results will be used in the next sections where (P_0) appears as a blow-up limit. The proofs are very similar to those of Lemmas 4.1, 4.3, and 4.4 in [7]. We leave the details to the reader.

LEMMA 3.1. *Let $a, b \geq 0$, and let ψ be the classical solution to*

$$\begin{aligned} &\psi_{ss} = (\psi + \omega_0) f(\psi) \quad \textit{for } s > 0, \\ &\psi(0) = a, \quad \psi_s(0) = -\sqrt{2b}. \end{aligned} \tag{3.1}$$

Let $B(\tau) = \int_{-\omega_0}^{\tau} (\rho + \omega_0) f(\rho)\, d\rho$.

(3.2) *If $b = 0$ and $a \in \{-\omega_0\} \cup [1, +\infty)$, then $\psi \equiv a$.*

(3.3) *If $b = 0$ and $a \in (-\omega_0, 1)$, then $\lim_{s\to+\infty} \psi(s) = +\infty$.*

(3.4) *If $b \in \big(0,\, B(a)\big)$, then $\lim_{s\to+\infty} \psi(s) = +\infty$.*

(3.5) *If $0 < b = B(a)$, then $\psi_s < 0$ and $\lim_{s\to+\infty} \psi(s) = -\omega_0$.*

(3.6) *If $b \in \big(B(a),\, +\infty\big)$, then $\psi_s < 0$ and $\lim_{s\to+\infty} \psi(s) = -\infty$.*

LEMMA 3.2. *Let $B(\tau)$ be as in the previous lemma, let $\mathcal{R}_\gamma = \big\{(x,t) \in \mathbb{R}^{N+1} / x_1 > 0\,,\, -\infty < t \leq \gamma\big\}$, $0 \leq \theta < 1 + \omega_0$, and let $U \in C^{2+\alpha, 1+\frac{\alpha}{2}}(\overline{\mathcal{R}_\gamma})$ be such that*

$$\begin{aligned} \Delta U - U_t = (U + \omega_0) f(U) \quad &\textit{in } \mathcal{R}_\gamma, \\ U = 1 - \theta \quad &\textit{on } \{x_1 = 0\}, \\ -\omega_0 \leq U \leq 1 - \theta \quad &\textit{in } \overline{\mathcal{R}_\gamma}. \end{aligned}$$

(1) *If $\theta = 0$, then $|\nabla U| \leq \sqrt{2B(1)}$ on $\{x_1 = 0\}$.*

(2) *If $0 < \theta < 1 + \omega_0$ and $0 < \sigma < B(1)$ are such that $\int_{-\omega_0}^{1-\theta} (\rho + \omega_0) f(\rho)\, d\rho = B(1) - \sigma$, then $|\nabla U| = \sqrt{2(B(1) - \sigma)}$ on $\{x_1 = 0\}$.*

Finally, we state a compactness result.

LEMMA 3.3. *Let ε_j, γ_{ε_j}, and τ_{ε_j} be sequences such that $\varepsilon_j > 0$, $\varepsilon_j \to 0$, $\gamma_{\varepsilon_j} > 0$, $\gamma_{\varepsilon_j} \to \gamma$, with $0 \le \gamma \le +\infty$, $\tau_{\varepsilon_j} > 0$, $\tau_{\varepsilon_j} \to \tau$ with $0 \le \tau \le +\infty$, and such that $\tau < +\infty$ implies that $\gamma = +\infty$. Assume that $w^{\varepsilon_j}/\varepsilon_j$ converge to w_0 uniformly in compact sets of $\mathbb{R}^N \times [0,T]$. Let $\rho > 0$ and*

$$\mathcal{A}_{\varepsilon_j} = \left\{ (x,t) \ / \ |x| < \frac{\rho}{\varepsilon_j} \, , \ -\min\left(\tau_{\varepsilon_j}, \frac{\rho^2}{\varepsilon_j^2}\right) < t < \min\left(\gamma_{\varepsilon_j}, \frac{\rho^2}{\varepsilon_j^2}\right) \right\}.$$

Let $(x_0,t_0) \in \mathbb{R}^N \times [0,T]$. Assume that $0 \le \theta < 1 + w_0(x_0,t_0)$, and let $\bar{u}^{\varepsilon_j}$ be weak solutions to

$$\begin{aligned}
&\Delta \bar{u}^{\varepsilon_j} - \bar{u}_t^{\varepsilon_j} = \left(\bar{u}^{\varepsilon_j} + \frac{w^{\varepsilon_j}(\varepsilon_j x + x_{\varepsilon_j}, \varepsilon_j^2 t + t_{\varepsilon_j})}{\varepsilon_j} \right) f(\bar{u}^{\varepsilon_j}) \text{ in } \{x_1 > \bar{h}_{\varepsilon_j}(x',t)\} \cap \mathcal{A}_{\varepsilon_j}, \\
&\bar{u}^{\varepsilon_j} = 1 - \theta \quad \text{on } \{x_1 = \bar{h}_{\varepsilon_j}(x',t)\} \cap \mathcal{A}_{\varepsilon_j}, \\
&-\frac{w^{\varepsilon_j}(\varepsilon_j x + x_{\varepsilon_j}, \varepsilon_j^2 t + t_{\varepsilon_j})}{\varepsilon_j} \le \bar{u}^{\varepsilon_j} \le 1 - \theta \quad \text{in } \{x_1 \ge \bar{h}_{\varepsilon_j}(x',t)\} \cap \overline{\mathcal{A}_{\varepsilon_j}},
\end{aligned}$$

where $(x_{\varepsilon_j}, t_{\varepsilon_j}) \to (x_0,t_0)$, with $\bar{u}^{\varepsilon_j} \in C(\{x_1 \ge \bar{h}_{\varepsilon_j}(x',t)\} \cap \overline{\mathcal{A}_{\varepsilon_j}})$, and $\nabla \bar{u}^{\varepsilon_j} \in L^2$. Here $\bar{h}_{\varepsilon_j}$ are continuous functions such that $\bar{h}_{\varepsilon_j}(0,0) = 0$ with $\bar{h}_{\varepsilon_j} \to 0$ uniformly on compact subsets of $\mathbb{R}^{N-1} \times (-\tau, \gamma)$. Moreover, we assume that $\|\bar{h}_{\varepsilon_j}\|_{C^1(K)} + \|\nabla_{x'} \bar{h}_{\varepsilon_j}\|_{C^{\alpha, \frac{\alpha}{2}}(K)}$ are uniformly bounded for every compact set $K \subset \mathbb{R}^{N-1} \times (-\tau, \gamma)$.

Then there exists a function $\bar{u}$ such that, for a subsequence,

$$\begin{aligned}
&\bar{u} \in C^{2+\alpha, 1+\frac{\alpha}{2}}\big(\{x_1 \ge 0, \, \gamma > t > -\tau\}\big), \\
&\bar{u}^{\varepsilon_j} \to \bar{u} \quad \text{uniformly on compact subsets of } \{x_1 > 0, \, \gamma > t > -\tau\}, \\
&\begin{aligned}
&\Delta \bar{u} - \bar{u}_t = (\bar{u} + w_0(x_0,t_0)) f(\bar{u}) && \text{in } \{x_1 > 0, \, \gamma > t > -\tau\}, \\
&\bar{u} = 1 - \theta && \text{on } \{x_1 = 0, \, \gamma > t > -\tau\}, \\
&-w_0(x_0,t_0) \le \bar{u} \le 1 - \theta && \text{in } \{x_1 \ge 0, \, \gamma > t > -\tau\}.
\end{aligned}
\end{aligned}$$

If $\gamma < +\infty$, we require, in addition, that

$$\|\bar{h}_{\varepsilon_j}(x', t + \gamma_{\varepsilon_j} - \gamma)\|_{C^1(K)} + \|\nabla_{x'} \bar{h}_{\varepsilon_j}(x', t + \gamma_{\varepsilon_j} - \gamma)\|_{C^{\alpha, \frac{\alpha}{2}}(K)}$$

be uniformly bounded for every compact set $K \subset \mathbb{R}^{N-1} \times (-\infty, \gamma]$. We deduce that

$$\bar{u} \in C^{2+\alpha, 1+\frac{\alpha}{2}}\big(\{x_1 \ge 0, \, t \le \gamma\}\big).$$

If $\tau < +\infty$, we let

$$\mathcal{B}_{\varepsilon_j} = \left\{ x \ / \ |x| < \frac{\rho}{\varepsilon_j} \, , \ x_1 > \bar{h}_{\varepsilon_j}(x', -\tau_{\varepsilon_j}) \right\},$$

and we require, in addition, that for every $R > 0$,

$$\|\bar{u}^{\varepsilon_j}(x, -\tau_{\varepsilon_j})\|_{C^\alpha\left(\overline{\mathcal{B}}_{\varepsilon_j} \cap \overline{B}_R(0)\right)} \le C_R,$$

and that there exists $r > 0$ such that

$$\|\bar{u}^{\varepsilon_j}(x, -\tau_{\varepsilon_j})\|_{C^{1+\alpha}\left(\overline{\mathcal{B}}_{\varepsilon_j} \cap \overline{B}_r(0)\right)} \le C_r.$$

Moreover, we assume that $\|\bar{h}_{\varepsilon_j}(x', t-\tau_{\varepsilon_j}+\tau)\|_{C^1(K)} + \|\nabla_{x'}\bar{h}_{\varepsilon_j}(x', t-\tau_{\varepsilon_j}+\tau)\|_{C^{\alpha,\frac{\alpha}{2}}(K)}$ *are uniformly bounded for every compact set* $K \subset \mathbb{R}^{N-1}\times[-\tau,+\infty)$.

Then there holds that

$$\bar{u} \in C^{\alpha,\frac{\alpha}{2}}\big(\{x_1 \geq 0,\, t \geq -\tau\}\big),\ \nabla\bar{u} \in C\big(\{0 \leq x_1 < r,\, t \geq -\tau\}\big),$$
$$\bar{u}^{\varepsilon_j}(x, -\tau_{\varepsilon_j}) \to \bar{u}(x,-\tau) \quad \textit{uniformly on compact subsets of } \{x_1 > 0\}.$$

In any case (τ, γ *be infinite or finite*)

$$|\nabla\bar{u}^{\varepsilon_j}(0,0)| \to |\nabla\bar{u}(0,0)|.$$

4. Approximation result. In this section we prove that, under certain assumptions, a classical supersolution to problem (P) is the uniform limit of a family of supersolutions to problem (P_ε) (Theorem 4.1), and we prove an analogous result for subsolutions (Theorem 4.2). Also, we prove that for compactly supported initial data, limit solutions have bounded support (Proposition 4.1).

The following construction follows the lines of Theorem 5.2 in [7]. In our case we have to be more careful with the construction of the initial data.

THEOREM 4.1. *Let* $\widetilde{u}$ *be a classical supersolution to* (P) *in* Q_T *with* $\widetilde{u} \in C^1(\overline{\{\widetilde{u}>0\}})$ *and such that* $\{\widetilde{u} > 0\}$ *is bounded. Assume, in addition, that there exist* $\delta_0, s_0 > 0$ *such that*

$$|\nabla\widetilde{u}^+| \leq \sqrt{2M(x,t)-\delta_0} \quad \textit{on } Q \cap \partial\{\widetilde{u}>0\},$$
$$|\nabla\widetilde{u}| > \delta_0 \quad \textit{in } Q \cap \{0 < \widetilde{u} < s_0\}.$$

Let w^ε *be a solution of the heat equation in* $\mathbb{R}^N\times(0,T)$ *such that* $\frac{w^\varepsilon(x,t)}{\varepsilon} \to w_0(x,t)$ *uniformly in* $\mathbb{R}^N\times[0,T]$ *with* $w_0 \in C(\mathbb{R}^N\times[0,T])$ *and* $w_0 \geq -1+\delta_1$ *for a certain positive constant* δ_1.

Then there exists a family $u^\varepsilon \in C(\overline{Q_T})$, *with* $\nabla u^\varepsilon \in L^2_{\mathrm{loc}}(\overline{Q_T})$, *of weak supersolutions to* (P_ε) *in* Q_T *such that, as* $\varepsilon \to 0$, $u^\varepsilon \to \widetilde{u}$ *uniformly in* $\overline{Q_T}$.

Proof.

Step 1. Construction of the family u^ε. Let $0 < \theta < \delta_1$ be such that

$$\int_{1-\theta}^{1}(s+W)f(s)\,ds = \frac{\delta_0}{8},$$

where W is a suitable uniform bound of $\|w^\varepsilon/\varepsilon\|_{L^\infty(\{\widetilde{u}>0\})}$. For every $\varepsilon > 0$ small, we define the domain $D^\varepsilon = \{\widetilde{u} < (1-\theta)\varepsilon\} \subset Q_T$.

Let z^ε be the bounded solution to

$$\Delta z^\varepsilon - z^\varepsilon_t = (z^\varepsilon + w^\varepsilon)f_\varepsilon(z^\varepsilon) \quad \text{in } D^\varepsilon,$$

with boundary data

$$z^\varepsilon(x,t) = \begin{cases} (1-\theta)\varepsilon & \text{on } \partial D^\varepsilon \cap t > 0, \\ z_0^\varepsilon(x) & \text{in } D^\varepsilon \cap \{t=0\}. \end{cases}$$

In order to give the initial data z_0^ε, we let $\psi^\varepsilon(s,x)$ be the solution to (3.1) with

$$a = 1-\theta, \quad b = \int_{-w^\varepsilon(x,0)/\varepsilon}^{1-\theta}\left(s + \frac{w^\varepsilon(x,0)}{\varepsilon}\right)f(s)\,ds, \quad \omega_0 = \frac{w^\varepsilon(x,0)}{\varepsilon}.$$

Assume first that $|\nabla\widetilde{u}|$ is smooth. Then, by extending $|\nabla\tilde{u}(x,0)|$ to the whole $\mathbb{R}^N$ as a positive function, we let

$$\varphi^\varepsilon(\xi,x)=\psi^\varepsilon\left(\frac{1-\theta-\xi}{|\nabla\widetilde{u}(x,0)|},x\right),$$

and we define

$$z_0^\varepsilon(x)=\varepsilon\varphi^\varepsilon\left(\frac{1}{\varepsilon}\widetilde{u}(x,0),x\right).$$

If $\widetilde{u}$ is not regular enough, we can replace $|\nabla\widetilde{u}(x,0)|$ by a smooth approximation $F_\varepsilon(x)$ so that the initial datum z_0^ε is $C^{1+\alpha}$. We leave the details to the reader.

Finally, we define the family u^ε as follows:

$$u^\varepsilon=\begin{cases}\widetilde{u} & \text{in } \{\widetilde{u}\geq(1-\theta)\varepsilon\},\\ z^\varepsilon & \text{in } \overline{D^\varepsilon}.\end{cases}$$

Step 2. Passage to the limit. If $(x,0)\in\overline{D^\varepsilon}$, then we have $0\leq\frac{1}{\varepsilon}\widetilde{u}(x,0)\leq 1-\theta$. Since, from Lemma 3.1, we know that $-w^\varepsilon(x,0)/\varepsilon\leq\psi^\varepsilon(s,x)\leq 1-\theta$ for $s\geq 0$, it follows that $-w^\varepsilon(x,0)\leq z^\varepsilon(x,0)\leq(1-\theta)\varepsilon$. Since $f_\varepsilon(s)\geq 0$, constant functions larger than $-w^\varepsilon(x,t)$ are supersolutions to (P_ε). Therefore, $(1-\theta)\varepsilon$ is a supersolution if $\varepsilon<\varepsilon_1$, for some $\varepsilon_1>0$, and we may apply the comparison principle for bounded super- and subsolutions of (P_ε) to conclude that $-w^\varepsilon\leq z^\varepsilon\leq(1-\theta)\varepsilon$.

Hence,

$$\sup_{\overline{Q_T}}|u^\varepsilon-\widetilde{u}|=\sup_{D^\varepsilon}|z^\varepsilon-\widetilde{u}|\leq C\varepsilon,$$

and therefore the convergence of the family u^ε follows.

Step 3. Let us show that there exists $\varepsilon_0>0$ such that the functions u^ε are supersolutions to (P_ε) for $\varepsilon<\varepsilon_0$.

If $u^\varepsilon>(1-\theta)\varepsilon$, then $u^\varepsilon=\widetilde{u}$, which by hypothesis is supercaloric. Since $f_\varepsilon(s)\geq 0$ and $(1-\theta)\varepsilon\geq-w^\varepsilon$ if $\varepsilon<\varepsilon_1$, it follows that u^ε are supersolutions to (P_ε) here.

If $u^\varepsilon<(1-\theta)\varepsilon$, then we are in D^ε, and therefore, by construction, u^ε are solutions to (P_ε).

That is, the u^ε's are continuous functions, and they are piecewise supersolutions to (P_ε). In order to see that u^ε are globally supersolutions to (P_ε), it suffices to see that the jumps of the gradients (which occur at smooth surfaces) have the right sign.

To this effect, we will show that there exists $\varepsilon_0>0$ such that

$$|\nabla u^\varepsilon|\geq\sqrt{2M(x,t)-\delta_0/2}\quad\text{on }\partial\{\widetilde{u}<(1-\theta)\varepsilon\}\ \text{ for }\varepsilon<\varepsilon_0. \tag{4.1}$$

Assume that (4.1) does not hold. Then, for every $j\in\mathbb{N}$, there exist $\varepsilon_j>0$ and $(x_{\varepsilon_j},t_{\varepsilon_j})\in Q$, with

$$\varepsilon_j\to 0\quad\text{and}\quad(x_{\varepsilon_j},t_{\varepsilon_j})\to(x_0,t_0)\in\partial\{\widetilde{u}>0\},$$

such that

$$u^{\varepsilon_j}(x_{\varepsilon_j},t_{\varepsilon_j})=(1-\theta)\varepsilon_j\quad\text{and}\quad|\nabla u^{\varepsilon_j}(x_{\varepsilon_j},t_{\varepsilon_j})|<\sqrt{2M(x_{\varepsilon_j},t_{\varepsilon_j})-\delta_0/2}. \tag{4.2}$$

From now on we will drop the subscript j when referring to the sequences defined above, and $\varepsilon \to 0$ will mean $j \to \infty$.

We can assume (performing a rotation in the space variables if necessary) that there exists a family g_ε of smooth functions such that, in a neighborhood of $(x_\varepsilon, t_\varepsilon)$,

$$\begin{aligned} &\{u^\varepsilon = (1-\theta)\varepsilon\} = \{(x,t) \,/\, x_1 - x_{\varepsilon 1} = g_\varepsilon(x' - x_\varepsilon{}', t - t_\varepsilon)\}, \\ &\{u^\varepsilon < (1-\theta)\varepsilon\} = \{(x,t) \,/\, x_1 - x_{\varepsilon 1} > g_\varepsilon(x' - x_\varepsilon{}', t - t_\varepsilon)\}, \end{aligned} \tag{4.3}$$

where there holds that

$$g_\varepsilon(0,0) = 0, \quad |\nabla_{x'} g_\varepsilon(0,0)| \to 0, \quad \varepsilon \to 0.$$

We can assume that (4.3) holds in $\big(B_\rho(x_\varepsilon) \times (t_\varepsilon - \rho^2, t_\varepsilon + \rho^2)\big) \cap \{0 \le t \le T\}$ for some $\rho > 0$.

Let us now define

$$\bar{u}^\varepsilon(x,t) = \frac{1}{\varepsilon} u^\varepsilon(x_\varepsilon + \varepsilon x, t_\varepsilon + \varepsilon^2 t), \quad \bar{g}_\varepsilon(x',t) = \frac{1}{\varepsilon} g_\varepsilon(\varepsilon x', \varepsilon^2 t),$$

and let

$$\tau_\varepsilon = \frac{t_\varepsilon}{\varepsilon^2}, \quad \gamma_\varepsilon = \frac{T - t_\varepsilon}{\varepsilon^2}.$$

We have, for a subsequence,

$$\tau_\varepsilon \to \tau, \quad \gamma_\varepsilon \to \gamma,$$

where $0 \le \tau, \gamma \le +\infty$, and τ and γ cannot be both finite.

We now let

$$\mathcal{A}_\varepsilon = \left\{ (x,t) \;/\; |x| < \frac{\rho}{\varepsilon}, \; -\min\left(\tau_\varepsilon, \frac{\rho^2}{\varepsilon^2}\right) < t < \min\left(\gamma_\varepsilon, \frac{\rho^2}{\varepsilon^2}\right) \right\}.$$

Then the functions $\bar{u}^\varepsilon$ are weak solutions to

$$\begin{aligned} &\Delta \bar{u}^\varepsilon - \bar{u}^\varepsilon_t = \left(\bar{u}^\varepsilon + \frac{w^\varepsilon(x_\varepsilon + \varepsilon x, t_\varepsilon + \varepsilon^2 t)}{\varepsilon}\right) f(\bar{u}^\varepsilon) && \text{in } \{x_1 > \bar{g}_\varepsilon(x',t)\} \cap \mathcal{A}_\varepsilon, \\ &\bar{u}^\varepsilon = 1 - \theta && \text{on } \{x_1 = \bar{g}_\varepsilon(x',t)\} \cap \mathcal{A}_\varepsilon, \\ &-\frac{w^\varepsilon(x_\varepsilon + \varepsilon x, t_\varepsilon + \varepsilon^2 t)}{\varepsilon} \le \bar{u}^\varepsilon \le 1 - \theta && \text{in } \{x_1 \ge \bar{g}_\varepsilon(x',t)\} \cap \overline{\mathcal{A}_\varepsilon}. \end{aligned}$$

Note that we are under the hypotheses of Lemma 3.3. Then there exists a function $\bar{u}$ such that, for a subsequence,

$$\begin{aligned} &\bar{u} \in C^{2+\alpha, 1+\frac{\alpha}{2}}\big(\{x_1 \ge 0, \, -\tau < t < \gamma\}\big), \\ &\bar{u}^\varepsilon \to \bar{u} \quad \text{uniformly on compact subsets of } \{x_1 > 0, \, -\tau < t < \gamma\}, \\ &\Delta \bar{u} - \bar{u}_t = (\bar{u} + w_0(x_0,t_0)) f(\bar{u}) && \text{in } \{x_1 > 0, \, -\tau < t < \gamma\}, \\ &\bar{u} = 1 - \theta && \text{on } \{x_1 = 0, \, -\tau < t < \gamma\}, \\ &-w_0(x_0,t_0) \le \bar{u} \le 1 - \theta && \text{in } \{x_1 \ge 0, \, -\tau < t < \gamma\}. \end{aligned}$$

We will divide the remainder of the proof into two cases, depending on whether $\tau = +\infty$ or $\tau < +\infty$.

Case 1. Assume $\tau = +\infty$.

In this case, Lemma 3.3 also gives

$$|\nabla \bar{u}^\varepsilon(0,0)| \to |\nabla \bar{u}(0,0)|.$$

On the other hand, $\bar{u}$ satisfies the hypotheses of Lemma 3.2, and therefore

$$|\nabla \bar{u}| \geq \sqrt{2M(x_0,t_0) - \delta_0/4} \quad \text{on } \{x_1 = 0\},$$

which yields

$$|\nabla \bar{u}^\varepsilon(0,0)| \geq \sqrt{2M(x_0,t_0) - 3\delta_0/8}$$

for ε small. But this gives

$$|\nabla u^\varepsilon(x_\varepsilon,t_\varepsilon)| \geq \sqrt{2M(x_\varepsilon,t_\varepsilon) - \delta_0/2}$$

for ε small. This contradicts (4.2) and completes the proof in case $\tau = +\infty$.

Case 2. Assume $\tau < +\infty$. (In this case $\gamma = +\infty$.)

There holds that $\bar{u}^\varepsilon(x,-\tau_\varepsilon) = \frac{1}{\varepsilon} u^\varepsilon(x_\varepsilon + \varepsilon x, 0)$; then

$$\bar{u}^\varepsilon(x,-\tau_\varepsilon) = \varphi^\varepsilon \left(\frac{1}{\varepsilon} \widetilde{u}(x_\varepsilon + \varepsilon x, 0), x_\varepsilon + \varepsilon x \right). \tag{4.4}$$

Here we want to apply the result of Lemma 3.3 corresponding to $\tau < +\infty$. In fact, we can see that there exist $C, r > 0$ such that $\|\bar{u}^\varepsilon(\cdot,-\tau_\varepsilon)\|_{C^{1+\alpha}(\overline{B}_r(0))} \leq C$.

Now Lemma 3.3 gives, for a subsequence,

$$\bar{u} \in C^{\alpha,\frac{\alpha}{2}}\big(\{x_1 \geq 0,\, t \geq -\tau\}\big),$$
$$\bar{u}^\varepsilon(x,-\tau_\varepsilon) \to \bar{u}(x,-\tau) \quad \text{uniformly on compact subsets of } \{x_1 > 0\}.$$

Therefore, we get that (recall that in the case we are considering $t_0 = 0$)

$$\bar{u}(x,-\tau) = \bar{\varphi}\Big(1 - \theta - |\nabla \widetilde{u}^+(x_0,t_0)|\, x_1, x_0\Big),$$

where $\bar{\varphi}(s,x) = \psi(\frac{1-\theta-s}{|\nabla \widetilde{u}(x,0)|}, x)$ and $\psi(s,x)$ is the solution of (3.1) with

$$a = 1-\theta, \quad b = \int_{-w_0(x,0)}^{1-\theta} (s + w_0(x,0)) f(s)\, ds, \quad \omega_0 = w_0(x,0).$$

Thus,

$$\bar{u}(x,-\tau) = \psi(x_1,x_0).$$

Since the function $\psi(x_1,x_0)$ is a stationary solution to equation (P_0), bounded for $x_1 \geq 0$, and $\bar{u} = \psi$ on the parabolic boundary of the domain $\{x_1 > 0,\, t > -\tau\}$, we conclude that

$$\bar{u}(x,t) = \psi(x_1,x_0) \quad \text{in } \{x_1 \geq 0,\, t \geq -\tau\}.$$

It follows from Lemma 3.1 and the choice of θ that

$$\frac{1}{2}|\nabla \bar{u}(0,0)|^2 = \frac{1}{2}\big(\psi_s(0,x_0)\big)^2 = \int_{-w_0(x_0,t_0)}^{1-\theta} (s + w_0(x_0,t_0)) f(s)\, ds \geq M(x_0,t_0) - \frac{\delta_0}{8}.$$

That is,

$$|\nabla \bar{u}| \geq \sqrt{2M(x_0,t_0) - \delta_0/4} \quad \text{on } \{x_1 = 0, t \geq -\tau\}.$$

But Lemma 3.3 gives

$$|\nabla \bar{u}^\varepsilon(0,0)| \to |\nabla \bar{u}(0,0)|,$$

which yields

$$|\nabla \bar{u}^\varepsilon(0,0)| \geq \sqrt{2M(x_0,t_0) - 3\delta_0/8}$$

for ε small. Then

$$|\nabla u^\varepsilon(x_\varepsilon,t_\varepsilon)| \geq \sqrt{2M(x_\varepsilon,t_\varepsilon) - \delta_0/2}$$

for ε small. This contradicts (4.2) and completes the proof in case $\tau < +\infty$. □

Remark 4.1. Observe that from the construction of u^ε done in the previous proof it follows that

$$u^\varepsilon \equiv \widetilde{u} \qquad \text{in } \{\widetilde{u} > (1-\theta)\varepsilon\}.$$

THEOREM 4.2. *Let $\widetilde{u}$ be a classical subsolution to* (P) *in Q_T with $\widetilde{u} \in C^1(\overline{\{\widetilde{u} > 0\}})$ such that $\{\widetilde{u} > 0\}$ is bounded. Assume, in addition, that there exist $\delta_0 > 0$ such that*

$$|\nabla \widetilde{u}^+| \geq \sqrt{2M(x,t) + \delta_0} \quad \text{on } Q \cap \partial\{\widetilde{u} > 0\}.$$

Let w^ε be a solution of the heat equation in $\mathbb{R}^N \times (0,T)$ such that $\frac{w^\varepsilon(x,t)}{\varepsilon} \to w_0(x,t)$ uniformly in $\mathbb{R}^N \times [0,T]$. Assume, moreover, that $w_0 \in C(\mathbb{R}^N \times [0,T])$ and $w_0(x,t) \geq -1 + \delta_1$ for a certain positive constant δ_1.

Then there exists a family $u^\varepsilon \in C(\overline{Q_T})$, with $\nabla u^\varepsilon \in L^2_{\mathrm{loc}}(\overline{Q_T})$, of weak subsolutions to (P_ε) *in Q_T such that, as $\varepsilon \to 0$, $u^\varepsilon \to \widetilde{u}$ uniformly in $\overline{Q_T}$.*

Proof. The proof is analogous to Theorem 4.1. See [7] for a similar result in the case $w^\varepsilon = 0$. □

Finally, we end this section by showing that, for compactly supported initial data, the support of a limit solution of problem (P) is bounded.

PROPOSITION 4.1. *Let $u_0 \in C(\mathbb{R}^N)$ with compact support. Let u_0^ε converge uniformly to u_0 with supports converging to the support of u_0, and let w^ε be a solution of the heat equation in $\mathbb{R}^N \times (0,T)$ such that $\frac{w^\varepsilon(x,t)}{\varepsilon} \to w_0(x,t)$ uniformly in $\mathbb{R}^N \times [0,T]$. Assume, moreover, that $w_0 \in C(\mathbb{R}^N \times [0,T])$ and $w_0(x,t) \geq -1 + \delta_1$ for a certain positive constant δ_1. Finally, let u^ε be the solution to* (P_ε) *with function w^ε and initial condition u_0^ε.*

Let $u = \lim u^{\varepsilon_j}$. Then $\{u > 0\}$ is bounded. Moreover, u vanishes in finite time.

Proof. Let $-1 < \omega_0 < w^\varepsilon(x,t)/\varepsilon$. Then it is easy to check that

$$M_{\omega_0} = \int_{-\omega_0}^1 (s+\omega_0) f(s)\, ds < M(x,t) = \int_{-w_0(x,t)}^1 (s + w_0(x,t)) f(s)\, ds. \tag{4.5}$$

Let us now consider the following self-similar function:

$$V(x,t;T) = (T-t)^{1/2} h(|x|(T-t)^{-1/2}),$$

where $h = h(r)$ is a solution of

$$
\begin{aligned}
&h'' + \left(\frac{N-1}{r} + \frac{1}{2}r\right)h' + \frac{1}{2}h = 0, \quad 0 < r < R,\\
&h'(0) = 0, \quad h(r) > 0, \quad 0 \le r < R,\\
&h(R) = 0, \quad h'(R) = -\sqrt{2M_{\omega_0}}.
\end{aligned} \tag{4.6}
$$

It is proved in [4, Proposition 1.1] that there exists a unique $R > 0$ and a unique h solution of (4.6).

Moreover, it can be checked that if one picks T sufficiently large, then

$$V(x, 0; T) \ge u_0 + 1 \quad \text{in } \{u_0 > 0\},$$

and so $V(x, t; T)$ is a strict classical supersolution of (P) with bounded support and a positive gradient near its free boundary.

Now let u^{ε_j} be solutions to $(\mathrm{P}_{\varepsilon_j})$, with initial data $u_0^{\varepsilon_j}$ converging uniformly to u_0 such that support $u_0^{\varepsilon_j} \to$ support u_0 such that $u = \lim u^{\varepsilon_j}$.

By Theorem 4.1, there exists a family v^{ε_j} of supersolutions of $(\mathrm{P}_{\varepsilon_j})$ such that $v^{\varepsilon_j} \to V$ uniformly on compact sets, and $v^{\varepsilon_j}(x, 0) \ge u^{\varepsilon_j}(x, 0)$. Therefore, by the comparison principle, we obtain $u^{\varepsilon_j} \le v^{\varepsilon_j}$ and, passing to the limit, $u(x, t) \le V(x, t; T)$; the result follows. □

5. Uniqueness of the limit solution. In this section we arrive at the main point of the article: we prove that, under certain assumptions, there exists at most one limit solution to the initial and boundary value problem associated with (P) as long as condition (1.2) is satisfied.

Let us begin with the following proposition, which is the key ingredient in the proof of our main result.

PROPOSITION 5.1. *Let $\widetilde{u}$ be a strict classical supersolution to* (P) *with bounded support in $\mathbb{R}^N \times (0, T)$ such that there exists $s_0 > 0$ so that $|\nabla \widetilde{u}| > 0$ in $\{0 < \widetilde{u} < s_0\}$, and let $w^\varepsilon/\varepsilon$ be solutions to the heat equation in $\mathbb{R}^N \times (0, T)$ converging to w_0 uniformly with $w_0 \in C(\mathbb{R}^N \times [0, T])$ and $w_0 \ge -1 + \delta_1$ for a certain positive constant δ_1.*

Let u^ε be solutions to (P_ε) *with function w^ε and initial condition u_0^ε, where u_0^ε are uniform approximations of u_0 with* support $u_0^\varepsilon \to$ support u_0. *Then*

$$\limsup_{\varepsilon \to 0+} u^\varepsilon(x, t) \le \widetilde{u}(x, t)$$

for every $(x, t) \in Q_T$.

Proof. Let $\widetilde{u}$ be a strict classical supersolution of (P). Let us first define the following regularization:

$$u(x, t) = (\widetilde{u}(x, t + h) - \eta)^+$$

for $h, \eta > 0$ small so that u is a strict classical supersolution of (P) with C_x^1 free boundary, $C^1(\overline{\{u > 0\}})$, and $|\nabla u| > \delta_0 > 0$ in a neighborhood of its free boundary. So, by Theorem 4.1, there exists a v^ε supersolution of (P_ε) such that $v^\varepsilon \to u$ uniformly in Q_{T-h}.

Now, using the comparison principle, we conclude that $u^\varepsilon \le v^\varepsilon$ in Q_{T-h}, and the proposition now follows letting first $\varepsilon \to 0+$ and then $h, \eta \to 0+$. □

Finally, we arrive at the main point of the paper: the uniqueness of limit solutions of (P).

THEOREM 5.1. *Let the initial datum u_0 be Lipschitz continuous with compact support and satisfy condition* (2.1). *Then there exists at most one limit solution such that its gradient does not vanish near its free boundary as long as the function w^ε in problem* (P_ε) *satisfies condition* (1.3).

More precisely, let $u_0^{\varepsilon_j}, \tilde{u}_0^{\varepsilon_k}$ be uniformly Lipschitz continuous in $\mathbb{R}^N$ with uniformly bounded Lipschitz norms and $\varepsilon_j, \varepsilon_k \to 0$. Assume that $u_0^{\varepsilon_j} \in C^1(\overline{\{u_0^{\varepsilon_j} > 0\}})$, $\tilde{u}_0^{\varepsilon_k} \in C^1(\overline{\{\tilde{u}_0^{\varepsilon_k} > 0\}})$, $u_0^{\varepsilon_j}, \tilde{u}_0^{\varepsilon_k} \to u_0$ uniformly and support $u_0^{\varepsilon_j}$, support $\tilde{u}_0^{\varepsilon_k} \to$ support u_0. *Let $w^{\varepsilon_j}/\varepsilon_j$ and $\tilde{w}^{\varepsilon_k}/\varepsilon_k$ be solutions of the heat equation converging to the same function $w_0 \in C(\overline{Q_T})$, uniformly bounded from below by $-1+\delta_1$ for a certain positive constant δ_1. Also, assume that w_0 satisfies the monotonicity condition* (2.2).

Let u^{ε_j} (resp., $\tilde{u}^{\varepsilon_k}$) be the solution to ($\mathrm{P}_{\varepsilon_j}$) *with function w^{ε_j} and initial datum $u_0^{\varepsilon_j}$ (resp., the solution to* ($\mathrm{P}_{\varepsilon_k}$) *with function $\tilde{w}^{\varepsilon_k}$ and initial datum $\tilde{u}_0^{\varepsilon_k}$). Let $u = \lim u^{\varepsilon_j}$ and $\tilde{u} = \lim \tilde{u}^{\varepsilon_k}$. If there exists $s_0 > 0$ such that $|\nabla \tilde{u}| > 0$ in $\{0 < \tilde{u} < s_0\}$, then $u \le \tilde{u}$.*

Proof. Since $\tilde{u}$ is a classical supersolution of (P), $\tilde{u} \in C^1(\{\tilde{u} > 0\})$, and, by Proposition 4.1, its support is bounded, the function $\tilde{u}_\lambda$ as defined in (2.3) satisfies the hypotheses of Proposition 5.1 in $Q_{T/\lambda^2} \supset Q_T$. So by letting $\lambda \to 1-$ we arrive at

$$u(x,t) \le \tilde{u}(x,t). \tag{5.1}$$

This finishes the proof. ∎

THEOREM 5.2. *Let the initial datum u_0 be as in Theorem* 5.1. *Assume that there exists a classical solution v to* (P) *with initial datum u_0, and let $u_0^{\varepsilon_j}$ be uniformly Lipschitz continuous in $\mathbb{R}^N$ with $\varepsilon_j \to 0$ such that $u_0^{\varepsilon_j} \in C^1(\overline{\{u_0^{\varepsilon_j} > 0\}})$, $u_0^{\varepsilon_j} \to u_0$ uniformly, and* support $u_0^{\varepsilon_j} \to$ support u_0. *Assume $w^{\varepsilon_j}/\varepsilon_j$ is a solution of the heat equation converging to w_0 uniformly with $w_0 \in C(\mathbb{R}^N \times [0,T])$ and $w_0 \ge -1+\delta_1$ in $\mathbb{R}^N \times (0,T)$ for a certain $\delta_1 > 0$. Also, assume that w_0 satisfies the monotonicity condition* (2.2).

Let u^{ε_j} be the solution to ($\mathrm{P}_{\varepsilon_j}$) *with function w^{ε_j} and initial datum $u_0^{\varepsilon_j}$, and let $u = \lim u^{\varepsilon_j}$. Then $u = v$.*

In particular, there exists at most one classical solution to (P).

Proof. Since u is a classical supersolution to (P) and v is a classical subsolution, Lemma 2.1 applies, and we get that $v \le u$.

On the other hand, if we define v_λ as in (2.3), with $0 < \lambda < \lambda' < 1$, we have that v_λ is a strict classical supersolution. Since v_λ has compact support (see Proposition 4.1) it satisfies the hypotheses of Proposition 5.1. Thus,

$$u = \lim u^{\varepsilon_j} \le v_\lambda.$$

Letting $\lambda \to 1^-$ we obtain the desired result. ∎

6. Conclusions. In this paper we have proved that the limits of sequences of solutions to (P_ε) with different constitutive functions w^ε and initial data u_0^ε coincide, as long as certain monotonicity assumptions are made, if the limits of $w^\varepsilon/\varepsilon$ and of u_0^ε are prescribed.

The monotonicity assumptions are necessary to provide strict classical supersolutions as close as we want to any classical supersolution. This kind of condition was also used with the same purpose, in the case in which $w^\varepsilon = 0$, in [9] and [7]. In the latter, a different geometry was considered; namely, the domain was a cylinder, Neumann boundary conditions were given on the boundary of the cylinder, and monotonicity in

the direction of the cylinder axis was assumed. In [7] it was proved that, if a classical solution exists and $w^\varepsilon = 0$, then it is equal to any limit of solutions to (P_ε).

In our case, this is with $w^\varepsilon \neq 0$ satisfying (1.3) and nondecreasing in the direction of the cylinder axis; the uniqueness result in the presence of a classical solution still holds.

The cylindrical geometry has the advantage of giving the condition of nonvanishing gradient in the positivity set of any limit solution. Since in dimension 2 one can prove that limit solutions are classical supersolutions up to the fixed boundary, the uniqueness of limit solutions follows in this case without further assumptions.

REFERENCES

[1] J. D. Buckmaster and G. S. S. Ludford, *Theory of Laminar Flames*, Cambridge University Press, Cambridge, UK, 1982.

[2] L. A. Caffarelli, C. Lederman, and N. Wolanski, *Uniform estimates and limits for a two phase parabolic singular perturbation problem*, Indiana Univ. Math. J., 46 (1997), pp. 453–490.

[3] L. A. Caffarelli, C. Lederman, and N. Wolanski, *Pointwise and viscosity solutions for the limit of a two phase parabolic singular perturbation problem*, Indiana Univ. Math. J., 46 (1997), pp. 719–740.

[4] L. A. Caffarelli and J. L. Vazquez, *A free boundary problem for the heat equation arising in flame propagation*, Trans. Amer. Math. Soc., 347 (1995), pp. 411–441.

[5] J. Fernández Bonder and N. Wolanski, *A free boundary problem in combustion theory*, Interfaces Free Bound., 2 (2000), pp. 381–411.

[6] A. Langlois, *Sur l'ètude asymptotique d'un système parabolique modèlisant des flames presque èquidiffusives*, Ph.D. dissertation, Ecole Central de Lyon, Lyon, France, 2000.

[7] C. Lederman, J. L. Vazquez, and N. Wolanski, *Uniqueness of solution to a free boundary problem from combustion*, Trans. Amer. Math. Soc., 353 (2001), pp. 655–692.

[8] C. Lederman, J. L. Vazquez, and N. Wolanski, *Uniqueness of solution to a two-phase free boundary problem from combustion*, Adv. Differential Equations, 6 (2001), pp. 1409–1442.

[9] A. Petrosyan, *On existence and uniqueness in a free boundary problem from combustion*, Comm. Partial Differential Equations, 27 (2002), pp. 763–789.

SIAM J. MATH. ANAL.
Vol. 36, No. 1, pp. 186–203

SUFFICIENT CONDITIONS FOR THE EXISTENCE OF VISCOSITY SOLUTIONS FOR NONCONVEX HAMILTONIANS*

GIOVANNI PISANTE†

Abstract. We study a sufficient geometric condition for the existence of a $W^{1,\infty}(\Omega)$ *viscosity solution* of the Hamilton–Jacobi equation

$$\begin{cases} F(Du) = 0 & \text{in} \quad \Omega, \\ u = \varphi & \text{on} \quad \partial\Omega, \end{cases}$$

where $\Omega \subset \mathbb{R}^n$ and $F : \mathbb{R}^n \to \mathbb{R}$ are not necessarily convex.

Key words. Hamilton–Jacobi equations, viscosity solutions

AMS subject classifications. 49L25, 70H20

DOI. 10.1137/S0036141003426902

1. Introduction.

In this paper we consider the Dirichlet problem

$$\begin{cases} F(Du) = 0 & \text{in} \quad \Omega, \\ u = \varphi & \text{on} \quad \partial\Omega, \end{cases} \tag{1.1}$$

where $\Omega \subset \mathbb{R}^n$ is a bounded open set, $F : \mathbb{R}^n \to \mathbb{R}$ is continuous, and $\varphi \in Lip(\partial\Omega)$ (by the notation $\varphi \in Lip(\partial\Omega)$ we mean that there exists a constant $C \geq 0$ such that $|\varphi(x) - \varphi(y)| \leq C|x - y|$ for all $x, y \in \partial\Omega$). In particular, we are interested in the existence of viscosity solutions $u \in W^{1,\infty}(\Omega) \cap C(\overline{\Omega})$ of problem (1.1).

The study of Hamilton–Jacobi equations arises from classical problems in calculus of variations, and the notion of viscosity solution has aroused much interest since its introduction by Crandall and Lions in [11]. In particular, the interest of finding viscosity solutions of problem (1.1) is well known and studied in optimal control theory, differential games theory, etc. (see [1, 2, 9, 10, 21] for further details).

We should remark that the notion of viscosity solution is stronger than that of the almost everywhere solution: indeed, the viscosity method, when it establishes the existence of solutions, at the same time, gives a criterion of selection among them. Moreover, under appropriate hypotheses, we have uniqueness, maximality, stability, and explicit formulas (see [9, 10, 21]).

Here we want to investigate some sufficient geometrical conditions for the existence of $W^{1,\infty}(\Omega) \cap C(\overline{\Omega})$ viscosity solutions of (1.1).

This study has been motivated by a recent paper of Cardaliaguet et al. [6], where they gave a necessary and sufficient geometric condition for the problem (1.1) to admit a $W^{1,\infty}(\Omega)$ viscosity solution, under some restrictive hypotheses on Ω and φ. In particular, they showed that if Ω is convex, $\varphi \in C^1(\overline{\Omega})$, and verifies the compatibility condition

$$D\varphi(x) \in E \cup \operatorname{int} \operatorname{co} E \quad \forall\, x \in \Omega, \tag{1.2}$$

*Received by the editors April 28, 2003; accepted for publication (in revised form) November 14, 2003; published electronically June 22, 2004. This work was supported by the Ph.D. program in Mathematics of the University of Naples "Federico II" and partially supported by GNAMPA through the research project "Calcolo delle Variazioni, Teoria del Controllo e Ottimizzazione."
http://www.siam.org/journals/sima/36-1/42690.html

†Dipartimento di Matematica e Applicazioni "Renato Caccioppoli":—Universita Degli, Studi di Napoli "Federico II" Via Cintia, Monte S. Angelo I-80126 Napoli, Italy (pisante@unina.it) and Institut de Mathematiques, EPFL, 1015 Lausanne, Switzerland (giovanni.pisante@epfl.ch).

where $E = \{\xi \in \mathbb{R}^n \mid F(\xi) = 0\}$ and $\operatorname{int}\operatorname{co} E$ is the interior of the convex envelope of E, then the condition

- (G1) $\forall\, y \in \partial\Omega$, where the inward normal, $\nu(y)$, is uniquely defined, there exists $\lambda(y) > 0$ such that

$$D\varphi(y) + \lambda(y)\nu(y) \in E$$

is necessary and sufficient for the existence of $W^{1,\infty}(\Omega)$ viscosity solution of (1.1).

We should remark that, as shown in [15], the compatibility condition (1.2) is sufficient for the existence of infinitely many $W^{1,\infty}(\Omega)$ almost everywhere solutions of problem (1.1); in fact, the aim of [6] was to compare the theory for existence of almost everywhere solutions of implicit partial differential equations developed by Dacorogna and Marcellini (see [12, 13, 14, 15]) with the classical method of viscosity and to investigate the existence of $W^{1,\infty}$ viscosity solutions under assumption (1.2) only.

Our aim goes in a different direction: we want to show that the same type of techniques used in [6] can be refined to obtain a more general result in a more general framework. Moreover, we will see that the compatibility condition (1.2) can also be weakened in order to obtain a condition for the existence of viscosity solutions of equation (1.1).

We will prove that if Ω is bounded and connected, not necessarily convex, $\varphi \in Lip(\partial\Omega)$, and verifies a compatibility condition like (1.2) only on the boundary $\partial\Omega$ (the precise meaning of this condition will also be clarified in what follows), then the geometrical condition (G1) can be replaced by

- (G2) $\forall\, y \in \partial\Omega$, where $N^N_{\mathbb{R}^n\setminus\Omega}(y) \neq \emptyset$ there exists $h \in D^+\varphi(y)$ such that $\forall \nu \in N^N_{\mathbb{R}^n\setminus\Omega}(y)$ there exists a unique $\lambda_{\nu,h} > 0$ such that

$$h + \lambda_{\nu,h}\nu \in E,$$

  where $N^N_{\mathbb{R}^n\setminus\Omega}(y)$ is the normal cone to the set $\mathbb{R}^n \setminus \Omega$ and $D^+\varphi(y)$ is the superdifferential of φ in y (see Definitions 2.1 and 2.5).

In particular, we will see that (G2) is a sufficient condition for the existence of $W^{1,\infty}(\Omega)$ viscosity solutions of (1.1).

We should remark that (G2) strictly extends (G1): indeed, if Ω is convex, $\forall\, y \in \partial\Omega$, where the inward normal, $\nu(y)$, is uniquely defined, we have $N^N_{\mathbb{R}^n\setminus\Omega}(y) = \{\nu(y)\}$, and if $\varphi \in C^1$, then $D^+\varphi(y) = \{D\varphi(y)\}$ (see Remark 2.2 and Proposition 2.6).

REMARK 1.1. *If φ is an affine function, then the condition* (G2) *is also necessary for the existence of viscosity solutions, as it can be deduced by the last section of* [6].

To better understand the conditions (G1) and (G2) one should keep in mind the following examples.

EXAMPLE 1.2. *Let*

$$F_1(\xi_1, \xi_2) = -(\xi_1^2 - 1)^2 - (\xi_2^2 - 1)^2\ ;\ \varphi = 0.$$

Clearly,

$$\begin{cases} E_1 = \{\xi \in \mathbb{R}^2\ :\ \xi_1^2 = \xi_2^2 = 1\} = \{\xi \in \mathbb{R}^2\ :\ F_1(\xi) = 0\}, \\ \operatorname{co} E_1 = \{\xi \in \mathbb{R}^2\ :\ |\xi_1| \le 1\,,\ |\xi_2| \le 1\}, \\ E_1 \subset \partial(\operatorname{co} E_1)\ \ \textit{and}\ \ E_1 \neq \partial(\operatorname{co} E_1). \end{cases}$$

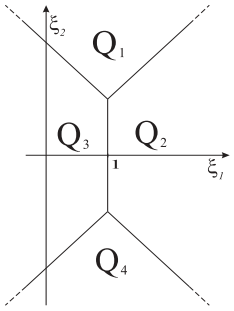

FIG. 1. *Partition of the plane in the definition of f.*

For this classical example the condition (G1) allows us to say that the only convex Ω for which there exists a $W^{1,\infty}(\Omega)$ viscosity solution of

$$\begin{cases} F_1(Du) = 0 & \text{in} \quad \Omega, \\ u = 0 & \text{on} \quad \partial\Omega \end{cases} \tag{1.3}$$

are rectangles whose normals are in E_1. The condition (G2) instead allows us to make this selection among all the sets Ω, convex and not; in particular, there are no $W^{1,\infty}(\Omega)$ viscosity solutions of problem (1.3) if Ω is a nonconvex domain.

EXAMPLE 1.3. *Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a positive continuous function which is zero only on the vertical segment $S = \{(\xi_1, \xi_2) \ : \ \xi_1 = 1, \xi_2 \in [-1, 1]\}$; for instance, we can consider*

$$f(\xi_1, \xi_2) = \begin{cases} \xi_2 - 1 \text{ if } (\xi_1, \xi_2) \in Q_1, \\ \xi_1 - 1 \text{ if } (\xi_1, \xi_2) \in Q_2, \\ -\xi_1 + 1 \text{ if } (\xi_1, \xi_2) \in Q_3, \\ -\xi_2 - 1 \text{ if } (\xi_1, \xi_2) \in Q_4, \end{cases}$$

where Q_i, $i = 1, \dots, 4$, is a partition of the plane as in Figure 1.

Let

$$F_2(\xi_1, \xi_2) = f(\xi_1, \xi_2) F_1(\xi_1, \xi_2) = f(\xi_1, \xi_2)[-(\xi_1^2 - 1)^2 - (\xi_2^2 - 1)^2],$$

where $F_1(\xi_1, \xi_2)$ is the function defined in the previous example and $\varphi = 0$.

Clearly, we have

$$\begin{cases} E_2 = E_1 \cup S = \{\xi \in \mathbb{R}^2 \ : \ F_2(\xi) = 0\}, \\ \operatorname{co} E_2 = \{\xi \in \mathbb{R}^2 \ : \ |\xi_1| \le 1 \, , \, |\xi_2| \le 1\}, \\ E_2 \subset \partial(\operatorname{co} E_2) \ \text{ and } \ E_2 \neq \partial(\operatorname{co} E_2). \end{cases}$$

If we consider the problem

$$\begin{cases} F_2(Du) = 0 & \text{in} \quad \Omega, \\ u = 0 & \text{on} \quad \partial\Omega, \end{cases} \tag{1.4}$$

where Ω is the nonconvex domain as in Figure 2*, we can easily verify the condition* (G2) *that ensures the existence of viscosity solutions. Indeed, since $\varphi = 0$, to verify* (G2) *it is sufficient to show that the sets of directions of the internal normal cone to*

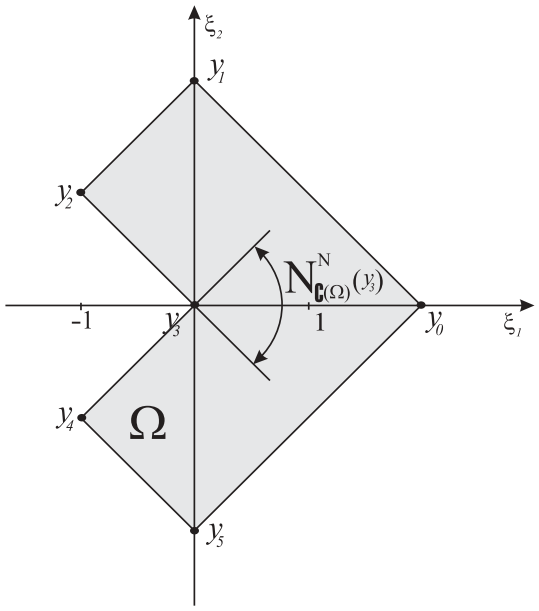

FIG. 2. *Domain of problem* (1.4).

$\partial\Omega$ at y, $N^N_{\mathbb{R}^n\setminus\Omega}(y)$, are contained in E_2 for every $y \in \partial\Omega$. In order to see this, we start by observing that in the points of regularity of $\partial\Omega$ the inward unit normal is in E_1. Then we have to consider $N^N_{\mathbb{R}^n\setminus\Omega}(y_i)$ for $i = 0, \dots, 5$. The only point at which $N^N_{\mathbb{R}^n\setminus\Omega}(y_i) \neq \emptyset$ is y_3, since at the other points Ω is convex and $N^N_{\mathbb{R}^n\setminus\Omega}(y_i)$ is empty; moreover, we can see that

$$N^N_{\mathbb{R}^n\setminus\Omega}(y_3) = S,$$

and this proves (G2).

2. Preliminaries. This section is divided into two parts. In the first part we recall several definitions of normal and tangent cones to a compact set that generalize the notions of normal and tangent vectors in the case where the set is not regular. In the second part, after recalling the definition of a viscosity solution of a Hamilton–Jacobi equation, we state some preliminary results on the existence of viscosity solutions of a Dirichlet problem with a convex Hamiltonian.

2.1. Normal and tangent cones. We start by giving some definitions.

DEFINITION 2.1. *Let K be a locally compact subset of $\mathbb{R}^n$ and $x \in K$. A vector $v \in \mathbb{R}^n$ is a* generalized tangent *to K at x if there are $h_n \to 0^+$, $v_n \to v$ such that $x + h_n v_n \in K \ \forall\, n \in \mathbb{N}$. The set of all generalized tangent vectors to K at x is denoted by $T_K(x)$, that is,*

$$T_K(x) := \left\{ v \in \mathbb{R}^n \,|\, \exists\, h_n \to 0^+ \,,\, v_n \to v \ :\ x + h_n v_n \in K \right\}.$$

A vector $\nu \in \mathbb{R}^n$ is a generalized outward normal *to K at x if, for every generalized tangent v to K at x, $\langle v, \nu \rangle \leq 0$. We denote by $N_K(x)$ the set of generalized normals to K at x. That is,*

$$N_K(x) := \left\{ \nu \in \mathbb{R}^n \,|\, \langle v, \nu \rangle \leq 0 \ \ \forall\, v \in T_K(x) \right\}.$$

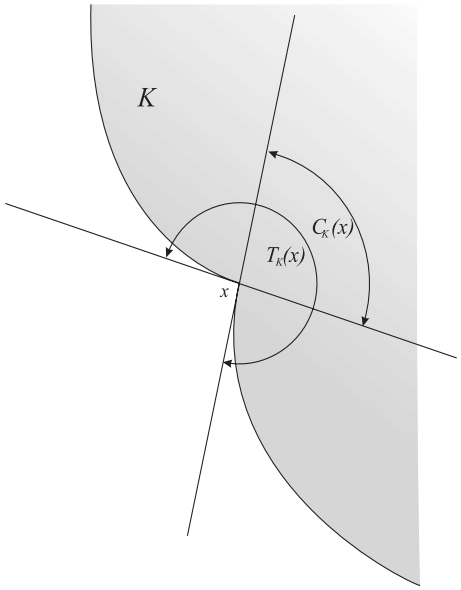

FIG. 3. *Different tangent cones.*

The set $T_K(x)$ is a closed cone containing the origin, and we will refer to it as *tangent cone*[1] to K at x; by duality we will call $N_K(x)$ the *normal cone* to K at x. Moreover, we denote by $N_K^N(x)$ the set of directions of the normal cone to K at x, that is,

$$N_K^N(x) := \left\{ \frac{\nu}{|\nu|}, \nu \in N_K(x) \setminus \{0\} \right\}.$$

REMARK 2.2. *Let K be a locally compact subset of $\mathbb{R}^n$ and $x \in K$.*

(i) *If the boundary of K is piecewise C^1, then $N_K^N(x)$ reduces to a single vector ν_x, where ν_x is the usual outward normal at any $x \in \partial K$, where the normal exists.*

(ii) *If Ω is an open subset of $\mathbb{R}^n$ and $x \in \partial\Omega$, then a generalized normal $\nu \in N_{\mathbb{R}^n \setminus \Omega}(x)$ can be regarded as an interior normal to Ω at x.*

Another useful set that can be defined is *Clarke's tangent cone* to K at x (see [8, 22]). It is defined by[2]

$$C_K(x) := \left\{ v \in \mathbb{R}^n \,|\, \forall x_n \to x\,, \forall t_n \to 0^+\,, \exists\, v_n \to v \;:\; x_n + t_n v_n \in K\,, \forall\, n \in \mathbb{N} \right\}.$$

DEFINITION 2.3. *A set K is said to be* regular *in the sense of Clarke at x, provided $T_K(x) = C_K(x)$.*

To have an idea of the relations between the two definitions of tangent cones $T_K(x)$ and $C_K(x)$ we can take a look at Figure 3.

[1] The set $T_K(x)$ was introduced in 1932 by Bouligand [4] with the name of *contingent cone* and it was studied for the theory of derivations of functions on $\mathbb{R}^2$. Later, in the theory of optimal control it was called simply *tangent cone* (see, for example, [19, 23, 24]).

[2] The original definition of $C_K(x)$ was given by Clarke in a slightly different way, more indirectly, but the two definitions are equivalent (see [7]).

REMARK 2.4. *Let K be a locally compact subset of $\mathbb{R}^n$ and $x \in K$.*

(i) *$C_K(x)$ is always a closed convex cone contained in $T_K(x)$; for this reason some authors prefer $C_K(x)$ instead of $T_K(x)$ as definition of tangent cone in many applications (see, for example,* [22]*).*

(ii) *If $T_K(x)$ is convex, then $N_K(x)$ is, in fact, the polar cone of $T_K(x)$ in the sense of convex analysis. It is the case, for example, of a set K regular in Clarke's sense at x for which we have*

$$N_K(x) = T_K^0(x) = C_K^0(x), \tag{2.1}$$

where $C_K^0(x)$ and $T_K^0(x)$ denote the polar cones of $C_K(x)$ and $T_K(x)$ in the sense of convex analysis.

(iii) *Any convex set is regular in the sense of Clarke.*

2.2. Viscosity solutions and convex Hamiltonians. Let us start by giving the definition of subdifferential and superdifferential of continuous functions defined on an open set $\Omega \subseteq \mathbb{R}^n$ (see [1, 2, 11, 17]).

DEFINITION 2.5. *Let $u \in C(\Omega)$; we define for $x \in \Omega$ the following sets:*

$$D^+u(x) = \left\{ p \in \mathbb{R}^n \ : \ \limsup_{y \to x, y \in \Omega} \frac{u(y) - u(x) - \langle p, y - x \rangle}{|x - y|} \leq 0 \right\},$$

$$D^-u(x) = \left\{ p \in \mathbb{R}^n \ : \ \liminf_{y \to x, y \in \Omega} \frac{u(y) - u(x) - \langle p, y - x \rangle}{|x - y|} \geq 0 \right\}.$$

$D^+u(x)$ ($D^-u(x)$) is called the superdifferential (subdifferential) of u at x.

In the following proposition we recall some useful properties of $D^+u(x)$ and $D^-u(x)$ that we will need in what follows.

PROPOSITION 2.6. *Let $u \in C(\Omega)$ and $x \in \Omega$. Then we have the following.*

(i) *$D^+u(x)$ and $D^-u(x)$ are closed, convex (possibly empty) subsets of $\mathbb{R}^n$.*

(ii) *If u is differentiable at x, then*

$$D^+u(x) = D^-u(x) = \{Du(x)\}. \tag{2.2}$$

(iii) *If for some x both $D^+u(x)$ and $D^-u(x)$ are nonempty, then* (2.2) *holds.*

(iv) *If $u \in W^{1,\infty}(\Omega)$, then*

$$D^+u(x) \cup D^-u(x) \subseteq \operatorname{co} \left\{ p \in \mathbb{R}^n \ \mid \ p = \lim_{n \to \infty} Du(x_n) \, , \, x_n \to x \right\}, \tag{2.3}$$

where the limit is taken over all the sequence $x_n \to x$ such that $Du(x_n)$ exists and the sequence $\{Du(x_n)\}$ converges.

There are many ways to define the $W^{1,\infty}$ viscosity solution of a differential equation (see [1, 11, 17]). Here we give a definition of such a solution in terms of sub- and superdifferential. We use this definition since it is more convenient for our purposes.

DEFINITION 2.7.

(i) *$u \in C(\Omega)$ is a viscosity subsolution of $F(Du(x)) = 0$ in Ω if and only if $F(p) \leq 0$ for every $x \in \Omega$ $\forall p \in D^+u(x)$.*

(ii) *$u \in C(\Omega)$ is a viscosity supersolution of $F(Du(x)) = 0$ in Ω if and only if $F(p) \geq 0$ for every $x \in \Omega$ $\forall p \in D^-u(x)$.*

A function $u \in C(\Omega)$ is a viscosity solution of $F(Du(x)) = 0$ if u is a viscosity subsolution and supersolution.

REMARK 2.8. *The definition of viscosity solution was originally given in terms of test functions (see, for example,* [21]*). The equivalence of the two definitions can be found in* [1] *or* [21].

For stating the main result we need to recall some preliminary results on viscosity solutions of Hamilton–Jacobi equations with convex Hamiltonians (cf. [18, 21] for further details).

We focus our attention on the problem

$$\begin{cases} H(Du) = n(x) & \text{in} \quad \Omega, \\ u(x) = \varphi(x) & \text{on} \quad \partial\Omega, \end{cases} \tag{2.4}$$

where $H : \mathbb{R}^n \to \mathbb{R}$ is convex, continuous, and satisfies

$$H(p) \to \infty \text{ as } |p| \to \infty,$$

and $n \in C(\overline{\Omega})$ is such that $n \geq \inf_{\mathbb{R}^n} H(p)$ in $\overline{\Omega}$.

We first define the function $L(x, y)\ \forall\, (x, y) \in \overline{\Omega} \times \overline{\Omega}$ as

$$L(x,y) := \inf_{\xi \,\in\, S_{x,y}} \left\{ \int_0^1 \max_{p \,\in\, P_{\xi,t}} \left\langle -\frac{d\xi}{dt}, p \right\rangle dt \right\}, \tag{2.5}$$

where

$$P_{\xi,t} := \Big\{ p \in \mathbb{R}^n \mid H(p) = n\big(\xi(t)\big) \Big\},$$

$$S_{x,y} := \left\{ \xi : [0,1] \to \bar{\Omega} \mid \xi(0) = x\,,\, \xi(1) = y\,,\, \frac{d\xi}{dt} \in L^\infty(0,1) \right\}.$$

REMARK 2.9. *We should point out that L can be written also in terms of the Lagrangian (i.e., the dual convex function) of H (see* [21, *section* 5.3]*).*

REMARK 2.10. *Many authors refer to the function L as optical length; let us point out why. For an admissible path ξ (i.e., a function $\xi : [0,1] \to \overline{\Omega}$ such that $\xi(0) = x$ and $\xi(1) = y$) we define the* optical length *of ξ as*

$$L(\xi) = \int_0^1 \max_{p \,\in\, P_{\xi,t}} \left\langle -\frac{d\xi}{dt}, p \right\rangle dt,$$

and this denomination introduced by Kružkov in [20] *is motivated by the fact that in the very special case $H(p) = |p|^2$, $n(x) = \text{const}$, this coincides with the optical length introduced by Born and Wolf in* [5].

Now we can state the classical (cf. [21, Theorem 5.2]).

THEOREM 2.11 (Hopf–Lax formula). *Let Ω be a bounded, connected domain of $\mathbb{R}^n$ with Lipschitz boundary $\partial\Omega$. Let $\varphi \in Lip(\partial\Omega)$. If φ verifies the compatibility condition*

$$\varphi(x) - \varphi(y) \leq L(x,y) \ \ \forall\, x, y \,\in\, \partial\Omega, \tag{2.6}$$

then the function

$$u(x) = \inf_{y \in \partial\Omega} \{\varphi(y) + L(x,y)\}$$

is the unique $W^{1,\infty}(\Omega)$ viscosity solution of the problem (2.4).

In section 3 we will apply Theorem 2.11 in the particular case where the Hamiltonian H is the gauge function of a convex set. For this reason we now want to investigate how L can be rewritten in the special case where $n(x) = 1$ and H is a gauge function; that is, H is convex and

$$\begin{cases} H(\xi) > 0 & \forall\, \xi \neq 0, \\ H(t\xi) = tH(\xi) & \forall\, \xi \in \mathbb{R}^n, \quad \forall\, t > 0. \end{cases}$$

Under these assumptions the function L can be rewritten as

$$L(x,y) := \inf_{S_{x,y}} \left\{ \int_0^1 \max_{H(p)=1} \left\langle -\frac{d\xi}{dt}, p \right\rangle dt \right\}, \tag{2.7}$$

and, by the definition of the polar[3] function of a gauge, (2.7) is equivalent to

$$L(x,y) := \inf_{S_{x,y}} \left\{ \int_0^1 H^0\left(-\frac{d\xi}{dt}\right) dt \right\}, \tag{2.8}$$

where H^0 is the polar function of H.

In the last part of this section we recall a Mac–Shane-type extension lemma which is, in fact, a consequence of the Hopf–Lax formula (for more details see, for example, [15]).

LEMMA 2.12. *Let $\Omega \subset \mathbb{R}^n$ be a bounded closed set. Let $H : \mathbb{R}^n \to \mathbb{R}$ be a gauge function, that is, a positively homogeneous convex function, and let H^0 be its polar. If $\varphi : \partial\Omega \to \mathbb{R}$ satisfies*

$$\varphi(x) - \varphi(y) \leq H^0(x-y) \quad \forall\, x, y \in \partial\Omega, \tag{2.9}$$

then the function

$$\tilde{\varphi}(x) = \inf_{y \in \partial\Omega} \left\{ \varphi(y) + H^0(x-y) \right\}$$

is a Lipschitz extension of φ to the whole $\mathbb{R}^n$, and, moreover, it satisfies

$$\tilde{\varphi}(x) - \tilde{\varphi}(y) \leq H^0(x-y) \quad \forall\, x, y \in \mathbb{R}^n$$

and

$$H^0(D\tilde{\varphi}(x)) \leq 1 \ \text{ a.e. in } \mathbb{R}^n. \tag{2.10}$$

REMARK 2.13. *The condition* (2.9) *is more restrictive than* (2.6) *since using Jensen's inequality we can easily prove that*

$$L(x,y) \geq H^0(x-y).$$

Moreover, we should note that if the segment $[x,y]$ is an admissible path for the definition of L (that is, it is completely contained in $\overline{\Omega}$), then $L(x,y) = H^0(x-y)$; this is the case, for example, when Ω is convex.

---

[3]The polar of a gauge H is defined as

$$H^0(\xi^*) = \inf \{ \lambda \geq 0 \ : \ \langle \xi, \xi^* \rangle \leq \lambda H(\xi) \ \ \forall\, \xi \in \mathbb{R}^n \}$$

and is characterized by

$$H^0(\xi^*) = \sup_{\xi \neq 0} \left\{ \frac{\langle \xi, \xi^* \rangle}{H(\xi)} \right\}.$$

3. Main result. In this section we establish a sufficient condition for the existence of a $W^{1,\infty}(\Omega)$ viscosity solution of the problem (1.1) under the following hypotheses:

- (H1) Let $F : \mathbb{R}^n \longrightarrow \mathbb{R}$ be continuous and such that

$$E = \{\xi \in \mathbb{R}^n \,:\, F(\xi) = 0\} \subset \partial(\operatorname{co} E),$$

with E bounded, $0 \in \operatorname{int}\operatorname{co} E$, and $F(\xi) < 0$ for every $\xi \in \operatorname{int}\operatorname{co} E$.

REMARK 3.1. *If F is convex and coercive, as in the classical literature, then*

$$\operatorname{co} E := \{\xi \in \mathbb{R}^n \,:\, F(\xi) \leq 0\},$$

and (H1) *is satisfied with* $E = \partial(\operatorname{co} E)$.

Following an idea used in [6], we want to compare the solution of (1.1) with the viscosity solution of the equation

$$\begin{cases} \rho(Du) = 1 & \text{in} \quad \Omega, \\ u = \varphi & \text{on} \quad \partial\Omega, \end{cases} \tag{3.1}$$

where ρ is the gauge associated with $\operatorname{co} E$ defined as

$$\rho(\xi) = \inf\{\lambda \geq 0 \mid \xi \in \lambda \operatorname{co} E\}.$$

We start by observing that ρ is well defined since by (H1) $0 \in \operatorname{int}\operatorname{co} E$ and $\operatorname{co} E$ is compact; moreover, ρ is, by definition, convex and positively homogeneous of degree 1. Therefore we have the right hypotheses to apply the preliminary work done in the previous section for the convex Hamiltonian; in particular, we can write the "optical length" $L(x, y)$ related to problem (3.1) as follows:

$$L(x,y) := \inf_{S_{x,y}} \left\{ \int_0^1 \rho^0 \left(-\frac{d\xi}{dt} \right) dt \right\}, \tag{3.2}$$

where ρ^0 is the polar function of ρ in the sense of convex analysis and, therefore, ρ^0 is convex and positively homogeneous.

Before stating the main result, we need to set our hypotheses on φ.

- (H2) Let $\varphi \in Lip(\partial\Omega)$, with

$$\emptyset \neq D^+\varphi(x) \subseteq E \cup \operatorname{int}\operatorname{co} E \;\; \forall\, x \in \partial\Omega \tag{3.3}$$

and satisfying the compatibility condition

$$\varphi(x) - \varphi(y) \leq \rho^0(x,y) \;\; \forall\, x, y \in \partial\Omega. \tag{3.4}$$

REMARK 3.2. *We should note that in condition* (H2) *we refer to $D^+\varphi(x)$ as the superdifferential of the Lipschitz extension of φ given by Lemma* 2.12. *Moreover, we can prove that $D^+\varphi(x) \subseteq \overline{\operatorname{co} E}\; \forall\, x \in \Omega$ (see the proof of Theorem* 3.3*).*

Finally, Theorem 2.11, Remark 2.13, and (H2) allow us to write the $W^{1,\infty}(\Omega)$ viscosity solution of (3.1) as follows:

$$u(x) = \inf_{y \in \partial\Omega} \{\varphi(y) + L(x,y)\}, \quad x \in \overline{\Omega}. \tag{3.5}$$

Now we are in the position to state the main theorem of this section.

THEOREM 3.3. *Let $\Omega \subset \mathbb{R}^n$ be a bounded connected set. Let F and φ satisfy* (H1) *and* (H2). *If $\forall\, y \in \partial\Omega$, where $N^N_{\mathbb{R}^n\setminus\Omega}(y) \neq \emptyset$, there exists $h \in D^+\varphi(y)$ such that $\forall \nu \in N^N_{\mathbb{R}^n\setminus\Omega}(y)$ there exists a unique $\lambda_{\nu,h} > 0$ that verifies*

$$h + \lambda_{\nu,h}\nu \in E,$$

then there exists $u \in W^{1,\infty}(\Omega)$ viscosity solution of (1.1).

REMARK 3.4. *Let $h \in D^+\varphi(y)$ be as in Theorem* 3.3 *and $\nu \in N^N_{\mathbb{R}^n\setminus\Omega}(y)$; then, since $E \subset \partial(\operatorname{co} E)$, the unique $\lambda_{\nu,h} > 0$ such that $h + \lambda_{\nu,h}\nu \in E$ is determined by the equality*

$$\rho(h + \lambda_{\nu,h}\nu) = 1.$$

We will prove that, under the hypotheses of Theorem 3.3, the function $u : \overline{\Omega} \to \mathbb{R}$ defined by (3.5) is actually the viscosity solution of (1.1). Before starting the proof we need to investigate the properties of u. Let us start by proving the following key lemma and making some remarks.

LEMMA 3.5. *Let Ω be a bounded connected open set of $\mathbb{R}^n$ with Lipschitz boundary $\partial\Omega$, and let $\varphi \in Lip(\partial\Omega)$ verify* (H2). *Let u be defined by* (3.5) *and $y(x) \in \partial\Omega$ be such that $u(x) = \varphi(y(x)) + L(x, y(x))$. Then $\forall\, p \in D^-u(x)$ and $\forall\, h \in D^+\varphi(y(x))$,*

$$\langle p - h, q\rangle \leq 0 \quad \forall\, q \in T_{\mathbb{R}^n\setminus\Omega}(y(x)),$$

that is, $p - h \in N_{\mathbb{R}^n\setminus\Omega}(y(x))$.

Proof. Let $x_0 \in \Omega$, $y_0 \in \partial\Omega$ such that $u(x_0) = \varphi(y_0) + L(x_0, y_0)$ and $q \in T_{\mathbb{R}^n\setminus\Omega}(y_0)$. Let $q_k \to q$, as in Definition 2.1, such that $y_0 + \varepsilon_k q_k \notin \Omega$ and $x_0 + \varepsilon_k q_k \in \Omega$. By definition of $L(x_0, y_0)$ for every $\varepsilon > 0$ we can find $\xi_0 \in S_{x_0,y_0}$ (that is, $\xi_0 : [0,1] \to \overline{\Omega} \mid \xi(0) = x_0\,,\ \xi(1) = y_0\,,\ \frac{d\xi_0}{dt} \in L^\infty(0,1)$) such that

$$L(x_0, y_0) + \varepsilon \geq \int_0^1 \rho^0\left(-\frac{d\xi_0}{dt}(t)\right) dt. \tag{3.6}$$

Next we define, for every $k \in \mathbb{N}$, $\xi_k(t) = \xi_0(t) + \varepsilon_k q_k$; clearly, we have $\xi_k(0) = x_0 + \varepsilon_k q_k$, $\xi_k(1) = y_0 + \varepsilon_k q_k$, and $\frac{d\xi_k}{dt} = \frac{d\xi_0}{dt}$.

Since ξ_k and $\partial\Omega$ are continuous, there exist $t_k \in (0,1)$ and $y_k \in \partial\Omega$ such that $\xi_k(t_k) = y_k$ and $\xi_k(t) \in \overline{\Omega} \quad \forall\, t < t_k$ (see Figure 4).

Using (3.6), the properties of ξ_k, and the definition (3.5) of u, we have

$$\begin{aligned}
u(x_0) &= \varphi(y_0) + L(x_0, y_0) \\
&\geq \varphi(y_0) + \int_0^1 \rho^0\left(-\frac{d\xi_0}{dt}(t)\right) dt - \varepsilon \\
&= \varphi(y_0) - \varphi(y_0 + \varepsilon_k q_k) \\
&+ \varphi(y_0 + \varepsilon_k q_k) - \varphi(y_k) + \int_{t_k}^1 \rho^0\left(-\frac{d\xi_0}{dt}(t)\right) dt \\
&+ \varphi(y_k) + \int_0^{t_k} \rho^0\left(-\frac{d\xi_0}{dt}(t)\right) dt - \varepsilon \\
&\geq \varphi(y_0) - \varphi(y_0 + \varepsilon_k q_k) + u(x_0 + \varepsilon_k q_k) \\
&+ \varphi(y_0 + \varepsilon_k q_k) - \varphi(y_k) + \int_{t_k}^1 \rho^0\left(-\frac{d\xi_0}{dt}(t)\right) dt - \varepsilon,
\end{aligned} \tag{3.7}$$

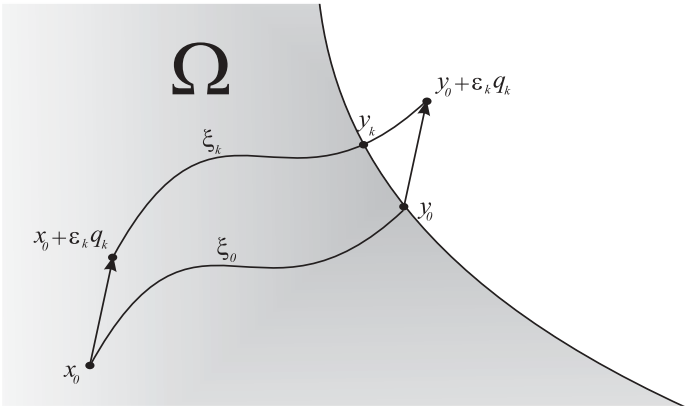

FIG. 4. *Geometrical construction of path* ξ_k.

where we have used the homogeneity of ρ^0 to establish

$$\int_0^{t_k} \rho^0\left(-\frac{d\xi_0}{dt}(t)\right)dt = \int_0^1 \rho^0\left(-\frac{d(\xi_k(t_k s))}{ds}\right)ds \geq L(x_0+\varepsilon_k q_k, y_k).$$

We claim that

$$\varphi(y_0+\varepsilon_k q_k) - \varphi(y_k) + \int_{t_k}^1 \rho^0\left(-\frac{d\xi_0}{dt}(t)\right)dt \geq 0. \tag{3.8}$$

Indeed, Lemma 2.12 ensures us that

$$\varphi(y_0+\varepsilon_k q_k) - \varphi(y_k) \geq -\rho^0(y_k - y_0 - \varepsilon_k q_k); \tag{3.9}$$

moreover, by Jensen's inequality we have

$$\begin{aligned}\int_{t_k}^1 \rho^0\left(-\frac{d\xi_0}{dt}(t)\right)dt &= \int_0^1 \rho^0\left(-\frac{d\xi_k((1-t_k)s+t_k)}{ds}\right)ds \\ &\geq \rho^0(y_k - y_0 - \varepsilon_k q_k).\end{aligned} \tag{3.10}$$

Combining (3.9) and (3.10) we obtain the claim.

Now using (3.7) and (3.8) we can write, letting $\varepsilon \to 0$,

$$u(x_0) \geq u(x_0+\varepsilon_k q_k) - \big(\varphi(y_0+\varepsilon_k q_k) - \varphi(y_0)\big). \tag{3.11}$$

Therefore, taking $h \in D^+\varphi(y_0)$ and $p \in D^- u(x_0)$, we have by definition that

$$\begin{aligned}\varphi(y_0+\varepsilon_k q_k) - \varphi(y_0) &\leq \langle h, \varepsilon_k q_k\rangle + o(\varepsilon_k),\\ u(x_0+\varepsilon_k q_k) - u(x_0) &\geq \langle p, \varepsilon_k q_k\rangle + o(\varepsilon_k),\end{aligned}$$

and in light of (3.11), we can say that

$$\begin{aligned}\langle p, \varepsilon_k q_k\rangle &\leq u(x_0+\varepsilon_k q_k) - u(x_0) + o(\varepsilon_k)\\ &\leq \varphi(y_0+\varepsilon_k q_k) - \varphi(y_0) + o(\varepsilon_k)\\ &\leq \langle p, \varepsilon_k q_k\rangle + o(\varepsilon_k).\end{aligned}$$

Finally, dividing both sides of the last inequality by ε_k and taking the limit for $k \to \infty$, we obtain

$$\langle p - h, q \rangle \leq 0.$$

This proves the lemma. □

REMARK 3.6. *If we fix $p \in D^- u(x)$ and $h \in D^+\varphi(y(x))$ with $h \neq p$, then there exist $\nu_{p,h} \in N^N_{\mathbb{R}^n \setminus \Omega}(y(x))$ and a unique $\lambda_{p,h} > 0$ such that $p = h + \lambda_{p,h}\nu_{p,h}$.*

We now give the proof of the main theorem.

Proof of Theorem 3.3. Let u be defined as in (3.5); by definition, u is a viscosity solution of (3.1). We claim that u is also a viscosity solution of (1.1). We divide the proof into two steps: first, we show that u is in fact a supersolution of (1.1) and then show that u is also a subsolution.

- We start by observing that $\forall\, x \in \Omega$ and $\forall\, p \in D^- u(x)$ we have $\rho(p) = 1$ (see also [3]). Indeed, since u is a supersolution of (3.1), we have that $\forall\, x \in \Omega$ and $\forall\, p \in D^- u(x)$, $\rho(p) \geq 1$. Moreover, since u is also a viscosity subsolution of (3.1), in particular we have $\rho(Du(x)) \leq 1$ (i.e., $Du(x) \in \overline{\operatorname{co} E}$) $\forall\, x \in \Omega$, where $Du(x)$ exists, since in such points $D^+ u(x) = \{Du(x)\}$. The continuity of ρ ensures us that

$$\rho\left(\lim_{n\to\infty} Du(x_n)\right) \leq 1$$

$\forall\, x_n \to x$ such that $Du(x_n)$ is well defined and $Du(x_n)$ converges; that is, the following inclusion holds:

$$\left\{p \in \mathbb{R}^n \mid p = \lim_{n\to\infty} Du(x_n) \ : \ x_n \to x\right\} \subseteq \overline{\operatorname{co} E}. \tag{3.12}$$

Therefore, by Proposition 2.6(iv) and (3.12) we can say that

$$D^- u(x) \subseteq \operatorname{co}\left\{p \in \mathbb{R}^n \mid p = \lim_{n\to\infty} Du(x_n) \ : \ x_n \to x\right\} \subseteq \overline{\operatorname{co} E},$$

that is, $\rho(p) \leq 1\ \forall\, p \in D^- u(x)$, and this proves the claim.
Now let $y(x) \in \partial\Omega$ be such that $u(x) = \varphi(y(x)) + L(x, y(x))$ and $h \in D^+\varphi(y(x))$ as in the hypotheses. We distinguish two cases.

(1) If $h = p$, then $\rho(h) = 1$; since $h \in E \cup \operatorname{int}\operatorname{co} E$, we have $h \in E$, and so $p \in E$, that is, $F(p) = 0$.

(2) If $h \neq p$, by Remark 3.6, there exist $\nu_{p,h} \in N^N_{\mathbb{R}^n \setminus \Omega}(y(x))$ and a unique $\lambda_{p,h} > 0$ such that

$$p = h + \lambda_{p,h}\nu_{p,h}; \tag{3.13}$$

moreover, $\lambda_{p,h}$ is uniquely determined by $\rho(h + \lambda_{p,h}\nu_{p,h}) = 1$. The hypothesis made on h and (3.13) imply $p \in E$; that is, as before, $F(p) = 0$.
In particular, u is a viscosity supersolution of (1.1).

- Since u is also a viscosity subsolution of (3.1), then for every $x \in \Omega$ and $p \in D^+ u(x)$ we have $p \in \overline{\operatorname{co} E}$ (i.e., $\rho(p) \leq 1$). As (H1) is satisfied and F is continuous, it follows that $F(p) \leq 0$. So u is a viscosity subsolution of (1.1).

The two above observations complete the proof. □

4. Corollaries. This section is divided into two parts. In the first part we focus our attention on the differentiability properties of Lipschitz and semiconcave functions with the aim of relating the notions of normal and tangent cones to the sets described by such types of functions (e.g., epigraphs or level-sets) to their generalized gradients. In the second part we state two corollaries of Theorem 3.3 in which the hypotheses on the geometry of the domain Ω can be written in a nicer way in terms of the differential property of the functions that represent the boundary $\partial\Omega$.

4.1. Lipschitz continuity and semiconcavity. Let us recall briefly some definitions and relevant differential properties of locally Lipschitz continuous functions that we will use in what follows. By the Rademacher theorem such functions are almost everywhere differentiable with locally bounded gradients (see [16]). Hence, if $u \in Lip_{\mathrm{loc}}(\Omega)$, we can consider the set

$$D^*u(x) := \Big\{ p \in \mathbb{R}^n \,:\, p = \lim_{n\to\infty} Du(x_n), x_n \to x \Big\},$$

where x_n is a sequence of points of differentiability for u. We note that $D^*u(x)$ is nonempty and closed for any $x \in \Omega$.

Let $u : \Omega \to \mathbb{R}$ be Lipschitz in a neighborhood of a given point x, and let $q \in S^{n-1}$ be a direction in $\mathbb{R}^n$. We define

- the *one-sided directional derivative* of u at x in the direction q as

$$u'(x,q) = \lim_{t\to 0^+} \frac{u(x+tq)-u(x)}{t},$$

- the *generalized directional derivatives* of u at x in the direction q as

$$u^0(x,q) = \limsup_{y\to x, t\to 0^+} \frac{u(y+tq)-u(y)}{t},$$

$$u_0(x,q) = \liminf_{y\to x, t\to 0^+} \frac{u(y+tq)-u(y)}{t},$$

- the *generalized gradient* (or Clarke's gradient) of u at x as

$$\begin{aligned} \partial u(x) &= \{ p \in \mathbb{R}^n \,:\, u^0(x,q) \geq p\cdot q \ \forall q \in \mathbb{R}^n \} \\ &= \{ p \in \mathbb{R}^n \,:\, u_0(x,q) \leq p\cdot q \ \forall q \in \mathbb{R}^n \}. \end{aligned}$$

In the following proposition we collect some well-known properties of Lipschitz functions (see [1, 8]).

PROPOSITION 4.1. *Let $u : \Omega \to \mathbb{R}$ be locally Lipschitz continuous in the open set Ω; then*

(i) $u_0(x,q) = -u^0(x,-q)\ \forall\, x \in \Omega,\ q \in \mathbb{R}^n$;

(ii) $\forall\, x \in \Omega$ *the function $q \mapsto u^0(x,q)$ is finite, positively homogeneous, subadditive, and convex (and locally Lipschitz continuous);*

(iii) *the map $(x,q) \mapsto u^0(x,q)$ is upper semicontinuous;*

(iv) $\forall\, x \in \Omega$ *we have $\operatorname{co} D^*u(x) = \partial u(x)$;*

(v) *$D^+u(x)$ and $D^-u(x)$ are bounded $\forall\, x \in \Omega$ and*

$$D^+u(x) \cup D^-u(x) \subseteq \partial u(x);$$

(vi) $\forall\, q \in S^{n-1}$ *there exists the classical one-sided directional derivative* $u'(x,q)$ *at any* $x \in \Omega$, *where* $D^+u(x) = \partial u(x)$ *and the following equality holds:*

$$u'(x,q) = \min_{p \in D^+u(x)} p \cdot q = u_0(x,q). \tag{4.1}$$

REMARK 4.2. *Looking at the definition of* $D^*u(x)$ *and Proposition* 4.1(iv), *one can observe that Proposition* 4.1(v) *is just a reformulation of Proposition* 2.6(iv).

Now we introduce a definition of regularity of functions that is in some way related to regularity of sets in Clarke's sense (from which the name derives). It will be useful for stating some hypotheses that allow us to write the normal cone of a set in a nicer way.

DEFINITION 4.3. *A function* $u : \Omega \to \mathbb{R}$ *is said to be* regular *at* x *(in the sense of Clarke), provided*

(i) $\forall\, q \in \mathbb{R}^n$ *the one-sided directional derivative* $u'(x,q)$ *exists;*

(ii) $\forall\, q \in \mathbb{R}^n$ *the equality* $u'(x,q) = u^0(x,q)$ *holds.*

The following theorem (a proof of which can be found in [8]) and its corollaries give us a useful characterization of normal cone to the level sets of regular functions.

THEOREM 4.4. *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be Lipschitz near a given point* x, *and suppose that* $0 \notin \partial f(x)$. *If* K *is defined as*

$$K := \{y \in \mathbb{R}^n \,:\, f(y) \le f(x)\},$$

then

$$C_K^0(x) \subset \bigcup_{\lambda \ge 0} \lambda \partial f(x).$$

If, in addition, f *is regular in the sense of Clarke at* x, *then equality holds, and* K *is Clarke regular at* x, *that is,*

$$N_K(x) = C_K^0(x) = \bigcup_{\lambda \ge 0} \lambda \partial f(x). \tag{4.2}$$

REMARK 4.5. *The first equality in* (4.2) *follows by* (2.1) *of Remark* 2.4(iii), *since* K *is regular.*

REMARK 4.6. *The above proposition holds also in a more general framework, that is, for functions defined in a general Banach space, as stated in* [8].

COROLLARY 4.7. *Let* $\Omega := \{y \in \mathbb{R}^n \,:\, f(y) > 0\}$, *where* f *is a Lipschitz continuous function. Let* $y_0 \in \partial\Omega$, *and suppose that* f *verifies the following properties:*

(i) f *is regular in Clarke's sense at* y_0;

(ii) $0 \notin \partial f(y_0) = D^- f(y_0) \cup D^+ f(y_0)$;

then

$$N^N_{\mathbb{R}^n \setminus \Omega}(y_0) = \left(D^- f(y_0) \cup D^+ f(y_0)\right)^N. \tag{4.3}$$

Proof. We note first that $\partial\Omega \subseteq \{y \in \mathbb{R}^n \,:\, f(y) = 0\}$; then $y_0 \in \partial\Omega$ imply $f(y_0) = 0$. So we can write

$$\mathbb{R}^n \setminus \Omega := \{y \in \mathbb{R}^n \,:\, f(y) \le f(y_0)\}.$$

Hence we can apply Theorem 4.4, and, in particular, since f is Clarke regular, by (4.2) we have

$$N_{\mathbb{R}^n \setminus \Omega}(y_0) = \bigcup_{\lambda \ge 0} \lambda \partial f(y_0).$$

Finally, we can conclude, using hypothesis (ii) of Corollary 4.7, that

$$N^N_{\mathbb{R}^n\setminus\Omega}(y_0) = \left(\bigcup_{\lambda\geq 0}\lambda\partial f(y_0)\right)^N = (\partial f(y_0))^N = \left(D^- f(y_0)\cup D^+ f(y_0)\right)^N. \qquad \square$$

In order to prove a second corollary of Theorem 4.4 that is equally useful, we need to recall the definition and some relevant properties of semiconcave and semiconvex functions (see [1] for further details).

DEFINITION 4.8. *We say that* $u:\Omega\to\mathbb{R}$ *is* semiconcave *on an open convex set* Ω *if there exists a constant* $C>0$ *such that*

$$\lambda u(x)+(1-\lambda)u(y)\leq u\left(\lambda x+(1-\lambda)y\right)+\frac{1}{2}C\lambda(1-\lambda)|x-y|^2 \tag{4.4}$$

or, equivalently, if the application $x\mapsto u(x)-\frac{1}{2}C|x|^2$ *is concave.*

We say that $u:\Omega\to\mathbb{R}$ *is semiconvex if* $-u$ *is semiconcave.*

If u is continuous, an equivalent way to express condition (4.4) is to require that

$$u(x+h)-2u(x)+u(x-h)\leq C|h|^2$$

for any $x\in\Omega$ and $h\in\mathbb{R}^n$ with sufficiently small $|h|$.

REMARK 4.9. *It can be proved (see, for example,* [1]*) that a semiconcave function* u *in* Ω *is in fact locally Lipschitz continuous and* $\forall\, x\in\Omega$ *we have*

$$D^+u(x)=\partial u(x)=co\, D^*u(x),$$

while

$$D^-u(x)\neq 0\ \Rightarrow\ u \text{ is differentiable in } x.$$

Now we can prove the following corollary.

COROLLARY 4.10. *Let* $\Omega:=\{y\in\mathbb{R}^n\,:\,f(y)\leq 0\}$*, where* f *is a semiconcave function, if* $y_0\in\partial\Omega$ *and* $0\notin D^+f(y_0)$*; then*

$$N^N_{\mathbb{R}^n\setminus\Omega}(y_0)=-(D^+f(y_0))^N.$$

Proof. We first note that from Remark 4.9 f is locally Lipschitz continuous and $D^+f(y_0)=\partial f(y_0)$. So we can say, by Proposition 4.1(vi), that

$$f'(y_0,q)=f_0(y_0,q)\quad\forall\, q\in S^{n-1}.$$

Moreover, using the definition of generalized derivatives we have

$$-\left(-f'(y_0,q)\right)=f'(y_0,q)=f_0(y_0,q)=-\left(-f^0(y_0,q)\right)\quad\forall\, q\in S^{n-1};$$

that is, $-f$ is regular at y_0 in the sense of Clarke. We now observe that, since

$$-D^-(-f)(y_0)=D^+f(y_0)=\partial f(y_0)=-\partial(-f)(y_0),$$

f verifies the hypothesis of Corollary 4.7 with $\Omega:=\{y\in\mathbb{R}^n\,:\,-f(y)>0\}$, and so we have

$$N^N_{\mathbb{R}^n\setminus\Omega}(y_0)=\left(D^-(-f)(y_0)\right)^N=-(D^+f(y_0))^N. \qquad \square$$

REMARK 4.11. *The two above corollaries hold also if the hypotheses are verified only locally, that is, if for* $y_0 \in \partial\Omega$ *there exists a ball* $B(y_0, r)$ *centered in* y_0 *such that* $\Omega \cap B(y_0, r)$ *can be represented as the sublevel or superlevel set of a function defined on* $B(y_0, r)$ *satisfying the hypotheses required.*

The last result that we want to recall can be found in [8], and it gives us a useful relation between the generalized gradient of a locally Lipschitz function f and Clarke's normal cone $C^0_{epi\, f}$ to its epigraph.

PROPOSITION 4.12. *Let* $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$ *be Lipschitz continuous near a given point* x*; then* $\xi \in \mathbb{R}^n$ *belongs to* $\partial f(x)$ *if and only if* $(\xi, -1)$ *belongs to* $C^0_{epi\, f}(x, f(x))$.

4.2. Corollaries. In the two following corollaries we consider some hypotheses on the geometry of the domain Ω that allow us to write Theorem 3.3 in a nicer way.

Let Ω be a Lipschitz domain, and we have that Ω can be locally represented as the epigraph of a Lipschitz function; that is, $\forall y \in \partial\Omega$ there exists a direction ν_y and a function ω_y defined on the hyperplane orthogonal to ν_y such that in a neighborhood of y, Ω is the epigraph of ω_y.

DEFINITION 4.13. *We will say that* Ω *is convex (concave) at* $y \in \partial\Omega$ *if there exists a* $\nu_y \in S^{n-1}$ *such that the function* ω_y*, which represents* Ω *in the direction* ν_y*, is convex (concave).*

COROLLARY 4.14. *Let* Ω *be a locally Lipschitz domain, and denote by* J *the set of the points of nondifferentiability of* $\partial\Omega$*. Suppose that* Ω *is convex or concave at* $y\ \forall\, y \in J$*. Let* F *and* φ *satisfy* (H1) *and* (H2).

If $\forall y \in \partial\Omega$*, where* $D^+\omega_y(y) \neq \emptyset$*, there exists* $h \in D^+\varphi(y)$ *such that* $\forall \xi \in D^+(\omega_y)(y)$ *there exists a unique* $\lambda_{h,\xi}$ *that verify*

$$h - \lambda_{h,\xi}(\xi + \nu_y) \in E, \tag{4.5}$$

then there exists a $u \in W^{1,\infty}(\Omega)$ *viscosity solution of* (1.1).

REMARK 4.15. *We have to note that in* (4.5) ξ *has to be considered as a point of* $\mathbb{R}^n$ *using the classical immersion in* $\mathbb{R}^n$ *of the hyperplane orthogonal to* ν_y *to which* ξ *belongs by definitions.*

REMARK 4.16. *In the statement of Corollary* 4.14 *we have used the functions* ω_y*; with this notation it seems that we have to change* $\omega_y\ \forall\, y \in \partial\Omega$*, but we can simply observe that the compactness of* $\partial\Omega$ *ensures us that we need only a finite number of* ω_y*. In fact, we can consider for every* $y \in \partial\Omega$ *a neighborhood* Ω_y *of* y *in which* Ω *is represented by the function* ω_y*. From this cover we can extract a finite one* $\cup_{i=1}^k \Omega_{y_i}$*, where* $\omega_y = \omega_{y_i}$ *for every* $y \in \Omega_{y_i} \cap \partial\Omega$.

REMARK 4.17. *If we consider an orthogonal basis* $\{e_1, \ldots, e_n\}$ *for* $\mathbb{R}^n$*, with* $e_n = \nu_y$*, we note that* ξ *lives in the space spanned by* $\{e_1, \ldots, e_{n-1}\}$*, and* (4.5) *can be rewritten as*

$$h - \lambda_{h,\xi}(\xi, 1) \in E \quad \forall \xi \in D^+(f_y)(y).$$

Proof of Corollary 4.14. Looking at the proof of Theorem 3.3 we need only to work with the points on $\partial\Omega$ that realize the minimum in definition (3.5). Now let $x \in \Omega$ and $y \in \partial\Omega$ be such that $u(x) = \varphi(y) + L(x, y)$. If $D^+\omega_y(y) \neq \emptyset$, then Ω is convex in y, and we can prove, using the same argument of Lemma 2.9 in [6], that y must be a point of differentiability for $\partial\Omega$, and this is a contradiction. Hence we have that all the points that realize the minimum in (3.5) have $D^+\omega_y \neq \emptyset$.

Now we want to identify the set $N_{\mathbb{R}^n \setminus \Omega}(y)$ and write it in terms of the superdifferential of ω_y in order to apply Theorem 3.3.

We first observe that if ω_y is differentiable in y, then $N_{\mathbb{R}^n\setminus\Omega}(y)$ reduces to the classic interior normal to $\partial\Omega$ given by $(D\omega(y)+\nu_y)$, and there is nothing to prove.

The last case that we have to consider is if Ω is concave at y and $D^+\omega_y(y)$ does not reduce to a single vector. In this case we have that $-\omega_y$ is convex near y, and it represents $\mathbb{R}^n\setminus\Omega$ in the direction $-\nu_y$. Hence $\mathbb{R}^n\setminus\Omega$ is convex near y, and so, by Remark 2.4(iii), is Clarke regular, and we have

$$N_{\mathbb{R}^n\setminus\Omega}(y)=C^0_{\mathbb{R}^n\setminus\Omega}(y)=C^0_{epi\,(-\omega_y)}(y). \tag{4.6}$$

Moreover, by Proposition 4.12 we can write

$$C^0_{epi\,(-\omega_y)}(y)=\{(\xi,-1)\,:\,\xi\in\partial(-\omega_{y(y)})(y)\}. \tag{4.7}$$

We now observe that, since $-\omega_y$ is convex, we have

$$\partial(-\omega_{y(y)})(y)=D^-(-\omega_y)(y)=-D^+\omega_y(y), \tag{4.8}$$

and, finally, by (4.6), (4.7), and (4.8) we have

$$N^N_{\mathbb{R}^n\setminus\Omega}(y)=\{(\xi,-1)\,:\,\xi\in-D^+\omega_y(y)\}^N.$$

The conclusion follows by Theorem 3.3. □

Another way to represent a domain is like a sublevel- or superlevel-set of a given function. Also in such a case, we have an appropriate version of Theorem 3.3. It is clear that if $\partial\Omega$ is regular in a neighborhood of a point $y\in\partial\Omega$, we can locally (near y) write Ω as the sublevel-set of a regular function f^Ω_y; suppose, moreover, that Ω verifies the following hypothesis:

- (H3) Let Ω be a locally Lipschitz domain, and denote by J the set of the points of nondifferentiability of $\partial\Omega$. Suppose that if for $y\in J$ there exists an $x\in\Omega$ such that $u(x)=\varphi(y)+L(x,y)$ (that is, y realizes the minimum in definition (3.5)), then Ω can be represented near y as the sublevel-set of a semiconcave function f^Ω_y (see Remark 4.9).

The following corollary is an easy consequence of Theorem 3.3 and Corollary 4.10.

COROLLARY 4.18. *Let Ω, F, and φ satisfy* (H1), (H2), *and* (H3).

If $\forall y\in\partial\Omega$, where $D^+f^\Omega_y(y)\neq\emptyset$, there exists $h\in D^+\varphi(y)$ such that $\forall\xi\in D^+f^\Omega_y(y)$ there exists a unique $\lambda_{h,\xi}$ that verify

$$h-\lambda_{h,\xi}\xi\in E, \tag{4.9}$$

then there exists a $u\in W^{1,\infty}(\Omega)$ viscosity solution of (1.1).

Acknowledgments. The author would like to thank Bernard Dacorogna for stimulating discussions and useful suggestions. He also wants to thank the Institute of Mathematics of the EPFL in Lausanne, Switzerland, where part of this work was done.

REFERENCES

[1] M. BARDI AND I. CAPUZZO-DOLCETTA, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*. Systems and Control: Foundations and Applications, Birkhäuser Boston, Boston, 1997.

[2] G. BARLES, *Solutions de viscosité des équations de Hamilton-Jacobi*, Math. Appl. 17, Springer-Verlag, Paris, 1994.

[3] E. N. Barron and R. Jensen, *Semicontinuous viscosity solutions for Hamilton-Jacobi equations with convex Hamiltonians*, Comm. Partial Differential Equations, 15 (1990), pp. 1713–1742.
[4] G. Bouligand, *Sur la semi-continuité d'inclusions et quelques sujets connexes,* Enseign. Math., 31 (1932), pp. 11–22.
[5] M. Born and E. Wolf, *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light*, 3rd ed., Pergamon Press, Oxford, UK, 1965.
[6] P. Cardaliaguet, B. Dacorogna, W. Gangbo, and N. Georgy, *Geometric restrictions for the existence of viscosity solutions*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 16 (1999), pp. 189–220.
[7] F. H. Clarke, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247–262.
[8] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Canadian Mathematical Society Series of Monographs and Advanced Texts, John Wiley and Sons, New York, 1983.
[9] M. G. Crandall, L. C. Evans, and P.-L. Lions, *Some properties of viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487–502.
[10] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.
[11] M. G. Crandall and P.-L. Lions, *Viscosity solutions of Hamilton-Jacobi equations,* Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.
[12] B. Dacorogna and P. Marcellini, *Sur le problème de Cauchy-Dirichlet pour les systèmes d'équations non linéaires du premier ordre*, C. R. Acad. Sci. Paris Sér. I Math., 323 (1996), pp. 599–602.
[13] B. Dacorogna and P. Marcellini, *Théorèmes d'existence dans les cas scalaire et vectoriel pour les équations de Hamilton-Jacobi*, C. R. Acad. Sci. Paris Sér. I Math., 322 (1996), pp. 237–240.
[14] B. Dacorogna and P. Marcellini, *General existence theorems for Hamilton-Jacobi equations in the scalar and vectorial cases*, Acta Math., 178 (1997), pp. 1–37.
[15] B. Dacorogna and P. Marcellini, *Implicit Partial Differential Equations*, Progr. Nonlinear Differential Equations Appl. 37, Birkhäuser Boston, Boston, 1999.
[16] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions,* Studies in Advanced Mathematics, CRC Press, Boca Raton, FL, 1992.
[17] H. Frankowska, *Hamilton-Jacobi equations: Viscosity solutions and generalized gradients,* J. Math. Anal. Appl., 141 (1989), pp. 21–26.
[18] N. Georgy, *Equations de type implicite du premier ordre*, Ph.D. thesis, EPFL, Lausanne, Switzerland, 1999.
[19] M. R. Hestenes, *Optimization Theory. The Finite Dimensional Case. Pure and Applied Mathematics*, Wiley-Interscience (John Wiley and Sons), New York, 1975.
[20] S. N. Kružkov, *Generalized solutions of Hamilton-Jacobi equations of eikonal type.* I. *Statement of the problems; existence, uniqueness and stability theorems; certain properties of the solutions*, Mat. Sb. (N.S.), 98 (1975), pp. 450–493; 496. (In Russian.)
[21] P.-L. Lions, *Generalized Solutions of Hamilton-Jacobi Equations*, Res. Notes Math. 69, Pitman, Boston, 1982.
[22] R. T. Rockafellar, *La théorie des sous-gradients et ses applications à l'optimisation: Fonctions convexes et non convexes*, Presses de l'Université de Montréal, Montreal, Quebec, Canada, 1979.
[23] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Grundlehren Math. Wiss. 317, Springer-Verlag, Berlin, 1998.
[24] S. Saks, *Theory of the Integral*, 2nd ed., Dover, New York, 1964.

SIAM J. MATH. ANAL.
Vol. 36, No. 1, pp. 204–213

ON SQUIRT SINGULARITIES IN HYDRODYNAMICS*

DIEGO CÓRDOBA†, CHARLES FEFFERMAN‡, AND RAFAEL DE LA LLAVE§

Abstract. We consider certain singularities of hydrodynamic equations that have been proposed in the literature. We present a kinematic argument that shows that if a volume preserving field presents these singularities, certain integrals related to the vector field have to diverge. We also show that if the vector fields satisfy certain partial differential equations (Navier–Stokes, Boussinesq), then the integrals have to be finite. As a consequence, these singularities are absent in the solutions of the above equations.

Key words. singularities, Boussinesq equations, Navier–Stokes equations

AMS subject classifications. 76D03, 76D05, 35Q35

DOI. 10.1137/S0036141003424095

1. Introduction.

One way to make progress towards settling the question of existence of singularities in incompressible fluid motion is to conjecture plausible scenarios for the formation of singularities supported by numerical evidence. Then, it becomes a natural object to develop mathematically rigorous arguments that derive quantitative consequences of the different scenarios and possibly show that these singularities cannot occur in solutions of hydrodynamic equations.

In this note we introduce some classes of singularities which we call *"squirt"* singularities in which some portion of material is ejected from a set of positive measure. These squirt singularities include as particular cases several other singularities that had been considered in the literature (for example, the *"potato chip"* singularities, the *"saddle collapse,"* and the *"tube collapse"*; see section 2 for precise definitions).

In section 3 we present a very simple argument that shows that if a volume preserving vector field u presents a squirt singularity at time T, then

$$\int_0^T ||u||_{L^\infty}\, dt = \infty. \tag{1.1}$$

In section 4 we show that if the vector field satisfies certain partial differential equations (e.g., Navier–Stokes in two and three dimensions, Boussinesq equations in two and three dimensions with positive viscosity), then

$$\int_0^T ||u||_{L^\infty}\, dt < \infty. \tag{1.2}$$

As a consequence of the results in sections 3 and 4, we conclude that volume preserving vector fields satisfying the partial differential equations considered in section 4 do not experience any of the squirt singularities.

*Received by the editors March 11, 2003; accepted for publication (in revised form) November 14, 2003; published electronically June 22, 2004.
http://www.siam.org/journals/sima/36-1/42409.html

†IMAFF, Consejo Superior de Investigaciones Científicas, Madrid, 28006, Spain (dcg@imaff.cfmac.csic.es). The work of this author was partially supported by Ministerio de Ciencia y Tecnología, BFM2002-02042.

‡Department of Mathematics, Princeton University, Princeton, NJ 08540 (cf@math.princeton.edu). The work of this author was supported by the NSF.

§Department of Mathematics, University of Texas at Austin, Austin, TX 78712-1802 (llave@math.utexas.edu). The work of this author was supported by the NSF.

The above results include as a particular case a partial answer to a question proposed by Moffatt in [20]. We show that if a two-dimensional fluid satisfies the Boussinesq equation describing a fluid moving under buoyancy forces with positive fluid viscosity $\nu > 0$, but possibly with zero molecular diffusivity $\kappa = 0$, then it cannot have a saddle collapse.

Even if the arguments presented here exclude that the singularities happen, they give little information on how fast the singular terms may grow. In some of the cases discussed here, these more quantitative arguments are available in the literature (see [8]). They, of course, require using more heavily the details of the equation and the singularity.

2. Squirt singularities. In this section we collect the definitions of the different types of singularities that we will be considering in this paper.

2.1. Notation. We denote the Lebesgue measure of a set A by $|A|$ and the ball centered at $\mathbf{x}^0$ with radius r by $B_r(\mathbf{x}^0)$.

Let $\Omega \subset \mathbb{R}^n$ be an open set. We consider a C^1 time dependent vector field $u : \Omega \times [0, T) \to \mathbb{R}^n$.

This vector field defines an evolution for trajectories $\Phi_t(x)$, where $\Phi_t(x)$ denotes the position at time t of the trajectory with initial condition x at time $t = 0$. More generally, we denote by $\Phi_{t,a}(x)$ the position at time t of the trajectory which at time $t = a$ is in x. Note that, when both sides of the formulas make sense, $\Phi_t(x) = \Phi_{t,0}(x)$, $\Phi_{t,a} = \Phi_t \circ \Phi_a^{-1}(x)$, $\Phi_{t,a} \circ \Phi_{a,b}(x) = \Phi_{t,b}(x)$.

For $\mathcal{S} \subset \Omega$, we denote by

$$\Phi_{t,a}^{\Omega}\mathcal{S} = \{x \in \Omega \mid x = \Phi_t(y),\ y \in \mathcal{S},\ \Phi_s(y) \in \Omega,\ 0 \le s \le t\}.$$

That is, $\Phi_{t,a}^{\Omega}$ is the evolution of the set $\mathcal{S}$, starting at time a, after we eliminate the trajectories which step out of Ω at some time.

We will henceforth assume that u is divergence free. Given the fact that u has zero divergence, we have that $|\Phi_{t,s}\mathcal{S}|$ is independent of t and $|\Phi_{t,a}^{\Omega}\mathcal{S}|$ is nonincreasing in t.

2.2. Definition of singularities.

2.2.1. Squirt singularities. The following definition will be the hypothesis of the main kinematic result of this paper, Theorem 3.1.

DEFINITION 1. *Let Ω_-, Ω_+ be open and bounded sets. $\overline{\Omega_-} \subset \Omega_+$. Therefore,* $\operatorname{dist}(\Omega_-, \mathbb{R}^d - \Omega_+) \ge r > 0$.

We say that u experiences a squirt singularity in Ω_-, at time $T > 0$, when for every $0 \le s < T$, we can find a set $\mathcal{S}_s \subset \Omega_+$ such that

- *$\mathcal{S}_s \cap \Omega_-$ has positive measure, $0 \le s < T$,*
- $\lim_{t \to T} |\Phi_{t,s}^{\Omega_+}\mathcal{S}_s| = 0$.

The physical intuition is that there is a region of positive volume so that all the fluid occupying it gets ejected from a slightly bigger region in a finite time.

As we see in the following subsections, Definition 1 includes as particular cases other singularities that have been considered in the literature. Of course the conclusions of Theorem 3.1, which uses only Definition 1 as hypothesis, are a fortiori valid when we use as hypothesis the existence of the other singularities that we now formulate.

2.2.2. Potato chip singularities.

DEFINITION 2. *We say that u experiences a potato chip singularity when we can find continuous functions*

$$f_\pm : \mathbb{R}^{n-1} \times [0,T) \to \mathbb{R}$$

such that

$$\begin{aligned} &f_+(x_1,\dots,x_{n-1},t) \geq f_-(x_1,\dots,x_{n-1},t), \quad t \in [0,T], x_1,\dots,x_{n-1} \in B_{2r}(\Pi \mathbf{x}^0),\\ &f_+(x_1,\dots,x_{n-1},0) > f_-(x_1,\dots,x_{n-1},0), \quad x_1,\dots,x_{n-1} \in B_r(\Pi \mathbf{x}^0),\\ &\lim_{t\to T^-} (f_+(x_1,\dots,x_{n-1},t) - f_-(x_1,\dots,x_{n-1},t)) = 0 \quad \forall \quad x_1,\dots,x_{n-1} \in B_{2r}(\Pi \mathbf{x}^0) \end{aligned}$$

and such that the surfaces

$$\Sigma_{\pm,t} = \{x_n = f_\pm(x_1,\dots,x_{n-1},t)\} \subset \Omega$$

are transformed into each other by the flow

$$\Phi_t(\Sigma_{\pm,0}) \supset \Sigma_{\pm,t} \ .$$

Note that in Definition 2 we are not requiring that the functions are C^1 as was done in [12]. For us, it suffices that $f_\pm$ are continuous. That is, we allow the singularities to be ruffled potato chips. Since the arguments we will present in section 3 do not depend on calculus identities, there is no need for the boundaries of the sets to be differentiable.

If we denote by $\mathcal{S}_t^{(f)} = \{x \mid f_-(x_1,t) \leq x_n \leq f_+(x_1,t)\}$, we have, by the intermediate value theorem and the continuity of the trajectories

$$\Phi_{t,s}^{B_r(\mathbf{x}^0)}(\mathcal{S}_s^{(f)}) \subset \mathcal{S}_t. \tag{2.1}$$

In particular, if $f_\pm$ verify Definition 2, then $\mathcal{S}_s^{(f)}$ verifies the assumptions of Definition 1.

Hence, if a system satisfies Definition 2, it also satisfies Definition 1.

Potato chip singularities were introduced as a conjectural mechanism (see [17] and [19]) of singularities for a three-dimensional ideal magnetohydrodymamic flow in which two linked flux rings approach each other forming two-dimensional current sheets. In two-dimensional cases, similar singularities were proposed by [21], [22].

The two-dimensional potato chip singularities were considered in [9], where they were called *"sharp fronts."* Using calculus identities and the fact that the fluid admits a stream function representation, it was shown that if a sharp front exits, then (1.1) holds. This result was generalized to three dimensions in [12]. Both of these results follow from Theorem 3.1.

2.2.3. Tube collapse singularities. The following definition appears for the case $n = 3$ in [11]. In the case $d = 2$ the concept was introduced in [9], [10].

Let $I_i \subset \mathbb{R}$, $i = 1,\dots,n$, be bounded intervals. Let $Q = \times_i I_i \subset \mathbb{R}^n$ be a cube.

DEFINITION 3. *A regular tube is a relatively open set $\mathcal{S} \subset Q$, characterized as*

$$\mathcal{S} = \{x \in Q \mid f(x) < 0\}, \tag{2.2}$$

where $f : Q \to \mathbb{R}$ is a C^1 function that satisfies

$$f(x) = 0 \implies \nabla_{x_1,\dots,x_{n-1}} f \neq 0.$$

For every $x_n \in I_n$, the set

$$\mathcal{S}(x_n) = \mathcal{S} \cap I_1 \times \cdots \times I_{n-1} \times \{x_n\}$$

is nonempty, and its closure is contained in the interior of $I_1 \times \cdots \times I_{n-1} \times \{x_n\}$.

We will also consider the situation when f_t is a family of functions indexed by time $t \in [0, T)$.

DEFINITION 4. *We say that the vector field u experiences a tube collapse singularity at time T when the boundaries of the tube evolve with the velocity field u and* $\liminf_{t\to T} |\mathcal{S}_t| = 0$.

An example worth keeping in mind is when $f_t(x) = \operatorname{dist}(x, \gamma) + r(t)$, where γ is a curve, dist denotes the distance, and $r(t) \to 0$ as $t \to 0$. S_t is the set of points which are at a distance less than $r(t)$ from the curve γ. (Of course, we could let the curve γ depend on time, provided that it does not become too pathological.)

Again, we point out that Definition 4 implies Definition 1. We can take $\Omega_- = \times_{i=1,\dots,n-1} I_i \times J$, $\Omega_- = \times_{i=1,\dots,n-1} I_i \times I_d$, where $J \subset I_d$ is an interval contained in the interior of I_d.

2.2.4. Saddle collapse singularity. This singularity is specific for two-dimensional flows. We follow the definition in [8]. We refer to that paper for a comparison with alternative definitions in the literature.

DEFINITION 5. *We consider foliations of a neighborhood of the origin (with coordinates x_1, x_2) whose leaves are given by equations of the form*

$$\rho \equiv (y_1\beta(t) + y_2) \cdot (y_1\delta(t) + y_2) = \text{cte} \tag{2.3}$$

and $(y_1, y_2) = F_t(x_1, x_2)$, where $\beta, \delta : [0, T) \to \mathbb{R}^+$ are C^1 functions, F is a C^2 function of x, t, for a fixed t, and F_t is an orientation preserving diffeomorphism.

We say that the foliation experiences a saddle collapse when

$$\liminf_{t\to T} \beta(t) + \delta(t) = 0.$$

If the leaves of the foliation are transported by a vector field u, we say that the vector field u experiences a saddle collapse.

If we take as $\Omega_\pm$ balls centered at $F(0, 0, T)$ and as the set $\mathcal{S}_s$ a connected component of the set $\rho < 0$, we see that Definition 5 implies Definition 1.

3. Kinematic arguments. The main result of this section follows.

THEOREM 3.1. *If u as before has a squirt singularity, then*

$$\int_s^T \sup_x |u(x, t)|\, dt = \infty \quad \forall s \in (0, T). \tag{3.1}$$

Moreover, if u has a potato chip singularity, then

$$\int_s^T \sup_x |\Pi u(x, t)|\, dt = \infty, \tag{3.2}$$

where Π *is the projection on the first* $n-1$ *coordinates.*

REMARK 1. *We note that in the argument for Theorem* 3.1*, some of the hypotheses can be somewhat weakened.*

For example, using the theory of [13]*, the hypothesis that* $u \in C^1$ *can be weakened to* $u \in H^1$*.*

We also note that strict volume preservation is not needed. It suffices that the volume contraction remains bounded. That is, for some constant $C \geq 1$ *and all* $M \subset \mathbb{R}^n$ *measurable,* $C^{-1}|M| \leq |\Phi_t(M)| \leq C|M|$*.*

REMARK 2. *We note that if* $u(x,t)$ *experiences a squirt singularity at* $t = T$ *and* $\Gamma : [0,T) \hookleftarrow [0,T)$ *is a reparameterization, then*

$$\tilde{u}(x,t) = u(x,\Gamma(t))\Gamma'(t)$$

also has a potato chip singularity.

It is reassuring to note that the conclusions of Theorem 3.1 *remain true for* $\tilde{u}$*. But the observation that the existence of potato chip singularities is invariant under time reparameterizations shows that, with the present assumptions, one cannot obtain more precise rates of the blow-up of* $\sup_x |u(x,t)|$ *than* (3.1)*.*

In case we assume that singularities are somewhat more uniform, it is possible to develop more quantitative information about the rates of collapse.

Roughly speaking, we just need to assume that the exit area of the set $\mathcal{S}_s$ controls the volume of the set.

For example, in potato chip singularities (Definition 2), we say that the collapse is uniform when

$$\max_{x_1,x_2}(f_+(x_1,x_2,t) - f_-(x_1,x_2,t)) \leq M \min_{x_1,x_2}(f_+(x_1,x_2,t) - f_-(x_1,x_2,t)),$$

where M is a constant independent of time.

In tube collapse singularities (Definitions 4 and 3) we say that the collapse is uniform when

$$\max |S(x_n)|_{n-1} \leq M \min |S(x_n)|,$$

where M is a constant independent of time and $|\cdot|_{n-1}$ denotes the $n-1$ dimensional area.

Given a S_t a C^1 set, we denote by $\tilde{\partial} S_t$ the portion of the boundary which is not evolving with the fluid.

We note that by zero divergence of the fluid the change of volume is the integral of u over $\tilde{\partial} S_t$. Hence, we always have

$$\frac{d}{dt}|\mathcal{S}_t| \geq -||u||_{L^\infty} |\tilde{\partial}\mathcal{S}_t|_{n-1}.$$

In the uniform cases, we have

$$\frac{d}{dt}|\mathcal{S}_t| \geq -M||u||_{L^\infty} |\mathcal{S}_t|.$$

Integrating the above equation we have

$$|\mathcal{S}_t| \geq |\mathcal{S}_0| \exp\left(-M \int_0^t ||u(s)||_{L^\infty}\, ds\right)$$

so that uniform collapses cannot happen too fast.

3.1. Proof of Theorem 3.1. From the assumption that $|\Phi_{T,s}^{\Omega_+}\mathcal{S}_s| \to 0$, we conclude that almost all the trajectories starting in $\mathcal{S}_s$ at time s leave the set Ω_+ at a time $\tau \in (s,T)$.

Therefore, we conclude that for any trajectory $x(t)$ starting in $\Omega_- \cap \mathcal{S}_s$ at time s we have

$$\left|\int_s^\tau u(\Phi_t(x),t)\,dt\right| \geq r > 0. \tag{3.3}$$

Therefore,

$$\int_s^T \sup_x |u(x,t)|\,dt \geq r > 0. \tag{3.4}$$

Since (3.4) holds for every $s \in (0,T)$ we conclude that (3.1) holds.

To establish (3.2) we observe that in the classical potato chip singularity, since the escape can happen only by increasing the $n-1$ first components, we can sharpen (3.3) to

$$\left|\int_s^T \Pi u(\Phi_t(x),t)\,dt\right| \geq r/2$$

again for all $s \in [0,T)$.

4. A priori bounds. In this section, we show how if the vector field u satisfies certain partial differential equations, then (1.2) holds. By Theorem 3.1, we conclude immediately that these equations do not exhibit any of the singularities considered in Definition 1.

We consider two- and three-dimensional Boussinesq equations and Navier–Stokes equations in three dimensions.

We note that the results on two-dimensional Boussinesq equations are closely related to the problem proposed by Moffatt [20]:

XXI Century Problem 3: *The problem is to examine the evolution of the θ-field for Boussinesq equations (see* [1]*) in the neighborhood of its saddle points, to determine whether singularities of $\nabla\theta$ can develop, and to examine the influence of weak molecular diffusivity k in controlling the approach to such singularities.*

We show that the saddle collapse singularities and similar ones cannot occur. We do not exclude the possibility that singularities other than squirt singularities could also occur. For example, the argument presented here does not exclude singularities in which the surfaces evolve until they touch at just one point other than the origin.

The proof of (1.2) for the case of Navier–Stokes equations has been in the literature for a long time. See, for example, [16], [23] and the exposition in [14], where it is called the *"second F_N ladder."* Hence, we will present only the proof in the case of the Boussinesq equation in two and three dimensions.

The proofs are based on very elementary arguments, basically, integration by parts without boundary terms, Sobolev inequalities and interpolation inequalities. Hence, they remain valid for all the boundary conditions that allow us to carry out these operations. These include problems defined in the whole space and in a bounded domain with periodic as well as Neumann and Dirichlet conditions with respect to appropriate fields or their gradients. We will henceforth assume that the boundary conditions are such that they allow integration by parts without boundary terms.

4.1. Two-dimensional Boussinesq equations. The Boussinesq equations are

$$\frac{\partial u}{\partial t} + u \cdot \nabla u = -\nabla p + \nu \Delta u + (0, \theta), \tag{4.1}$$

$$\nabla \cdot u = 0, \tag{4.2}$$

$$(\partial_t + u \cdot \nabla)\, \theta = \kappa \Delta \theta, \tag{4.3}$$

with $u = (u_1, u_2)$, $x = (x_1, x_2) \in \mathbb{R}^2$ or $\mathbb{R}^2/\mathbb{Z}^2$, and finite energy at initial time.

The Cauchy problem for the system (4.1), (4.2), and (4.3) has been extensively studied in the literature; see [5], [18], and [24]. In the case $\kappa > 0$ it is known that the equation does not develop singularities in finite time. In order to study the evolution of the level sets of θ it is reasonable to take $\kappa = 0$, where the collapse of the saddle would produce a singularity on $\nabla\theta$. This is a two-dimensional potato chip singularity; for more details, see [9].

In [6], [7], and [15] the two-dimensional Boussinesq convection in the absence of viscous effects was studied numerically and analytically.

THEOREM 4.1. *If u satisfies the two-dimensional Boussinesq equation with $\nu > 0$ and $||\theta(0)||_{L^2} \le A < \infty$, then* (1.2) *holds.*

In particular, using Theorem 3.1*, u does not exhibit any singularity satisfying Definition* 1.

Proof. We denote by C_1, C_2 constants that depend only on ν, A and the initial conditions. In particular, they can change the meaning from line to line.

From (4.3) we obtain that the L^p norms $p \ge 1$ are nonincreasing—they are conserved if $\kappa = 0$:

$$\|\theta(\cdot, t)\|_{L^p} \le \|\theta(\cdot, 0)\|_{L^p} \quad \text{for } 1 \le p \le \infty \quad \forall t \ge 0.$$

Taking the curl of (4.1) of the velocity field we get

$$(\partial_t + u \cdot \nabla)\, \omega = \theta_{x_1} + \nu \Delta \omega, \tag{4.4}$$

where $\omega = \operatorname{curl}(u)$.

We multiply (4.4) by ω and integrate by parts to obtain

$$\frac{1}{2}\frac{d}{dt}\int |\omega|^2\, dx + \nu \int |\nabla \omega|^2 dx = \int \omega \theta_{x_1} dx. \tag{4.5}$$

Integration by parts and the Hölder inequality gives

$$\frac{1}{2}\frac{d}{dt}\int |\omega|^2\, dx + \nu ||\nabla \omega||_{L^2}^2 \le ||\nabla \omega||_{L^2} ||\theta||_{L^2}.$$

This implies that

$$\frac{d}{dt}\int |\omega|^2\, dx \le C_1,$$

and therefore $||\omega||_{L^2} \le C_1 t + C_2$.

Substituting this into (4.5) gives

$$\nu ||\nabla \omega||_{L^2}^2 \le A ||\nabla \omega||_{L^2} - \frac{1}{2}\frac{d}{dt}||\omega||_{L^2}.$$

An integration with respect to time and a Hölder inequality for the integration with respect to time yields

$$\nu \int_0^t ds\, ||\nabla\omega(s)||_{L^2}^2 \le A \int_0^t ds\, ||\nabla\omega(s)||_{L^2} - \frac{1}{2}||\omega(t)||_{L^2}^2 + \frac{1}{2}||\omega(0)||_{L^2}^2$$
$$\le At^{1/2}\left(\int_0^t ds\, ||\nabla\omega(s)||_{L^2}^2\right)^{1/2} + \frac{1}{2}||\omega(0)||_{L^2}^2.$$

This yields

$$\int_0^t ds\, ||\nabla\omega(s)||_{L^2}^2 \le C_1 t + C_2, \tag{4.6}$$

and, using Hölder inequality again,

$$\int_0^t ds\, ||\nabla\omega(s)||_{L^2} \le C_1 t + C_2. \tag{4.7}$$

The well-known Biot–Savart law recovers the velocity field from the vorticity by the integral operator

$$u(x,t) = \frac{1}{2\pi}\int K(x-y)\omega(y,t)dy,$$

with $K(x) = (-\frac{x_2}{x_1^2+x_2^2}, \frac{x_1}{x_1^2+x_2^2})$ for $x \in \mathbb{R}^2$, and a similar formula holds for $\mathbb{R}^2/\mathbb{Z}^2$.

Furthermore, ∇u is a singular integral operator of ω, and Δu is a singular integral operator of $\nabla\omega$ (for details, see [3]). From the classical Calderon–Zygmund theory we have

$$\|\nabla u\|_{L^2} \le C\|\omega\|_{L^2}, \qquad \|\Delta u\|_{L^2} \le C\|\nabla\omega\|_{L^2}. \tag{4.8}$$

Combining estimates (4.6), (4.7), and (4.8) and using Sobolev inequalities we finally get

$$\int_0^t \|u\|_{L^\infty} ds \le C\int_0^t (\|u\|_{L^2} + \|\Delta u\|_{L^2})ds$$
$$\le C_1 t + C_2. \qquad \square$$

REMARK 3. *The argument above works for $\nu > 0$. For $\nu = 0$ we do not have control on any norm of the derivatives of the vorticity.*

4.2. Three-dimensional Boussinesq equations. In this section we adapt Theorem 4.1 to three dimensions.

THEOREM 4.2. *If u satisfies the three-dimensional Boussinesq equation with $\nu > 0$ and $||\theta_0||_{L^2} < \infty$, then* (1.2) *holds.*

In particular, using Theorem 3.1*, u does not exhibit any singularity satisfying Definition* 1.

Compared with the proof of Theorem 4.1, the proof of Theorem 4.2 requires an extra estimate on the nonlinear term that appears on the vorticity equation. Below we give the argument which is based on the argument in [16] for Navier–Stokes.

By the usual integration by parts

$$\frac{1}{2}\frac{d}{dt}\int |u|^2\, dx + \nu\int |\nabla u|^2 dx \le C\int |u\theta| dx.$$

Therefore, proceeding as before,

$$\|u\|_{L^2}^2 \le C_1 t + C_2,$$
$$\int_0^t \|\nabla u\|_{L^2}^2 ds \le \tilde{C}_1 t + \tilde{C}_2. \tag{4.9}$$

The vorticity equation is

$$(\partial_t + u \cdot \nabla)\, \omega = \omega \cdot \nabla u + \theta_{x_1} - \theta_{x_2} + \nu \Delta \omega.$$

Multiply the vorticity equation by ω and integrate by parts

$$\frac{1}{2}\frac{d}{dt}\int |\omega|^2\, dx + \nu \int |\nabla \omega|^2 dx \le \int |(\omega \cdot \nabla u)\omega| dx + \frac{1}{2\nu}\int |\theta|^2\, dx + \frac{\nu}{2}\int |\nabla \omega|^2 dx.$$

The nonlinear term can be bounded by (see [16])

$$\begin{aligned}\int |(\omega \cdot \nabla u)\omega| dx &\le C \|\omega\|_{L^2}^{\frac{3}{2}} \|\nabla \omega\|_{L^2}^{\frac{3}{2}} \\ &\le \tilde{C}\|\omega\|_{L^2}^6 + \frac{\nu}{4}\|\nabla \omega\|_{L^2}^2;\end{aligned}$$

then

$$\frac{1}{2}\frac{d}{dt}\int |\omega|^2\, dx + \frac{\nu}{4}\int |\nabla \omega|^2 dx \le \tilde{C}(1 + \|\omega\|_{L^2}^6)$$

and

$$\frac{\frac{1}{2}\frac{d}{dt}\|\omega\|_{L^2}^2}{(1+\|\omega\|_{L^2}^2)^2} + \nu \frac{\|\nabla \omega\|_{L^2}^2}{(1+\|\omega\|_{L^2}^2)^2} \le \tilde{C}(1 + \|\omega\|_{L^2}^2),$$

and we get

$$\int_0^t \frac{\|\nabla \omega\|_{L^2}^2}{(1+\|\omega\|_{L^2}^2)^2} ds \le \tilde{\tilde{C}}(1+t). \tag{4.10}$$

Finally, we estimate $\int_0^t \|u\|_{L^\infty} ds$ applying Sobolev inequalities, Calderon–Zygmund theory, (4.9), and (4.10):

$$\begin{aligned}\int_0^t \|u\|_{L^\infty} ds &\le C \int_0^t \|\nabla u\|_{L^2}^{\frac{1}{2}} \|\Delta u\|_{L^2}^{\frac{1}{2}} ds \\ &\le C\left(\int_0^t \|\nabla u\|_{L^2}^2 ds + \int_0^t \|\Delta u\|_{L^2}^{\frac{2}{3}} ds\right) \\ &\le C\left[\int_0^t \|\omega\|_{L^2}^2 ds + \left(\int_0^t \frac{\|\nabla \omega\|_{L^2}^2}{(1+\|\omega\|_{L^2}^2)^2} ds\right)^{\frac{1}{3}} \left(\int_0^t (1+\|\omega\|_{L^2}^2) ds\right)^{\frac{2}{3}}\right] \\ &\le C(1+t),\end{aligned}$$

where C depends on the initial data and on the viscosity.

Acknowledgment. We thank W. Strauss for several comments.

REFERENCES

[1] G.K. Batchelor, V.M. Canuto, and J.R. Chasnov, *Homogeneous buoyancy generated turbulence,* J. Fluid Mech., 212 (1990), pp. 337–363.
[2] J.T. Beale, T. Kato, and A. Majda, *Remarks on the breakdown of smooth solutions for the* 3*D Euler equations,* Comm. Math. Phys., 94 (1984), pp. 61–64.
[3] A.L. Bertozzi and A.J. Majda, *Vorticity and Incompressible Flow,* Cambridge University Press, Cambridge, UK, 2002.
[4] L. Caffarelli, R. Kohn, and L. Nirenberg, *Partial regularity of suitable weak solutions of the Navier-Stokes equations,* Comm. Pure Appl. Math., 35 (1982), pp. 711–831.
[5] J.R. Cannon and E. Dibenedetto, *The initial problem for the Boussinesq equations with data in* L^p*,* in Approximation Methods for Navier–Stokes Problems, Lecture Notes in Math. 771, Springer, Berlin, 1980, pp. 129–144.
[6] D. Chae and O.Y. Imanuvilov, *Generic solvability of the axisymmetric* 3*-D Euler equations and the* 2*-D Boussinesq equations,* J. Differential Equations, 156 (1999), pp. 1–17.
[7] D. Chae, S.-K. Kim, and H.-S. Nam, *Local existence and blow-up criterion of Holder continuous solutions of the Boussinesq equations,* Nagoya Math. J., 155 (1999), pp. 55–80.
[8] D. Córdoba, *Nonexistence of simple hyperbolic blow-up for the quasi-geostrophic equation,* Ann. of Math. (2), 148 (1998), pp. 1135–1152.
[9] D. Córdoba and C. Fefferman, *Scalars convected by a* 2*D incompressible flow,* Comm. Pure Appl. Math., 55 (2002), pp. 255–260.
[10] D. Córdoba and C. Fefferman, *Behavior of several* 2*D fluid equations in singular scenarios,* Proc. Natl. Acad. Sci. USA, 98 (2001), pp. 4311–4312.
[11] D. Córdoba and C. Fefferman, *On the collapse of tubes carried by* 3*D incompressible flows,* Comm. Math. Phys., 222 (2001), pp. 293–298.
[12] D. Córdoba and C. Fefferman, *Potato chip singularities of* 3*D flows,* SIAM J. Math. Anal., 33 (2001), pp. 786–789.
[13] R.J. Diperna and P.L. Lions, *Ordinary differential equations, transport theory and Sobolev spaces,* Invent. Math., 98 (1989), pp. 511–547.
[14] C.R. Doering and J.D. Gibbon, *Applied Analysis of the Navier Stokes Equations,* Cambridge University Press, Cambridge, UK, 1995.
[15] W. E and C.-W. Shu, *Small-scale structures in Boussinesq convection,* Phys. Fluids, 6 (1994), pp. 49–58.
[16] C. Foias, C. Guillope, and R. Temam, *New a priori estimates for Navier-Stokes equations in dimension* 3, Comm. Partial Differential Equations, 6 (1981), pp. 329–359.
[17] R. Grauer and C. Marliani, *Current sheet formation in* 3*D ideal incompressible magnetohydrodynamics,* Phys. Rev. Lett., 84 (2000), pp. 4850–4853.
[18] B. Guo, *Spectral method for solving two-dimensional Newton-Boussinesq equation,* Acta Math. Appl. Sinica, 5 (1989), pp. 208–218.
[19] R. Kerr and A. Brandenburg, *Evidence for a singularity in ideal magnetohydrodynamics: Implications for fast reconnection,* Phys. Rev. Lett., 83 (1999), pp. 1155–1158.
[20] H.K. Moffatt, *Some remarks on topological fluid mechanics,* in An Introduction to the Geometry and Topology of Fluid Flows, R.L. Ricca, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001, pp. 3–10.
[21] E.N. Parker, *Spontaneous Current Sheets in Magnetic Fields,* Oxford University Press, New York, 1994.
[22] E.R. Priest and V.S. Titov, *Magnetic reconnection at three-dimensional null points,* Philos. Trans. Roy. Soc. London Ser. A, 354 (1996), pp. 2951–2992.
[23] L. Tartar, *Topics in Nonlinear Analysis,* Publ. Math. Orsay 78, Orsay, France, 1978.
[24] R. Temam, *Navier-Stokes Equations, Theory and Numerical Analysis,* North–Holland, Amsterdam, 1984.

SIAM J. MATH. ANAL.
Vol. 36, No. 1, pp. 214–232

A KINETIC FORMULATION FOR MULTIDIMENSIONAL SCALAR CONSERVATION LAWS WITH BOUNDARY CONDITIONS AND APPLICATIONS*

C. IMBERT† AND J. VOVELLE‡

Abstract. We state a kinetic formulation of weak entropy solutions of a general multidimensional scalar conservation law with initial and boundary conditions. We first associate with any weak entropy solution an entropy defect measure; the analysis of this measure at the boundary of the domain relies on the study of weak entropy sub- and supersolutions and implies the introduction of the notion of sided boundary defect measures. As a first application, we prove that any weak entropy subsolution of the initial-boundary value problem is bounded above by any weak entropy supersolution (comparison theorem). We next study a Bhatnagar–Gross–Krook-like kinetic model that approximates the scalar conservation law. We prove that such a model converges by adapting the proof of the comparison theorem.

Key words. conservation law, initial-boundary value problem, boundary defect measures, kinetic traces, weak entropy sub- and supersolutions, comparison theorem, generalized kinetic solutions, Bhatnagar–Gross–Krook-like kinetic model

AMS subject classifications. 35L65, 35B50, 35D99, 35F25, 35F30, 35A35

DOI. 10.1137/S003614100342468X

1. Introduction.

Let Ω be a strong Lipschitz open subset of $\mathbb{R}^d$. Let $\partial\Omega$ denote its boundary, $n(\overline{x})$ denote the outward unit normal to Ω at a point $\overline{x} \in \Omega$, $Q = (0,+\infty) \times \Omega$, and $\Sigma = (0,+\infty) \times \partial\Omega$. We consider the following multidimensional scalar conservation law:

$$\partial_t u + \operatorname{div}_x A(u) = 0 \text{ in } Q, \tag{1.1a}$$

with the initial condition

$$u(0,x) = u_0(x) \quad \forall x \in \Omega \tag{1.1b}$$

and the boundary condition

$$u(s,y) = u_b(s,y) \quad \forall (s,y) \in \Sigma. \tag{1.1c}$$

The first step in the understanding of (1.1c) is the work of Bardos, Le Roux, and Nédélec [1]: they show that if the initial datum u_0 is BV and the boundary datum is C^2-regular, there exists a unique (weak entropy) solution of (1.1). In particular, they show that an inequality must hold at the boundary. This inequality is known as the Bardos–Le Roux–Nédélec (BLN) condition (see (3.19)). Note that the BLN condition makes sense only if the solution u admits a trace on $\partial\Omega$. In the case of the Cauchy problem with merely essentially bounded (L^∞) data, some notions of a generalized solution have been defined. The measure-valued entropy solutions were introduced by DiPerna [9] and the entropy process solutions by Eymard, Gallouët,

*Received by the editors March 18, 2003; accepted for publication (in revised form) December 12, 2003; published electronically June 22, 2004.
http://www.siam.org/journals/sima/36-1/42468.html
†Laboratoire ACSIOM, Université Montpellier-II, Montpellier, France (imbert@mip.ups-tlse.fr).
‡Laboratoire IRMAR, Antenne de Bretagne de l'ENS Cachan, Rennes, France (Julien.Vovelle@bretagne.ens-cachan.fr).

and Herbin [11]. These notions of a very weak solution are well adapted to the study of the convergence of numerical schemes, and error estimates are also available. In the case of the Cauchy–Dirichlet problem with L^∞ data, Otto [25] proposed a notion of *weak entropy solution* $u \in \mathsf{L}^\infty(Q)$, relying on the notion of boundary entropy-flux pairs. An equivalent definition can be given by using "Kružkov semientropies" (see [8, 30, 34, 17]). An accurate notion of an entropy process solution can be given in order to prove the convergence of certain numerical methods [34], but it does not seem possible to get an error estimate with respect to the approximation by vanishing viscosity, for example. In order to fill this gap, we follow the ideas developed by Lions, Perthame, and Tadmor [18]. Their heuristic idea, which is, in part, a continuation of the works of Brenier [7] and Di Perna [9], is to take into account the decrease of the entropy by introducing an "entropy defect" measure. More precisely, a kinetic function f is associated with the macroscopic function u by setting

$$f(t,x,\xi) = \begin{cases} 1 & \text{if } 0 < \xi < u(t,x), \\ -1 & \text{if } u(t,x) < \xi < 0, \\ 0 & \text{otherwise.} \end{cases} \tag{1.2}$$

Such a kinetic function is a so-called equilibrium function. The kinetic formulation of Lions, Perthame, and Tadmor states that u is a weak entropy solution of the conservation law if and only if there exists a bounded nonnegative measure m such that

$$(\partial_t + a.\nabla_x) f = \partial_\xi m \text{ in } \mathcal{D}'((0,T) \times \mathbb{R}^d \times \mathbb{R}). \tag{1.3}$$

Next, Perthame [27] showed that these techniques supply a good technical framework to easily prove, for instance, the L^1-contraction property and the error estimate with respect to the parabolic approximation, without relying on the dedoubling variable technique.

We start from [27] and develop analogous techniques for a conservation law with boundary conditions. The main difficulty is to study how the weak entropy solution u and the defect measure m behave at the boundary of the domain. We handle this difficulty by considering the space kinetic trace f^τ of the kinetic function f [32, 33]. As far as the defect measure is concerned, two nonnegative measures $m^b_\pm$ supported by $\Sigma \times \mathbb{R}_\xi$ must therefore be considered. They are characterized by the formula

$$(-a \cdot n) f^\tau = M f^b + (-a \cdot n)\mathrm{sgn}_\mp + \partial_\xi m^b_\pm, \tag{1.4}$$

where the constant M is a Lipschitz constant of the flux A on a compact subset of $\mathbb{R}$ in which the data u_0 and u^b, which are supposed to be measurable essentially bounded functions, take a.e. their values (see section 2). Relation (1.4) can be understood as a kinetic analogue of the BLN condition.[1] Why do we need two nonnegative measures to describe the behavior of the entropy defect measure at the boundary? It is because the notion of weak entropy solution is "sided." Let us be more specific. We define weak entropy sub- and supersolutions for the initial-boundary value problem and give a kinetic formulation of them. Hence two different defect measures $m_\pm$ are a priori associated with each weak entropy solution. But, eventually, we prove they coincide in $Q \times \mathbb{R}_\xi$ and can be different at the boundary. Notions of weak entropy sub- and supersolutions for the Cauchy problem were previously considered [2, 15, 16, 26, 3, 4],

[1] It is a generalization of it in the sense that no strong traces are required; thus merely L^∞ data can be treated.

and comparison principles were established: any weak entropy subsolution of the Cauchy problem is bounded above by any weak entropy supersolution. Such results have also been proved by Terracina [31] for the initial-boundary value problem in the context of BV solutions. We state and prove an analogous result for the initial-boundary value problem in the context of L^∞ solutions. The L^1-contraction property and the maximum principle follow from it.

We then use our results to study an approximation of the conservation law, namely a kinetic model "à la Bhatnagar, Gross, and Krook" (BGK-like kinetic model for short). It was first introduced by Perthame and Tadmor [29] for the Cauchy problem and adapted by Nouri, Omrane, and Vila [22, 23, 24] to the initial-boundary value problem. Nouri, Omrane, and Vila prove the convergence of the BGK-like kinetic model whenever the data are at equilibrium or not. Here, we restrict our study to the case where the data are at equilibrium and show how, in this framework, the concept of a *generalized kinetic solution* can be used to prove the convergence of the BGK-like kinetic model. Such very weak solutions were introduced by Perthame [28] for the Cauchy problem. They can be viewed as the analogue of the measure-valued solutions of DiPerna [9] or the entropy process solutions of Eymard, Gallouët, and Herbin [11]. The definition of a generalized kinetic solution is based on the following kinetic formulation: instead of considering an equilibrium function, a solution can be a general kinetic function (see sections 2 and 5 for precise definitions). The proof of the comparison theorem is slightly modified in order to prove that there is at most one generalized kinetic solution of (1.1) and that it is in fact a weak entropy solution. Hence, it permits us to easily pass to the limit in the kinetic model.

To conclude this introduction, let us mention the recent work of Ben Moussa and Szepessy [6] in which the concept of measure-valued solution to deal with "very weak solutions" is used, and let us state some other occurrences of "kinetic methods" in the study of first-order problems with boundary conditions [5, 20]; see also [21, 13, 14].

The paper is organized as follows. Section 2 is devoted to notations and assumptions. In section 3, kinetic formulations of weak entropy solutions (Theorem 3.1) and entropy semisolutions (Proposition 3.3) are stated and proved. In particular, kinetic traces and boundary defect measures are constructed and characterized (Proposition 3.4). In section 4, the comparison theorem (Theorem 4.1) is proved. Section 5 is devoted to the study of the BGK-like kinetic model.

Finally, let us mention that in a forthcoming paper [10] we study another approximation of the initial-boundary value problem: the parabolic regularization of the conservation law by an artificial viscosity. We get an error estimate between the entropy solution of the conservation law and the regular solution of the parabolic equation. Even if we adapt once again the proof of the comparison theorem, additional difficulties arise, and the proof is rather long and technical.

2. Preliminaries. We give here some notations, assumptions, and basic properties that are used throughout the paper.

The space $\mathbb{R}^d$ is endowed with its usual Euclidean structure. The scalar product is denoted by $x \cdot y$ and the Euclidean norm by $|x|$. For the sake of clarity, $\mathbb{R}_t$ and $\mathbb{R}_\xi$ denote the lines of reals, respectively, related to the t and ξ variables.

Data. We assume u_0 and u_b to be essentially bounded measurable functions. Let $K > 0$ be a positive constant such that

$$-K \le u_0(x) \le K \text{ for a.e. } x \in \Omega \quad \text{and} \quad -K \le u_b(t,x) \le K \text{ for a.e. } (t,x) \in \Sigma\,.$$

The flux function A is assumed to be locally Lipschitz continuous. Let M be the

Lipschitz constant of the function A restricted to $[-K,K]$, and let $a(\xi)=A'(\xi)$.

Remark 1. We could as well consider the equation $\partial_t u+\operatorname{div}_x(A(t,x,u))=0$. All the results presented in this paper remain valid under the assumption that the function A is locally Lipschitz continuous with respect to $(t,x)\in[0,T]\times\overline{\Omega}$ uniformly with respect to the u variable, while for every $u\in\mathbb{R}$, $(t,x)\mapsto A(t,x,u)$ is in $\mathsf{C}^1([0,T]\times\overline{\Omega})$.

Kružkov semientropies. Define

$$\operatorname{sgn}_+(\xi)=\begin{cases}1 & \text{if} \quad \xi>0,\\ 0 & \text{if} \quad \xi\le 0\end{cases} \qquad \text{and} \quad \operatorname{sgn}_-(\xi)=\begin{cases}-1 & \text{if} \quad \xi<0,\\ 0 & \text{if} \quad \xi\ge 0\end{cases}$$

and $\xi^{\pm}=\operatorname{sgn}_{\pm}(\xi)\xi$. Let $a\top b$ denote $\max\{a,b\}$, and let $a\bot b$ denote $\min\{a,b\}$. The Kružkov semientropies are the convex functions $u\mapsto(u-\kappa)^{\pm}$ for $\kappa\in\mathbb{R}$. The corresponding entropy fluxes are given by the formula

$$\mathcal{F}^{\pm}(u,\kappa)=\operatorname{sgn}_{\pm}(u-\kappa)(A(u)-A(\kappa)).$$

Kinetic and equilibrium functions. We previously recalled what an equilibrium function is (see (1.2)). More generally, a kinetic function is a function $f(t,x,\xi)$ such that

$$\begin{aligned}&0\le f(t,x,\xi)\operatorname{sgn}(\xi)\le 1,\\ &\partial_\xi f(t,x,\xi)=\delta(\xi)-\nu_{t,x}(\xi),\end{aligned} \tag{2.1}$$

where ν is a Young measure. For an equilibrium function, $\nu_{t,x}(\xi)=\delta(\xi-u(t,x))$. In the following, we also consider two functions associated with any kinetic one:

$$\begin{aligned}f_+(t,x,\xi)&=f(t,x,\xi)-\operatorname{sgn}_-(\xi),\\ f_-(t,x,\xi)&=f(t,x,\xi)-\operatorname{sgn}_+(\xi).\end{aligned}$$

Notice that $\partial_\xi f_\pm=-\nu_{t,x}(\xi)$ and that these functions no longer have a bounded support with respect to the kinetic variable ξ. Nevertheless, and it is essential, there exists $\kappa\in\mathbb{R}_\xi$ such that $f_+(t,x,\xi)=0$ if $\xi\ge\kappa$, and there exists $\kappa'\in\mathbb{R}_\xi$ such that $f_-(t,x,\xi)=0$ if $\xi\le\kappa'$. We simply say that f_+ vanishes for $\xi\gg 1$ and f_+ vanishes for $\xi\ll -1$. For equilibrium functions, if (t,x) is fixed, then for a.e. $\xi\in\mathbb{R}_\xi$,

$$\begin{aligned}f_+(t,x,\xi)&=\operatorname{sgn}_+(u(t,x)-\xi),\\ f_-(t,x,\xi)&=\operatorname{sgn}_-(u(t,x)-\xi).\end{aligned}$$

Localization. The set Ω is assumed to be a strong Lipschitz open subset of $\mathbb{R}^d$, which means that, locally, Ω can be represented as the epigraph of a Lipschitz continuous function. More precisely, there exists a locally finite open cover $\{B_{\lambda_i}\}_{i\in I}$ of $\overline{\Omega}$ and a partition of unity $\{\lambda_i\}_{i\in I}$ of $\overline{\Omega}$ subordinate to $\{B_{\lambda_i}\}_{i\in I}$ such that for any λ,

$$\begin{aligned}\Omega_\lambda&:=\Omega\cap B_\lambda=\{x\in B_\lambda\ ;\ (A_\lambda\, x)_d>h_\lambda(\overline{A_\lambda\, x})\},\\ \partial\Omega_\lambda&:=\partial\Omega\cap B_\lambda=\{x\in B_\lambda\ ;\ (A_\lambda\, x)_d=h_\lambda(\overline{A_\lambda\, x})\},\end{aligned}$$

where $x\mapsto A_\lambda\, x$ is a change of coordinates of $\mathbb{R}^d$ (i.e., the composition of a translation and a rotation of $\mathbb{R}^d$) and where $\overline{y}$ stands for $(y_1,\dots,y_{d-1})$ if $y\in\mathbb{R}^d$. In the following, we also use the notations $Q_\lambda=(0,+\infty)\times\Omega_\lambda$ and $\Sigma_\lambda=(0,+\infty)\times\partial\Omega_\lambda$. When proving the comparison theorem and the error estimate, the problem is localized with the help of the functions λ_i. For the sake of clarity, we drop the index i and suppose that the change of coordinates is trivial: $A=\mathrm{Id}$. The open set $\Pi_\lambda=\{\overline{x}\,;\ x\in B_\lambda\}\subset\mathbb{R}^{d-1}$ is

used to parametrize $\partial\Omega_\lambda$. As a matter of fact, we even identify $\partial\Omega_\lambda$ with the graph of h restricted to Π_λ and Ω_λ with its epigraph. The outward unit normal to Ω_λ at any point $(\overline{x}, h(\overline{x}))$ of $\partial\Omega_\lambda$ is given by

$$n(\overline{x}) := n(\overline{x}, h(\overline{x})) = \frac{1}{\sqrt{1+|\nabla_{\overline{x}} h(\overline{x})|^2}}(\nabla_{\overline{x}} h(\overline{x}), -1).$$

Eventually, in order to make clearer integrations on $\partial\Omega_\lambda$, we use the notation

$$d\overline{\sigma}(\overline{x}) = \sqrt{1+|\nabla_{\overline{x}} h(\overline{x})|^2}\, d\overline{x}\,.$$

Regularization. Functions that are defined locally, i.e., that are defined on Ω_λ and $\partial\Omega_\lambda$, are regularized in the following way. Fix $\delta \in]0,1[$ and consider a smooth function $\theta : \mathbb{R} \to \mathbb{R}^+$ whose support is a subset of $[\delta, 1]$ and such that $\int \theta = 1$. Then define a (right-decentered) regularizing kernel $\theta_\varepsilon := \frac{1}{\varepsilon}\theta(\frac{\cdot}{\varepsilon})$ and set $\gamma_{\alpha,\varepsilon}(t, \bar{x}, x_d) = \theta_\alpha(t) \times \Pi_{i=1}^{d-1}\theta_{\overline{\varepsilon}}(x_i) \times \theta_{\varepsilon_d}(x_d)$. The space regularizing kernel $\Pi_{i=1}^{d-1}\theta_{\overline{\varepsilon}}(x_i) \times \theta_{\varepsilon_d}(x_d)$ is denoted by γ_ε. Consider now a function H defined on Q_λ and a function $\overline{H}$ defined on Σ_λ. Their (local) regularized functions are (both) defined on Q_λ by the following formulae:

$$\begin{cases} H^{\alpha,\varepsilon}(t,x) := (H \times 1_Q) \star \gamma_{\alpha,\varepsilon}(t,x) = \int_Q H(r,z)\gamma_{\alpha,\varepsilon}(t-r, x-z)\, dr\, dz, \\ \overline{H}^{\alpha,\varepsilon}(t,x) := (\overline{H} \times 1_\Sigma) \star \gamma_{\alpha,\varepsilon}(t,x) = \int_\Sigma \overline{H}(r,z)\gamma_{\alpha,\varepsilon}(t-r, x-z)\, dr\, d\sigma(z). \end{cases}$$

These two functions equal zero out of Q_λ as soon as $\delta\, \varepsilon_d \geq \sqrt{d}\, \mathrm{Lip} h\, \overline{\varepsilon}$, which is always assumed. Of course, if a function ψ is defined both on Q_λ and Σ_λ, then the two means of regularization described above do not lead to the same functions $\psi^{\alpha,\varepsilon}$; nevertheless, there will be no risk of confusion in the forthcoming proofs. Let us also point out the fact that this regularization is local and in fact depends on the map A_λ, even if it is hidden in computations in order to make them more readable.

3. A kinetic formulation of the Cauchy–Dirichlet problem. The main result of the paper is the following kinetic formulation of generalized entropy solutions. For any smooth test function $\phi \in \mathsf{C}_\mathsf{c}^\infty(\mathbb{R}^{d+2})$, $\phi^{(t=0)}$ and $\overline{\phi}$ denote, respectively, the restriction of ϕ to $\{0\} \times \Omega \times \mathbb{R}_\xi$ and to $\Sigma \times \mathbb{R}_\xi$.

THEOREM 3.1. *Consider a bounded function $u \in \mathsf{L}^\infty(Q)$. Let f^0 and f^b be the equilibrium functions associated with u_0 and u_b. Then u is a weak entropy solution of (1.1) if and only if there exists a bounded nonnegative measure $m \in \mathcal{M}^+(Q \times \mathbb{R}_\xi)$ and two nonnegative measurable functions $m_+^b, m_-^b \in \mathsf{L}^\infty_{\mathsf{loc}}(\Sigma \times \mathbb{R}_\xi)$ such that the function m_+^b vanishes for $\xi \gg 1$ (resp., the function m_-^b vanishes for $\xi \ll -1$) and such that the equilibrium function f associated with u satisfies for any $\phi \in \mathsf{C}_\mathsf{c}^\infty(\mathbb{R}^{d+2})$*

$$\begin{aligned} (3.1) \quad \int_{Q\times\mathbb{R}_\xi} f(\partial_t + a\cdot\nabla_x)\phi + \int_{\Omega\times\mathbb{R}_\xi} f^0\phi^{(t=0)} + \int_{\Sigma\times\mathbb{R}_\xi} (Mf_\pm^b + (-a\cdot n)\mathrm{sgn}_\mp)\overline{\phi} \\ = \int_{Q\times\mathbb{R}_\xi} \partial_\xi\phi\, dm + \int_{\Sigma\times\mathbb{R}_\xi} \partial_\xi\phi\, dm_\pm^b, \end{aligned}$$

where M is the Lipschitz constant of the flux function A on $\overline{Q} \times [-K, K]$.

In order to prove and understand this formulation, we define weak entropy sub- and supersolutions of the initial-boundary value problem (1.1) and exhibit a kinetic formulation for these semisolutions.

3.1. Weak entropy sub- and supersolutions. Let us define weak entropy sub- and supersolutions for the initial-boundary value problem (1.1).

DEFINITION 3.2. *Consider a bounded function* $u \in \mathsf{L}^\infty(Q)$.

1. *The function* u *is a* weak entropy subsolution *(resp.,* weak entropy supersolution*) of* (1.1) *if for any* $\kappa \in \mathbb{R}$ *and any* $\phi \in \mathsf{C}_c^\infty(\mathbb{R}_t \times \mathbb{R}^d)$, $\phi \geq 0$,

$$\begin{aligned}(3.2)\quad &\int_Q \left[(u(t,x)-\kappa)^\pm \partial_t \phi(t,x) + \mathcal{F}^\pm(u(t,x),\kappa).\nabla_x \phi(t,x)\right] dt\, dx \\ &+ \int_\Omega (u_0(x)-\kappa)^\pm \phi(0,x)dx + M\int_\Sigma (u_b(s,y)-\kappa)^\pm \phi(s,y)ds\, d\sigma(y) \geq 0.\end{aligned}$$

2. *The function* u *is a* weak entropy solution *of* (1.1) *if it is both a weak entropy subsolution and a supersolution.*

PROPOSITION 3.3. *Let* f^0 *and* f^b *be the equilibrium functions associated with* u_0 *and* u_b. *Consider a bounded function* $u \in \mathsf{L}^\infty(Q)$. *Then* u *is a weak entropy subsolution (resp., weak entropy supersolution) of* (1.1) *if and only if there exists* $m_\pm \in \mathsf{C}(\mathbb{R}_\xi; w - \mathcal{M}^+(\overline{Q}))$ *such that* m_ξ *vanishes for* $\xi \gg 1$ *(resp., for* $\xi \ll -1$*) and such that for any* $\phi \in \mathsf{C}_c^\infty(\mathbb{R}^{d+2})$,

$$\begin{aligned}(3.3)\quad &\int_{Q\times\mathbb{R}_\xi} f(\partial_t + a\cdot\nabla_x)\phi + \int_{\Omega\times\mathbb{R}_\xi} f^0\phi^{(t=0)} + \int_{\Sigma\times\mathbb{R}_\xi} (Mf^b_\pm + (-a\cdot n)\mathrm{sgn}_\mp)\overline{\phi} \\ &\qquad = \int_{\overline{Q}\times\mathbb{R}_\xi} \partial_\xi \phi dm_\pm.\end{aligned}$$

Remark 2. The function f satisfies (3.3) if and only if the function $f_\pm$ satisfies

$$(3.4)\quad \int_{Q\times\mathbb{R}_\xi} f_\pm(\partial_t + a\cdot\nabla_x)\phi + \int_{\Omega\times\mathbb{R}_\xi} f^0_\pm\phi^{(t=0)} + M\int_{\Sigma\times\mathbb{R}_\xi} f^b_\pm\overline{\phi} = \int_{\overline{Q}\times\mathbb{R}_\xi} \partial_\xi\phi dm_\pm.$$

Notice that here the expression of the boundary term is simplified. Moreover, (3.4) is the kinetic equation that appears in the construction $m_\pm$, and it is also the one we consider when proving the comparison theorem.

Proof of Proposition 3.3. Consider a weak entropy subsolution (resp., weak entropy supersolution) u of (1.1). Let us fix $\kappa \in \mathbb{R}$, and define a linear form $m^\kappa_\pm$ on $\mathsf{C}_c^\infty(\bar{Q})$ by

$$(3.5)\quad m^\kappa_\pm(\phi) = \int_Q (u-\kappa)^\pm \partial_t\phi + \mathcal{F}^\pm(u,\kappa)\cdot\nabla_x\phi + \int_\Omega (u_0-\kappa)^\pm\phi^{(t=0)} + M\int_\Sigma (u_b-\kappa)^\pm\overline{\phi}.$$

Since u is a weak entropy subsolution (resp., weak entropy supersolution), we know that $m^\kappa_\pm(\phi)$ is nonnegative for any κ and any ϕ. We conclude that for any κ, $m^\kappa_\pm$ is a nonnegative measure on $\overline{Q}$, and $m_\pm \in \mathsf{C}(\mathbb{R}_\xi, w - \mathcal{M}^+(\overline{Q}))$. Since $m_\pm \geq 0$, we have $||m_\pm|| = m_\pm(1) < +\infty$ by (3.5), and $m_\pm$ is bounded; moreover, $m_\pm$ vanishes for

$\kappa \gg 1$ (resp., $\kappa \ll 1$). Next, we compute

$$\begin{aligned}
&\int_{\bar{Q}\times\mathbb{R}_\xi} \partial_\xi\phi(t,x,\xi)dm_\pm(t,x,\xi)\\
&= \int_{Q\times\mathbb{R}_\xi} (u-\xi)^\pm \partial_t\partial_\xi\phi + \mathcal{F}^\pm(u,\xi)\cdot\nabla_x\partial_\xi\phi + \int_{\Omega\times\mathbb{R}_\xi} (u_0-\xi)^\pm\partial_\xi\phi^{(t=0)} + M\int_\Sigma (u_b-\xi)^\pm\overline{\partial_\xi\phi}\\
&= \int_{Q\times\mathbb{R}_\xi} \mathrm{sgn}_\pm(u-\xi)(\partial_t\phi + a\cdot\nabla_x\phi) + \int_{\Omega\times\mathbb{R}_\xi} \mathrm{sgn}_\pm(u_0-\xi)\phi^{(t=0)} + M\int_\Sigma \mathrm{sgn}_\pm(u_b-\xi)\overline{\phi}\\
&\qquad = \int_{Q\times\mathbb{R}_\xi} f_\pm(\partial_t\phi + a\cdot\nabla_x\phi) + \int_{\Omega\times\mathbb{R}_\xi} f^0_\pm\phi^{(t=0)} + M\int_\Sigma f^b_\pm\overline{\phi}\\
&\qquad = \int_{Q\times\mathbb{R}_\xi} f(\partial_t\phi + a\cdot\nabla_x\phi) + \int_{\Omega\times\mathbb{R}_\xi} f^0\phi^{(t=0)} + \int_\Sigma (Mf^b_\pm + (-a\cdot n)\mathrm{sgn}_\mp)\overline{\phi}.
\end{aligned}$$

Hence (3.3) is proved.

Conversely, consider $u \in \mathsf{L}^\infty(Q)$ and $g \in \mathsf{C}^\infty_{\mathsf{c}}(\mathbb{R}_t\times\mathbb{R}^d)$. Let $\xi \mapsto E_n(\xi)$ be a smooth approximation of $\xi \mapsto (\xi-\kappa)^\pm$ such that $|E_n'(\xi)| \le 1$ for any positive integer n. Let Ψ be a smooth function with support in $[-2,2]$, with values in $[0,1]$, and that equals 1 on $[-1,1]$. Next, define $\Psi_n(\xi) = \Psi(\xi/n)$. Now apply (3.4) to the test function $\phi(t,x,\xi) = g(t,x)\Psi_n(\xi)E_n'(\xi)$:

$$\begin{aligned}
&\int_Q \left[\int_{\mathbb{R}_\xi} \Psi_n E_n' f_\pm\right]\partial_t g + \left[\int_{\mathbb{R}_\xi} a\Psi_n E_n' f_\pm\right].\nabla_x g + \int_\Omega \left[\int_{\mathbb{R}_\xi} \Psi_n E_n' f^0_\pm\right] g^{(t=0)}\\
&\qquad + M\int_\Sigma \left[\int_{\mathbb{R}_\xi} \Psi_n E_n' f^b_\pm\right]\overline{g} = \int_{\bar{Q}\times\mathbb{R}_\xi} g[\Psi_n' E_n' + \Psi_n E_n'']dm_\pm.
\end{aligned}$$

Letting $n \to +\infty$, we get

(3.6)
$$\begin{aligned}
&\int_Q (u(t,x)-\kappa)^\pm\partial_t g(t,x) + \mathcal{F}^\pm(u(t,x),\kappa)\cdot\nabla_x g(t,x)dtdx + \int_\Omega (u_0(x)-\kappa)^\pm g(0,x)dx\\
&\qquad + M\int_\Sigma (u_b(s,y)-\kappa)^\pm g(s,y)ds\, d\sigma(y) = \int_Q g(t,x)dm_\pm(t,x,\kappa).
\end{aligned}$$

If, moreover, g is assumed to be nonnegative, (3.6) yields (3.2). ∎

3.2. Kinetic traces. In this subsection, we prove the following proposition. See [32, 33] and [19, Lemma 7.34, p. 115].

PROPOSITION 3.4. *Consider a function $f \in L^\infty(Q\times\mathbb{R}_\xi)$ satisfying* (3.3).

1. *There exist two kinetic functions $f^{\tau_0} \in \mathsf{L}^\infty(Q\times\mathbb{R}_\xi)$ and $f^\tau \in \mathsf{L}^\infty(\Sigma\times\mathbb{R}_\xi)$ such that*

$$\lim_{\alpha\to 0^+}\int_{\Omega\times\mathbb{R}_\xi}\left[\int_0^{+\infty} f(t)\theta_\alpha(t)dt\right]\phi = \int_{\Omega\times\mathbb{R}_\xi} f^{\tau_0}\phi, \tag{3.7}$$

$$\begin{aligned}
&\lim_{\varepsilon_d\to 0^+}\int_{[0;+\infty)\times\Pi_\lambda\times\mathbb{R}_\xi}(-a\cdot n)\left[\int_0^{+\infty} f(h(\overline{x})+r)\theta_{\varepsilon_d}(r)\,\lambda(h(\overline{x})+r)dr\right]\psi\\
&\qquad = \int_{[0;+\infty)\times\Pi_\lambda\times\mathbb{R}_\xi}(-a\cdot n)\, f^\tau\,\overline{\lambda}\,\psi
\end{aligned} \tag{3.8}$$

for any $\phi \in \mathsf{L}^1(\Omega \times \mathbb{R}_\xi)$ *and any* $\psi \in \mathsf{L}^1(\Sigma \times \mathbb{R}_\xi)$ *and any function* λ*, the element of the partition of unity* $\{\lambda_i\}_{i\in I}$.

2. *The time kinetic trace* f^{τ_0} *is bounded above (resp., bounded below) by* f^0*, and the space kinetic trace* f^τ *satisfies* (1.4)*, where* $m_\pm^b$ *denotes the restriction of* $m_\pm$ *to* $\Sigma \times \mathbb{R}_\xi$.

Proof. The proof of the existence of f^{τ_0} and of f^τ such that (3.7), (3.8) hold true can be found in [32, 33]. Let us prove that for any test function $\phi \in \mathsf{C}_\mathsf{c}^\infty(\mathbb{R}^{d+2})$,

$$\int_{Q\times\mathbb{R}_\xi} f(\partial_t + a\cdot\nabla_x)\phi + \int_{\Omega\times\mathbb{R}_\xi} f^{\tau_0}\phi^{(t=0)} + \int_{\Sigma\times\mathbb{R}_\xi} (-a\cdot n) f^\tau \overline{\phi} = \int_{Q\times\mathbb{R}_\xi} \partial_\xi \phi dm_\pm. \tag{3.9}$$

Let $\phi \in \mathsf{C}_\mathsf{c}^\infty([0;+\infty)\times\Omega\times\mathbb{R}_\xi)$; consider a right-decentered regularizing kernel $\theta_\alpha(r)$; define a cut-off function $w_\alpha(r) = \int_0^r \theta_\alpha(\tau)d\tau$ and apply (3.3) to the test function $w_\alpha(t)\phi(t,x,\xi)$:

$$\int_{Q\times\mathbb{R}_\xi} w_\alpha(t) f(\partial_t + a\cdot\nabla_x)\phi(t,x,\xi)dtdxd\xi + \int_{Q\times\mathbb{R}_\xi} \theta_\alpha(t) f(t,x,\xi)\phi(t,x,\xi)dtdxd\xi$$
$$= \int_{Q\times\mathbb{R}_\xi} w_\alpha(t)\partial_\xi\phi(t,x,\xi)dm(t,x,\xi).$$

Letting $\alpha \to 0+$ and using the Lebesgue dominated convergence theorem and (3.7), we obtain

$$\int_{Q\times\mathbb{R}_\xi} f(\partial_t + a\cdot\nabla_x)\phi + \int_{\Omega\times\mathbb{R}_\xi} f^{\tau_0}\phi^{(t=0)} = \int_{Q\times\mathbb{R}_\xi} \partial_\xi\phi dm. \tag{3.10}$$

Next, ϕ^λ denotes the function $\phi\,\lambda$, and we define a cut-off function

$$W_{\varepsilon_d}(x) = \int_0^{x_d - h(\bar{x})} \theta_{\varepsilon_d}(s)\,ds\,.$$

We apply (3.10) to the test function $\phi^\lambda W_{\varepsilon_d}$:

$$\int_{Q\times\mathbb{R}_\xi} W_{\varepsilon_d}(x) f(\partial_t + a\cdot\nabla_x)\phi^\lambda(t,x,\xi)dtdxd\xi + \int_{Q\times\mathbb{R}_\xi} f\phi^\lambda a\cdot\nabla_x W_{\varepsilon_d}$$
$$+ \int_{\Omega\times\mathbb{R}_\xi} f^{\tau_0}(x,\xi){\phi^\lambda}^{(t=0)} W_{\varepsilon_d}(x)dxd\xi = \int_{Q\times\mathbb{R}_\xi} \partial_\xi\phi^\lambda(t,x,\xi)W_{\varepsilon_d}(x)dm(t,x,\xi). \tag{3.11}$$

In (3.11), we can pass to the limit in each term, except from $\int_{Q\times\mathbb{R}_\xi} f\phi^\lambda a(\xi)\cdot\nabla_x W_{\varepsilon_d}$. Let us study it. Notice that

$$\nabla_x W_{\varepsilon_d}(x) = \theta_{\varepsilon_d}(x_d - h(\overline{x}))(-\nabla_{\overline{x}}h(\overline{x}),1) = -\theta_{\varepsilon_d}(x_d - h(\overline{x}))\sqrt{1+|\nabla_{\overline{x}}h(\overline{x})|^2}\; n(\overline{x}).$$

Hence,

$$\int_{Q\times\mathbb{R}_\xi} \phi^\lambda f a(\xi)\cdot\nabla_x W_{\varepsilon_d} dtdxd\xi$$
$$= \int_{Q\times\mathbb{R}_\xi} \phi^\lambda(t,x,\xi)(-a\cdot n) f(t,x,\xi)\theta_{\varepsilon_d}(x_d - h(\overline{x}))\sqrt{1+|\nabla_{\overline{x}}h(\overline{x})|^2}dtdxd\xi$$
$$= \int_{[0;+\infty)\times\Pi_\lambda\times\mathbb{R}_\xi} (-a\cdot n)\left[\int_{x_d=h(\overline{x})}^{+\infty} f(x_d)\theta_{\varepsilon_d}(x_d - h(\overline{x}))\,\lambda(x_d)dx_d\right]\phi^\lambda\; dtd\overline{\sigma}d\xi.$$

Using (3.8), we get (3.9) with ϕ^λ instead of ϕ as a test function. Recalling that the function λ is an element of the partition of unit $\{\lambda_i\}_{i\in I}$ and summing this previous inequality over $i \in I$ yields (3.9). We then deduce from (3.2) and (3.9) that (1.4) holds true and that $f^{\tau_0} = f^0 + \partial_\xi m^0_\pm$, where $m^0_\pm$ stands for the restriction of $m_\pm$ to $\{0\} \times \Omega \times \mathbb{R}_\xi$. It follows that

$$\int f^{\tau_0}_\pm(x,\xi)\mathrm{sgn}_\pm(\xi-\kappa)d\xi \le (u_0(x)-\kappa)^\pm.$$

Since f^{τ_0} is a kinetic function and $f^{\tau_0}(\xi) = 0$ for $\xi \gg 1$, we conclude that it can be written under the following form:

$$\begin{array}{ll} & f^{\tau_0}(x,\xi) = \nu^{\tau_0}_x(\xi,+\infty) + \mathrm{sgn}_- \\ (\text{resp.} & f^{\tau_0}(x,\xi) = \nu^{\tau_0}_x(-\infty;\xi) + \mathrm{sgn}_+). \end{array} \tag{3.12}$$

Next, replace κ with $u_0(x)$ and conclude that the support of $\nu^{\tau_0}_x$ lies in $(-\infty, u_0(x)]$ (resp., in $[u_0(x),+\infty)$). Finally, f^{τ_0} satisfies

$$f^{\tau_0}_+(x,\xi) = \nu^{\tau_0}_x\ (\xi \perp u_0(x), u_0(x)] \le \mathrm{sgn}_+(u_0(x)-\xi) \tag{3.13}$$

$$(\text{resp. } f^{\tau_0}_-(x,\xi) = -\nu^{\tau_0}_x\ [u_0(x), \xi \top u_0(x)) \ge \mathrm{sgn}_-(u_0(x)-\xi)). \tag{3.14}$$

This achieves the proof. □

Proof of Theorem 3.1. From Proposition 3.3, we get two measures $m_\pm$. If u is a weak entropy solution of the initial-boundary value problem, then m_+ and m_- coincide in $Q \times \mathbb{R}_\xi$. Indeed, from (3.5) we get

$$m_\pm(t,x,\kappa) = -\partial_t(u-\kappa)^\pm - \mathrm{div}_x\ \mathcal{F}^\pm(u,\kappa) \text{ in } \mathcal{D}'(Q\times\mathbb{R}_\xi). \tag{3.15}$$

Choosing κ large enough and $-\kappa$ large enough, respectively, we obtain that u is a weak solution of (1.1); i.e., $\partial_t u + \mathrm{div}_x\ A(u) = 0$ in $\mathcal{D}'(Q)$. Next, we conclude that $m_+ = m_-$ in $Q \times \mathbb{R}_\xi$:

$$m_\pm(t,x,\kappa) = -\frac{1}{2}\partial_t|u-\kappa| - \frac{1}{2}\mathrm{div}_x\ \mathcal{F}(u,\kappa) \text{ in } \mathcal{D}'(Q\times\mathbb{R}_\xi), \tag{3.16}$$

where $\mathcal{F} = \mathcal{F}^+ + \mathcal{F}^-$. Moreover, we proved in Proposition 3.4 that $f^{\tau_0} = f^0 + \partial_\xi m^0_\pm$ and that f^{τ_0} is bounded above and below by f^0. We then conclude that $\partial_\xi m^0_\pm = 0$, and hence that $m^0_\pm$ is constant in ξ. Since it equals 0 for large ξ, we conclude that $m^0_\pm = 0$. Eventually, the two measures $m^b_\pm$ are functions: indeed, since they satisfy (1.4) and vanish for $\xi \gg 1$ and $\xi \ll -1$, respectively, we have

$$m^b_+(s,y,\kappa) := M(u_b(s,y)-\kappa)^+ - \int_\kappa^{+\infty}(-a\cdot n)f^\tau_+(s,y,\xi)d\xi \ge 0, \tag{3.17}$$

$$m^b_-(s,y,\kappa) := M(u_b(s,y)-\kappa)^- + \int_{-\infty}^{\kappa}(-a\cdot n)f^\tau_-(s,y,\xi)d\xi \ge 0\,. \tag{3.18}$$

The proof of Theorem 3.1 is therefore achieved. □

Remark 3. Formula (3.16) appears in [18, p. 173]. Additional properties of m can be derived. See [18].

Link with the BLN condition. We detail here the link between the kinetic formulation of weak entropy solutions given in Theorem 3.1 and the BLN condition. Suppose that the function u is a weak entropy solution of problem (1.1) such that $u \in \mathsf{BV}(Q)$.

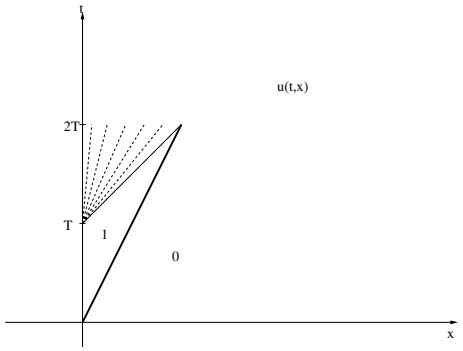

FIG. 3.1. *Weak entropy solution.*

Let u_τ denote the (strong) trace of the function u on Σ. Obviously, the space kinetic trace is the associated equilibrium function: $f^\tau = \chi_{u_\tau}$ (see Proposition 3.4). Next, remark that

$$\int_\kappa^{+\infty} a(\xi)\cdot n(y) f_+^\tau(s,y,\xi)d\xi = \mathcal{F}^+(u_\tau(s,y),\kappa)\cdot n(y)$$

and combine with (3.17) in order to get

$$m_+^b(s,y,\kappa) = M(u_b(s,y)-\kappa)^+ + \mathcal{F}^+(u_\tau(s,y),\kappa)\cdot n(y).$$

The fact that the function m_+^b is nonnegative is equivalent to the following condition:

$$\forall \kappa \in [u_b,u_\tau], \quad \operatorname{sgn}_+(u_\tau-u_b)[A(u_\tau)-A(\kappa)]\cdot n \geq 0.$$

Similarly, $m_-^b \geq 0$ if and only if the previous condition holds true replacing sgn_+ with sgn_-. Summing these two inequalities yields the well-known BLN condition [1]

$$\forall \kappa \in [u_b,u_\tau], \quad \operatorname{sgn}(u_\tau-u_b)[A(u_\tau)-A(\kappa)]\cdot n \geq 0. \tag{3.19}$$

3.3. An example. Let us detail the expressions of the entropy defect measure m and the boundary defect measures $m_\pm^b$ for the Burgers equation $\partial_t u + \partial_x(u^2/2) = 0$ considered on the domain $(0,2T)\times(0,+\infty)$ with data $u_0(x)=0$ and

$$u^b(t) = \begin{cases} 1 & \text{if} \quad 0<t<T, \\ -1 & \text{if} \quad T<t<2T. \end{cases}$$

A shock occurs at the time $t=0$, and a rarefaction wave appears at the time $t=T$. It collides with the shock at time $t=2T$. The corresponding weak entropy solution u is represented in Figure 3.1. Then the entropy defect measure is

$$m = \frac{1}{2}\left(\frac{1}{2}[|u-\xi|]_1^0 - [\operatorname{sgn}(u-\xi)(u^2/2-\xi^2/2)]_1^0\right)\delta_L,$$

where L is the line $t=2x$ in the (x,t)-plane and where $[G(u)]_1^0 := G(0)-G(1)$. In particular, the measure m is concentrated on the line of discontinuity of u, and the entropy criterion ensures that it is nonnegative. On the other hand, the boundary defect measures are given by

$$\begin{aligned} m_+^b(t,\xi) = (M(1-\xi)^+ - \operatorname{sgn}^+(1-\xi)(1/2-\xi^2/2))1_{(0,T)}(t) \\ + (M(1+\xi)^- - \operatorname{sgn}^-(\xi)\xi^2/2)1_{(T,2T)}(t) \end{aligned}$$

and

$$m_-^b(t,\xi) = (M(1-\xi)^- - \mathrm{sgn}^-(1-\xi)(1/2-\xi^2/2))\mathbf{1}_{(0,T)}(t) \\ + (M(1+\xi)^+ - \mathrm{sgn}^+(\xi)\xi^2/2)\mathbf{1}_{(T,2T)}(t),$$

where M is a constant greater than 1. The identity $a^2-b^2=(a+b)(a-b)$ ensures that the two functions are nonnegative. The reader can check that the expressions of m and $m_\pm^b$ are consistent with the formula (3.16) and (3.18)–(3.17), respectively.

4. A comparison theorem.

THEOREM 4.1. *Let $u \in \mathsf{L}^\infty(Q)$ be a weak entropy subsolution of* (1.1) *with data (u_0,u_b), and let $v \in \mathsf{L}^\infty(Q)$ be a weak entropy supersolution of* (1.1) *with data (v_0,v_b). Then*

$$\frac{1}{T}\int_0^T\int_\Omega (u(t,x)-v(t,x))^+dx\,dt \le \int_\Omega (u_0(x)-v_0(x))^+dx \\ + M\int_0^T\int_{\partial\Omega}(u_b(t,x)-v_b(t,x))^+dtd\sigma\,. \tag{4.1}$$

*In particular, $u \le v$ as soon as $u_0 \le v_0$ and $u_b \le v_b$ (*comparison principle*).*

Before proving Theorem 4.1, we state that the L^1-contraction property and the maximum principle follow from it.

COROLLARY 4.2.

1. *Let $u,v \in \mathsf{L}^\infty(Q)$ be two weak entropy solutions of* (1.1)*. Then*

$$\frac{1}{T}\int_0^T\int_\Omega |u(t,x)-v(t,x)|dx\,dt \le \int_\Omega |u_0(x)-v_0(x)|dx \\ + M\int_{(0;T)\times\partial\Omega}|u_b(t,y)-v_b(t,y)|dtd\sigma(y)$$

   (L^1-contraction property*).*

2. *Let u be a weak entropy solution of* (1.1)*, and suppose that there exists two constants $U_m, U_M \in \mathbb{R}$ such that*

$$U_m \le u_0 \le U_M \quad \text{a.e. in } \Omega \quad \text{and} \quad U_m \le u_b \le U_M \quad \text{a.e. in } \Sigma;$$

   *then $U_m \le u \le U_M$ a.e. in Q (*maximum principle*).*

Proof. The L^1-contraction property is obtained by combining the equations as (4.1) obtained successively with u as a weak entropy subsolution and v as a weak entropy supersolution and with v as a weak entropy subsolution and u as a weak entropy supersolution. In order to prove the maximum principle, one may remark that the constant function U_m is a weak entropy subsolution for data u_0,u_b and that the constant function U_M is a weak entropy supersolution for data u_0,u_b. ∎

Proof of Theorem 4.1. In order to prove Theorem 4.1, we show that

$$\int_Q (u-v)^+\partial_t\phi + \mathcal{F}^+(u,v)\cdot\nabla_x\phi + \int_\Omega (u_0-v_0)^+\phi^{(t=0)} + M\int_\Sigma (u_b-v_b)^+\overline{\phi} \ge 0 \tag{4.2}$$

holds true for any test function $\phi \in \mathsf{C}_\mathsf{c}^\infty(\mathbb{R}_t\times\mathbb{R}^d)$. Passing from (4.2) to (4.1) is classical. Let f, f^0, and f^b (resp., g, g^0, and g^b) denote the equilibrium functions

associated with u, u_0, and u_b (resp., with v, v_0, and v_b). The kinetic traces associated with u (resp., with v) are denoted by f^{τ_0} and f^{τ} (resp., g^{τ_0} and g^{τ}). Eventually, let m (resp., q) denote the entropy defect measure associated with u (resp., v) and set, for $(s, y, \xi) \in \Sigma \times \mathbb{R}_\xi$,

$$\overline{F_+}(s,y,\xi) = (-a(\xi)\cdot n(y)) f_+^{\tau}(s,y,\xi) \quad \text{and} \quad \overline{G_-}(s,y,\xi) = (-a(\xi)\cdot n(y)) g_-^{\tau}(s,y,\xi).$$

Since u is a weak entropy subsolution of (1.1), the following kinetic equation holds true:

$$\int_{Q\times\mathbb{R}_\xi} f_+(\partial_t + a\cdot\nabla_x)\phi + \int_{\Omega\times\mathbb{R}_\xi} f_+^{\tau_0}\ \phi^{(t=0)} + \int_{\Sigma\times\mathbb{R}_\xi} \overline{F_+}\phi = \int_{Q\times\mathbb{R}_\xi} \partial_\xi \phi dm \tag{4.3}$$

for any $\phi \in \mathsf{C}_{\mathsf{c}}^\infty(\mathbb{R}^{d+2})$. Let us fix a test function $\phi \in \mathsf{C}_{\mathsf{c}}^\infty(\mathbb{R}^{d+2})$ and apply (4.3) to the test function $\phi^\lambda \star \check{\gamma}_{\alpha,\varepsilon}$, where $\gamma_{\alpha,\varepsilon}$ denotes a right-decentered regularizing kernel and ϕ^λ denotes $\phi\lambda$:

$$\int_{\mathbb{R}^{d+2}} f_+^{\alpha,\varepsilon}(\partial_t + a\cdot\nabla_x)\phi^\lambda + f_+^{\tau_0\,\varepsilon}\theta_\alpha\phi^\lambda + \overline{F_+}^{\alpha,\varepsilon}\phi^\lambda = \int_{\mathbb{R}^{d+2}} \partial_\xi\phi^\lambda\ dm^{\alpha,\varepsilon}, \tag{4.4}$$

where $f_+^{\alpha,\varepsilon} = (f_+ \times 1_Q) \star_{t,x} \gamma_{\alpha,\varepsilon}$, $f_+^{\tau_0\,\varepsilon} = (f_+^{\tau_0} \times 1_{\Omega_\lambda}) \star_x \gamma_\varepsilon$, $m^{\alpha,\varepsilon} = (m \times 1_Q) \star_{t,x} \gamma_{\alpha,\varepsilon}$, and $\overline{F_+}^{\alpha,\varepsilon} = (\overline{F_+} \times 1_{\Sigma_\lambda}) \star_{t,x} \gamma_{\alpha,\varepsilon}$. Now, let us also regularize the kinetic equation satisfied by g but with different parameters:

$$\int_{\mathbb{R}^{d+2}} g_-^{\beta,\nu}(\partial_t + a\cdot\nabla_x)\phi^\lambda + g_-^{\tau_0\,\nu}\theta_\beta\phi^\lambda + \overline{G_-}^{\beta,\nu}\phi^\lambda = \int_{\mathbb{R}^{d+2}} \partial_\xi\phi^\lambda\ dq^{\beta,\nu}. \tag{4.5}$$

Now apply (4.4) to $-g_-^{\beta,\nu}(t,x,\xi)\phi^\lambda(t,x)$ and (4.5) to $-f_+^{\alpha,\varepsilon}(t,x,\xi)\phi^\lambda(t,x)$ and sum the two equations:

$$\begin{aligned}
&\int_{\mathbb{R}^{d+2}} -\phi^\lambda(\partial_t + a\cdot\nabla_x)(f_+^{\alpha,\varepsilon} g_-^{\beta,\nu}) + 2\int_{\mathbb{R}^{d+2}} (-f_+^{\alpha,\varepsilon} g_-^{\beta,\nu})(\partial_t + a\cdot\nabla_x)\phi^\lambda \\
&\quad - \int_{\mathbb{R}^{d+2}} \left[f_+^{\tau_0\,\varepsilon} g_-^{\beta,\nu}\theta_\alpha + g_-^{\tau_0\,\nu} f_+^{\alpha,\varepsilon}\theta_\beta \right] \phi^\lambda - \int_{\mathbb{R}^{d+2}} \left[\overline{F_+}^{\alpha,\varepsilon} g_-^{\beta,\nu} + \overline{G_-}^{\beta,\nu} f_+^{\alpha,\varepsilon} \right] \phi^\lambda \\
&\qquad\qquad = \int_{\mathbb{R}^{d+2}} \phi^\lambda [\delta_v^{\beta,\nu}\ dm^{\alpha,\varepsilon} + \delta_u^{\alpha,\varepsilon}\ dq^{\beta,\nu}],
\end{aligned} \tag{4.6}$$

where $\delta_u^{\alpha,\varepsilon} = (\delta(\xi - u(t,x)) \times 1_Q) \star \gamma_{\alpha,\varepsilon}$ and $\delta_v^{\beta,\nu} = (\delta(\xi - v(t,x)) \times 1_Q) \star \gamma_{\beta,\nu}$. Use the fact that the right-hand side of (4.6) is nonnegative and make an integration by parts in the first line:

$$\begin{aligned}
&\int_{\mathbb{R}^{d+2}} (-f_+^{\alpha,\varepsilon} g_-^{\beta,\nu})(\partial_t + a\cdot\nabla_x)\phi^\lambda \\
&\quad - \int_{\mathbb{R}^{d+2}} \left[f_+^{\tau_0\,\varepsilon} g_-^{\beta,\nu}\theta_\alpha + g_-^{\tau_0\,\nu} f_+^{\alpha,\varepsilon}\theta_\beta \right] \phi^\lambda - \int_{\mathbb{R}^{d+2}} \left[\overline{F_+}^{\alpha,\varepsilon} g_-^{\beta,\nu} + \overline{G_-}^{\beta,\nu} f_+^{\alpha,\varepsilon} \right] \phi^\lambda \geq 0.
\end{aligned}$$

Now let successively $\beta, \bar{\nu}$, and ν_d go to 0^+:

$$\int_{Q_\lambda\times\mathbb{R}_\xi} (-f_+^{\alpha,\varepsilon} g_-)(\partial_t + a\cdot\nabla_x)\phi^\lambda - \int_{Q_\lambda\times\mathbb{R}_\xi} f_+^{\tau_0\,\varepsilon} g_-\theta_\alpha\phi^\lambda - \int_{Q_\lambda\times\mathbb{R}_\xi} \overline{F_+}^{\alpha,\varepsilon} g_-\ \phi^\lambda \geq 0. \tag{4.7}$$

We used the fact that regularized functions equal zero at $t = 0$ and at the boundary. Next, let successively $\alpha, \overline{\varepsilon}$, and ε_d go to 0^+. The first limit is easy to compute:

(4.8)
$$\lim_{\varepsilon_d\to 0^+}\lim_{\overline{\varepsilon}\to 0^+}\lim_{\alpha\to 0^+}\int_{Q_\lambda\times\mathbb{R}_\xi}(-f_+^{\alpha,\varepsilon}g_-)(\partial_t + a\cdot\nabla_x)\phi^\lambda = \int_{Q\times\mathbb{R}_\xi}(-f_+g_-)(\partial_t + a\cdot\nabla_x)\phi^\lambda$$
$$= \int_Q (u-v)^+\partial_t\phi^\lambda + \mathcal{F}^+(u,v)\cdot\nabla\phi^\lambda.$$

Use (3.7) for g, (3.13) for f, and (3.14) for g:

(4.9)
$$\lim_{\varepsilon_d\to 0^+}\lim_{\overline{\varepsilon}\to 0^+}\lim_{\alpha\to 0^+} -\int_{Q_\lambda\times\mathbb{R}_\xi} f_+^{\tau_0\varepsilon}g_-\theta_\alpha\phi^\lambda = \lim_{\varepsilon_d\to 0^+}\lim_{\overline{\varepsilon}\to 0^+} -\int_{\Omega_\lambda\times\mathbb{R}_\xi} f_+^{\tau_0\varepsilon}g_-^{\tau_0}\left(\phi^\lambda\right)^{(t=0)}$$
$$= -\int_{\Omega_\lambda\times\mathbb{R}_\xi} f_+^{\tau_0}g_-^{\tau_0}\left(\phi^\lambda\right)^{(t=0)} \le -\int_{\Omega_\lambda\times\mathbb{R}_\xi} f_+^0 g_-^0\left(\phi^\lambda\right)^{(t=0)} = \int_\Omega (u_0-v_0)^+\left(\phi^\lambda\right)^{(t=0)}.$$

We proceed analogously with the boundary term:

$$\lim_{\varepsilon_d\to 0^+}\lim_{\overline{\varepsilon}\to 0^+}\lim_{\alpha\to 0^+} -\int_{Q_\lambda\times\mathbb{R}_\xi}\overline{F_+}^{\alpha,\varepsilon}g_-\,\phi^\lambda = \int_{\Sigma\times\mathbb{R}_\xi}(-a\cdot n)f_+^\tau\, g_-^\tau\overline{\phi^\lambda}$$
$$\le M\int_\Sigma (u_b-v_b)^+\overline{\phi^\lambda}. \tag{4.10}$$

Let us now justify the inequality in (4.10). In order to do so, we use (1.4) and represent f^τ and g^τ with their Young measures as in (3.12):

$$\int_{\mathbb{R}_\xi}(-a\cdot n)f_+^\tau\, g_-^\tau = -\int_{-\infty}^{v_b}\nu^\tau(\xi;+\infty)\partial_\xi q_-^b$$
$$+\int_{v_b}^{v_b\top u_b}(-a\cdot n)\nu^\tau(\xi;+\infty)\mu^\tau(-\infty;\xi) + \int_{v_b\top u_b}^{+\infty}\mu^\tau(-\infty;\xi)\partial_\xi m_+^b$$
$$\le -\int_{-\infty}^{v_b} q_-^b\, d\nu^\tau - \left[q_-^b\nu^\tau(\xi;+\infty)\right]_{-\infty}^{v_b} + M(u_b-v_b)^+$$
$$-\int_{v_b\top u_b}^{+\infty} m_+^b d\mu^\tau + \left[m_+^b\mu^\tau(-\infty;\xi)\right]_{v_b\top u_b}^{+\infty} \le M(u_b-v_b)^+.$$

Hence, we can pass to the limit in (4.7). By using (4.8), (4.9), and (4.10) and by summing over $i\in I$, (4.1) follows, and the proof of Theorem 4.1 is complete. ∎

5. Convergence of a BGK-like model. In this section, we present the first application of the kinetic formulation we introduced above. Let us consider the following BGK-like model:

$$(\partial_t + a\cdot\nabla_x)f_\varepsilon = \frac{\chi_{u_\varepsilon} - f_\varepsilon}{\varepsilon} \qquad \text{in } Q\times\mathbb{R}_\xi, \tag{5.1a}$$

$$u_\varepsilon(t,x) = \int_{\mathbb{R}} f_\varepsilon(t,x,\xi)d\xi, \quad (t,x)\in Q, \tag{5.1b}$$

$$f_\varepsilon(0,x,\xi) = f^0(x,\xi), \quad (x,\xi) \in \Omega \times \mathbb{R}_\xi, \tag{5.1c}$$

$$f_\varepsilon(t,y,\xi) = f^b(y,\xi), \quad (t,y,\xi) \in \Sigma^+, \tag{5.1d}$$

where f^0 and f^b are the equilibrium functions, respectively, associated with the initial and the boundary data and where $\Sigma^+ = \{(t,y,\xi) \in \Sigma \times \mathbb{R}_\xi : -a(\xi)\cdot n(y) > 0\}$. The approximation (5.1a)–(5.1c) for the Cauchy problem (i.e., when $\Omega = \mathbb{R}^n$) was first considered by Perthame and Tadmor [29]. They proved that the "hydrodynamic limit" as $\varepsilon \to 0$ is precisely the entropy solution of the initial value problem (1.1a)–(1.1b). Their study relies on the fact that the right-hand side of (5.1a) can be written as the derivative of a measure: $\partial_\xi m_\varepsilon$. This is a consequence of the following observation.

LEMMA 5.1 (see [18]). *Let $g \in \mathsf{L}^1(\mathbb{R})$ satisfy $0 \le \operatorname{sgn}(\xi) g(\xi) \le 1$ a.e. Then the function $m_g : \xi \mapsto \int_{-\infty}^{\xi} (\chi_{u_g} - g)(\zeta)\, d\zeta$ is nonnegative.*

As ε goes to 0, the measure m_ε converges to the entropy defect measure m. This kinetic model has been adapted by Nouri, Omrane, and Vila [22, 23] to take into account boundary conditions. In [22, 23], data at equilibrium as well as general kinetic ones are considered. The convergence of the kinetic model is proved and, particularly in the nonequilibrium case, the boundary conditions satisfied by the limit so obtained are discussed and compared to the BLN condition. In the present paper, we restrict ourselves to the case of data at equilibrium and show how the concept of boundary defect measures can help in the understanding of the "hydrodynamic limit"; more precisely, we define approximate boundary defect measures and prove that they converge to $m^b_\pm$ (see subsection 3.2). As in [28], we intend to show how a concept of a generalized kinetic solution can be used to prove the convergence of the kinetic model associated with (1.1) without "strong" (for instance BV) a priori estimates.

5.1. Solution of the kinetic model. We suppose that Ω is convex. The problem (5.1) admits an integral representation and is therefore solved by a fixed point method. The characteristic of the partial differential operator $\partial_t + a(\xi)\partial_x$ arriving at $(t,x) \in Q$ is the line of equation $X(\tau) = a(\xi)(\tau - t) + x$. If $u_\varepsilon \in \mathsf{C}(0,T;\mathsf{L}^1(\Omega))$, the solution f_ε of the linear equation $\partial_t f_\varepsilon + a(\xi)\cdot\nabla f_\varepsilon + \frac{1}{\varepsilon} f_\varepsilon = \frac{1}{\varepsilon}\chi_{u_\varepsilon}$ satisfies

$$f_\varepsilon(t,x,\xi) = f_\varepsilon(\tau, X(\tau), \xi) e^{\frac{\tau - t}{\varepsilon}} + \int_\tau^t \frac{1}{\varepsilon} \chi_{u_\varepsilon(s,X(s))}(\xi) e^{\frac{s-t}{\varepsilon}}\, ds \tag{5.2}$$

for any $\tau < t$ such that $X([\tau,t]) \subset \Omega$. Using the boundary condition (5.1d), we see that the computation of the value $f_\varepsilon(t,x,\xi)$ depends on the point of intersection of the characteristic line with the parabolic boundary:

- if $X([0,t]) \subset \Omega$, the characteristic starts from $\{0\} \times \Omega$ at $\tau = 0$, and we put $f_\varepsilon(\tau, X(\tau), \xi) = f^0(x - ta(\xi), \xi)$ in (5.2);
- if there exists $\tau^* \in [0,t]$ such that $X([\tau^*, t]) \subset \Omega$ and $X(\tau^* - 0) \notin \Omega$, the characteristic starts from the boundary Σ at $\tau = \tau^*$, and we put $f_\varepsilon(\tau, X(\tau), \xi) = f^b(\tau^*, X(\tau^*), \xi)$ in (5.2).

Thanks to the integral representation (5.2), it is therefore possible to build an operator T from $\mathsf{C}(0,T;\mathsf{L}^1(\Omega))$ to itself which maps u on $v : (t,x) \mapsto \int_{\mathbb{R}} f_\varepsilon(t,x,\xi)\, d\xi$. We then show that this operator is a contracting map, and the existence and the uniqueness of the solution f_ε of (5.1) follows [29, 22, 28]. This solution satisfies additional properties.

PROPOSITION 5.2 (see [29, 22, 28]). *Suppose that Ω is convex. Let $\varepsilon > 0$, and let $f_\varepsilon \in \mathsf{C}(0,T;\mathsf{L}^1(\Omega \times \mathbb{R}_\xi))$ be the solution of (5.1). Under the hypotheses of section 2, we have that*

1. f_ε *satisfies*

$$0 \le \operatorname{sgn}(\xi) f_\varepsilon(t,x,\xi) \le 1 \text{ for a.e. } (t,x,\xi) \in Q \times \mathbb{R}_\xi;$$

2. *there exists a nonnegative function* m_ε *such that*

$$\frac{\chi_{u_\varepsilon} - f_\varepsilon}{\varepsilon} = \partial_\xi m_\varepsilon; \tag{5.3}$$

3. *for every convex function* $\eta \in \mathsf{C}^2(\mathbb{R},\mathbb{R})$ *with a bounded derivative* η' *satisfying* $\eta'(0) = 0$,

$$\int_{Q\times\mathbb{R}_\xi} m_\varepsilon(t,x,\xi)\eta''(\xi)d\xi dx dt \le \int_{\Omega\times\mathbb{R}_\xi} f^0(\xi)\eta'(\xi)d\xi dx + \int_{\Sigma\times\mathbb{R}_\xi} (-a\cdot n)^+(s,y,\xi) f^b(\xi)\eta'(\xi)d\xi dt; \tag{5.4}$$

4. *there exists* $\mu \in \mathsf{L}^\infty(\mathbb{R})$ *independent of* ε *and such that* $\mu(\xi) = 0$ *if* $|\xi| \gg 1$ *and*

$$\int_Q m_\varepsilon(t,x,\xi)dx dt \le \mu(\xi); \tag{5.5}$$

5. *for a.e.* $(t,x,\xi) \in Q \times \mathbb{R}_\xi$: $f_\varepsilon(t,x,\xi) = 0$ *as soon as* $|\xi| > K$ *and*

$$\left| \int_{\mathbb{R}_\xi} f_\varepsilon(t,x,\xi)d\xi \right| \le K \text{ for a.e. } (t,x) \in Q. \tag{5.6}$$

Sketch of the proof. The fact that f_ε is a kinetic function follows from (5.2). We previously mentioned that (5.3) is a consequence of Lemma 5.1. A rigorous proof of (5.4) relies on the integral representation (5.2). Here is a formal argument: multiply the equation $\partial_t f_\varepsilon + a(\xi)\partial_x f_\varepsilon = \partial_\xi m_\varepsilon$ by $\eta'(\xi)$, integrate the result with respect to (t,x,ξ), and use the fact that $\eta'(\xi) f_\varepsilon(t,x,\xi) \ge 0$ (for $\operatorname{sgn}(\eta'(\xi)) = \operatorname{sgn}(\xi)$). Estimate (5.5) is a consequence of (5.4) with $\eta(\xi) = (\xi - \xi_0)^+$ if $\xi_0 > 0$ and $\eta(\xi) = (\xi - \xi_0)^-$ if $\xi_0 < 0$. It leads to the expression $\mu = \mu^+ + \mu^-$ with

$$\mu^\pm(\xi) = |\operatorname{sgn}_\pm(\xi)| \ \ (||(u_0 - \xi)^\pm||_{\mathsf{L}^1(\Omega)} + M||(u_b(t,y) - \xi)^\pm||_{\mathsf{L}^1(\Sigma)}).$$

Since f_ε is a kinetic function, (5.6) is a consequence of the fact that $f_\varepsilon(\cdot,\xi)$ vanishes for $|\xi| > K$. This argument also shows that the operator T maps

$$\left\{ u \in \mathsf{C}(0,T;\mathsf{L}^1(\Omega)),\ |u(t,x)| \le K \ \forall (t,x) \right\}$$

into itself: (5.6) follows from the uniqueness of the fixed point. ∎

5.2. Generalized kinetic solutions. In order to prove the convergence of the model, we need to introduce a very weak notion of solution of (1.1).

DEFINITION 5.3. *Consider a kinetic function* $f \in \mathsf{L}^\infty(Q \times \mathbb{R}_\xi)$. *We say that* f *is a* generalized kinetic solution *of* (1.1) *if there exists a bounded nonnegative measure* $m \in \mathcal{M}^+(Q \times \mathbb{R}_\xi)$ *and two nonnegative measurable functions* $m^b_+, m^b_- \in \mathsf{L}^\infty_{\mathsf{loc}}(\Sigma \times \mathbb{R}_\xi)$ *such that the function* m^b_+ *vanishes for* $\xi \gg 1$ *(resp., the function* m^b_- *vanishes for* $\xi \ll -1$*) and such that* (3.1) *holds true.*

The kinetic formulation can therefore be stated in the following terms: a function u is an entropy solution of (1.1) if and only if its associated equilibrium function is a generalized kinetic solution of (1.1).

THEOREM 5.4. *Any generalized kinetic solution of* (1.1) *is in fact an equilibrium function associated with an entropy solution of the initial-boundary value problem.*

Proof. We just adapt the proof of the comparison theorem. Consider a generalized kinetic solution f of the initial-boundary value problem. We can therefore easily prove that for a.e. $t > 0$:

$$\int_{\Omega\times\mathbb{R}_\xi} (-f^+ f^-)(t,x,\xi)dxd\xi \le 0.$$

Now use the fact that f is a kinetic function to get that for a.e. $(t,x) \in Q$:

$$f^-(t,x,\xi) = \nu_{t,x}(-\infty;\xi) \qquad \text{and} \qquad f^+(t,x,\xi) = \nu_{t,x}(\xi;+\infty).$$

Consequently, $\nu_{t,x}(-\infty;\xi) = 0$ or $\nu_{t,x}(\xi;+\infty) = 0$. It follows that $\nu_{t,x}$ is a Dirac mass. The proof is therefore complete. □

5.3. Proof of the convergence. We now state and prove a precise convergence result.

THEOREM 5.5. *Suppose that Ω is convex. Under the hypotheses of section 2, if f_ε denotes the solution of* (5.1)*, then the sequence of function u_ε defined by $u_\varepsilon(t,x) = \int_{\mathbb{R}} f_\varepsilon(t,x,\xi)d\xi$ converges as $\varepsilon \to 0$ to the entropy solution u of* (1.1) *in any $L^p((0,T)\times\Omega)$, $1 \le p < +\infty$.*

Proof. Let $\overline{f}_\varepsilon$ denote the space kinetic trace of f_ε, and consider $\varphi \in \mathsf{C}_c^\infty(\bar{Q}\times\mathbb{R}_\xi)$. By integrating the equation $\partial_t f_\varepsilon + a(\xi)\cdot\partial_x f_\varepsilon = \partial_\xi m_\varepsilon$ against φ we get

$$\int_{Q\times\mathbb{R}_\xi} f_\varepsilon(\partial_t\varphi + a\cdot\nabla_x\varphi) + \int_{\Omega\times\mathbb{R}_\xi} f^0\varphi^{(t=0)} + \int_{\Sigma\times\mathbb{R}_\xi} (-a\cdot n)\overline{f}_\varepsilon\overline{\varphi} = \int_{Q\times\mathbb{R}_\xi} \partial_\xi\varphi dm_\varepsilon. \tag{5.7}$$

By analogy with (3.17), define the function $m_+^{b,\varepsilon}$ by

$$m_+^{b,\varepsilon}(t,y,\xi) := M(u_b(t,y)-\xi)^+ - \int_\xi^{+\infty} (-a\cdot n)(\overline{f}_\varepsilon - \mathrm{sgn}_-)(\kappa)d\kappa$$

and get from (5.7)

$$\int_{Q\times\mathbb{R}_\xi} f_\varepsilon(\partial_t\varphi + a\cdot\nabla_x\varphi) + \int_{\Omega\times\mathbb{R}_\xi} f^0\varphi^{(t=0)} + \int_{\Sigma\times\mathbb{R}_\xi} (Mf_+^b + (-a\cdot n)\mathrm{sgn}_-)\overline{\varphi} = \int_{Q\times\mathbb{R}_\xi} \partial_\xi\varphi dm_\varepsilon + \int_{\Sigma\times\mathbb{R}_\xi} \partial_\xi\overline{\varphi} dm_+^{b,\varepsilon}. \tag{5.8}$$

Let us check that $m_+^{b,\varepsilon}(t,y,\xi)$ is a nonnegative function. Since $\overline{f}_\varepsilon$ is a kinetic function, $\overline{f}_\varepsilon - \mathrm{sgn}_-$ is nonnegative; hence,

$$\begin{aligned} m_+^{b,\varepsilon}(t,y,\xi) &\ge M(u_b(t,y)-\xi)^+ - \int_\xi^{+\infty} (-a\cdot n)^+(\overline{f}_\varepsilon(t,y,\kappa) - \mathrm{sgn}_-(\kappa))d\kappa \\ &= M(u_b(t,y)-\xi)^+ - \int_\xi^{+\infty} (-a\cdot n)^+(f^b(t,y,\kappa) - \mathrm{sgn}_-(\kappa))d\kappa \end{aligned}$$

$$\begin{aligned} &= \int_\xi^{+\infty} (M - (-a\cdot n)^+)(f^b(t,y,\kappa) - \mathrm{sgn}_-(\kappa))d\kappa \\ &\geq 0. \end{aligned}$$

Since f_ε is bounded in the L^∞-norm and m_ε is bounded in mass by (5.5), we have, up to subsequences,

$$\begin{array}{lll} f_\varepsilon \rightharpoonup f & \text{in} & w-*-L^\infty(Q\times\mathbb{R}_\xi), \\ \overline{f}_\varepsilon \rightharpoonup \widetilde{f} & \text{in} & w-*-L^\infty(\Sigma\times\mathbb{R}_\xi), \\ m_\varepsilon \rightharpoonup m & \text{in} & w-*-\mathcal{M}^+(Q\times\mathbb{R}_\xi), \end{array}$$

where f and $\widetilde{f}$ are, respectively, functions of $L^\infty(Q\times\mathbb{R}_\xi)$ and $L^\infty(\Sigma\times\mathbb{R}_\xi)$ such that (this property is preserved at the $w-*-$limit) $0\leq f(\cdot,\xi)\mathrm{sgn}(\xi)\leq 1$ and $0\leq \widetilde{f}(\cdot,\xi)\mathrm{sgn}(\xi)\leq 1$. We first deduce from Proposition 5.2 that

$$\int_\xi^{+\infty} (-a\cdot n)(\overline{f}_\varepsilon(t,y,\kappa) - \mathrm{sgn}_-(\kappa))d\kappa = \int_\xi^{K} (-a\cdot n)(\overline{f}_\varepsilon(t,y,\kappa) - \mathrm{sgn}_-(\kappa))d\kappa.$$

It follows that $m_+^{b,\varepsilon}(t,y,\xi)\rightharpoonup m_+^b$, where

$$m_+^b(t,y,\xi) := M(u_b(t,y)-\xi)^+ - \int_\xi^K (-a\cdot n)(f^\tau - \mathrm{sgn}_-(\kappa))d\kappa \tag{5.9}$$

so that, at the limit $\varepsilon\to 0$ in (5.8), we have

$$\begin{aligned} \int_{Q\times\mathbb{R}_\xi} f(\partial_t\varphi + a\cdot\nabla_x\varphi) + \int_{\Omega\times\mathbb{R}_\xi} f^0\varphi^{(t=0)} + \int_{\Sigma\times\mathbb{R}_\xi} (Mf_+^b + (-a\cdot n)\mathrm{sgn}_-)\overline{\varphi} \\ = \int_{Q\times\mathbb{R}_\xi} \partial_\xi\varphi dm + \int_{\Sigma\times\mathbb{R}_\xi} \partial_\xi\overline{\varphi} dm_+^b\,. \end{aligned} \tag{5.10}$$

Besides, it is clear from (5.9) that $m_+^b(t,y,\xi)$ vanishes for $\xi\gg 1$; moreover, (5.5) remains true at the limit. Derivating (5.1a) with respect to ξ gives

$$\partial_\xi f_\varepsilon = \partial_\xi\chi_{u_\varepsilon} + \alpha_\varepsilon = \delta_0(\xi) - \delta_{u_\varepsilon}(\xi) + \alpha_\varepsilon,$$

where $\alpha_\varepsilon = \varepsilon\,(\partial_{\xi t} f_\varepsilon + a(\xi)\partial_{\xi x} f_\varepsilon)$ tends to zero in $\mathcal{D}'(Q\times\mathbb{R}_\xi)$. We then define a Young measure $\nu_{t,x}(\xi)$ as an adherence value of $\delta(\xi - u_\varepsilon(t,x))$, and we obtain that

$$\partial_\xi f = \delta_0(\xi) - \nu_{t,x}(\xi) \quad \text{ in } \mathcal{D}'(Q\times\mathbb{R}_\xi).$$

Of course, the same arguments remain valid for m_-^b, and, consequently, f is a generalized kinetic solution of (1.1). By virtue of Theorem 5.4, it is therefore the equilibrium function associated with the unique entropy solution of (1.1). Since f is an equilibrium function, the weak-$*$ convergence of f_ε to f in $L^\infty(Q\times\mathbb{R}_\xi)$ implies the strong convergence of u_ε to u in $L^p(Q)$, $1\leq p<+\infty$. The proof is therefore complete. ∎

REFERENCES

[1] C. Bardos, A. Y. Le Roux, and J.-C. Nédélec, *First order quasilinear equations with boundary conditions*, Comm. Partial Differential Equations, 4 (1979), pp. 1017–1034.

[2] L. Barthélemy, *Problème d'obstacle pour une équation quasi-linéaire du premier ordre*, Ann. Fac. Sci. Toulouse Math. (5), 9 (1988), pp. 137–159.

[3] F. Benilan and S. N. Kruzhkov, *First-order quasilinear equations with continuous nonlinearities*, Dokl. Akad. Nauk, 339 (1994), pp. 151–154.

[4] P. Bénilan and S. Kružkov, *Conservation laws with continuous flux functions*, NoDEA Nonlinear Differential Equations Appl., 3 (1996), pp. 395–419.

[5] F. Berthelin and F. Bouchut, *Weak entropy boundary conditions for isentropic gas dynamics via kinetic relaxation*, J. Differential Equations, 185 (2002), pp. 251–270.

[6] B. Ben Moussa and A. Szepessy, *Scalar conservation laws with boundary conditions and rough data measure solutions*, Methods Appl. Anal., 9 (2002), pp. 579–598.

[7] Y. Brenier, *Résolution d'équations d'évolution quasilinéaires en dimension N d'espace à l'aide d'équations linéaires en dimension N+1*, J. Differential Equations, 50 (1983), pp. 375–390.

[8] J. Carrillo, *Entropy solutions for nonlinear degenerate problems*, Arch. Ration. Mech. Anal., 147 (1999), pp. 269–361.

[9] R. J. DiPerna, *Measure-valued solutions to conservation laws*, Arch. Rational Mech. Anal., 88 (1985), pp. 223–270.

[10] J. Droniou, C. Imbert, and J. Vovelle, *An error estimate for the parabolic approximation of multidimensional scalar conservation laws with boundary*, Ann. Inst. H. Poincaré Anal. Non Linéaire, to appear.

[11] R. Eymard, T. Gallouët, and R. Herbin, *Finite volume methods*, in Handbook of Numerical Analysis, Vol. VII, Handb. Numer. Anal. VII, North-Holland, P. G. Ciarlet and J. L. Lions, eds., Amsterdam, 2000, pp. 713–1020.

[12] Y. Giga and T. Miyakawa, *A kinetic construction of global solutions of first order quasilinear equations*, Duke Math. J., 50 (1983), pp. 505–515.

[13] S. Hwang and A. E. Tzavaras, *Kinetic decomposition of approximate solutions to conservation laws: Application to relaxation and diffusion-dispersion approximations*, Comm. Partial Differential Equations, 27 (2002), pp. 1229–1254.

[14] S. Hwang, *Kinetic decomposition for kinetic models of bgk type*, J. Differential Equations, 190 (2003), pp. 353–363.

[15] S. N. Kruzhkov and E. Y. Panov, *First-order conservative quasilinear laws with an infinite domain of dependence on the initial data*, Dokl. Akad. Nauk SSSR, 314 (1990), pp. 79–84.

[16] S. N. Kruzhkov and E. Y. Panov, *Osgood's type conditions for uniqueness of entropy solutions to Cauchy problem for quasilinear conservation laws of the first order*, Ann. Univ. Ferrara Sez. VII (N.S.), 40 (1994), pp. 31–54.

[17] S. N. Kružkov, *First order quasilinear equations with several independent variables*, Mat. Sb. (N.S.), 81 (1970), pp. 228–255.

[18] P.-L. Lions, B. Perthame, and E. Tadmor, *A kinetic formulation of multidimensional scalar conservation laws and related equations*, J. Amer. Math. Soc., 7 (1994), pp. 169–191.

[19] J. Málek, J. Nečas, M. Rokyta, and M. Růžička, *Weak and measure-valued solutions to evolutionary PDEs*, Appl. Math. Math. Comput. 13, Chapman and Hall, London, 1996.

[20] V. Milisic, *Stability and convergence of discrete kinetic approximations to an initial-boundary value problem for conservations laws*, Proc. Amer. Math. Soc., 131 (2003), pp. 1727–1737.

[21] R. Natalini and A. Terracina, *Convergence of a relaxation approximation to a boundary value problem for conservation laws*, Comm. Partial Differential Equations, 26 (2001), pp. 1235–1252.

[22] A. Nouri, A. Omrane, and J. P. Vila, *Boundary conditions for scalar conservation laws from a kinetic point of view*, J. Statist. Phys., 94 (1999), pp. 779–804.

[23] A. Nouri, A. Omrane, and J. P. Vila, *Erratum to "boundary conditions for scalar conservation laws from a kinetic point of view"*, J. Statist. Phys., submitted.

[24] A. Omrane and J. P. Vila, *On Two Kinetic Approaches for Scalar Conservation Laws*, preprint.

[25] F. Otto, *Initial-boundary value problem for a scalar conservation law*, C. R. Acad. Sci. Paris Sér. I Math., 322 (1996), pp. 729–734.

[26] E. Y. Panov, *On the theory of generalized entropy sub- and supersolutions of the Cauchy problem for a first-order quasilinear equation*, Differ. Uravn., 37 (2001), pp. 252–259, 287.

[27] B. Perthame, *Uniqueness and error estimates in first order quasilinear conservation laws via the kinetic entropy defect measure*, J. Math. Pures Appl. (9), 77 (1998), pp. 1055–1064.

[28] B. Perthame, *Kinetic Formulations of Conservation Laws*, Oxford University Press, Oxford, UK, 2002.
[29] B. Perthame and E. Tadmor, *A kinetic equation with kinetic entropy functions for scalar conservation laws*, Comm. Math. Phys., 136 (1991), pp. 501–517.
[30] D. Serre, *Systèmes de lois de conservation*. II, *Fondations. Structures géométriques, oscillation et problèmes mixtes*, Diderot Editeur, Paris, 1996.
[31] A. Terracina, *Comparison properties for scalar conservation laws with boundary conditions*, Nonlinear Anal., 28 (1997), pp. 633–653.
[32] A. Vasseur, *Strong traces for solutions of multidimensional scalar conservation laws*, Arch. Ration. Mech. Anal., 160 (2001), pp. 181–193.
[33] A. Vasseur, *Well-posedness of scalar conservation laws with singular sources*, Methods Appl. Anal., 9 (2002), pp. 291–312.
[34] J. Vovelle, *Convergence of finite volume monotone schemes for scalar conservation laws on bounded domains*, Numer. Math., 90 (2002), pp. 563–596.

SIAM J. MATH. ANAL.
Vol. 36, No. 1, pp. 233–255

NORMAL FORM OF REVERSIBLE SYSTEMS AND PERSISTENCE OF LOWER DIMENSIONAL TORI UNDER WEAKER NONRESONANCE CONDITIONS*

JUNXIANG XU†

Abstract. In this paper we give a normal form for reversible systems and then prove the persistence of lower dimensional invariant tori for integrable reversible systems under small perturbations with weaker nonresonance conditions than have previously been imposed. Our nonresonance conditions correspond to the first Melnikov's conditions in the case of Hamiltonian systems.

Key words. reversible systems, Kolmogorov–Arnold–Moser iteration, invariant tori, nonresonance conditions

AMS subject classifications. Primary, 58F14; Secondary, 34C35

DOI. 10.1137/S0036141003421923

1. Introduction.

Reversible systems form a class of special conservative systems with an involution structure. In studying persistence of invariant tori for reversible systems, there is a small divisor problem reminiscent of the case of Hamiltonian systems. Since reversible systems have many similar properties to Hamiltonian systems, we can apply some of the technique developed for Hamiltonian systems to them. Recently, there has been progress in the theory of invariant tori for Hamiltonian systems (see [1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 14, 19, 20]). In particular, under weaker and fewer nonresonance conditions, many KAM results have been improved for persistence of invariant tori of integrable Hamiltonian systems [1, 2, 3, 6, 21, 22]. Using the motivation of KAM theorems for Hamiltonian systems, we consider the persistence of invariant tori for reversible systems under weaker and fewer nonresonance conditions.

Consider the following dynamical system:

$$\begin{cases} \dot{x} = \omega + f(x,u,v;\omega), \\ \dot{u} = A(\omega)v + g_1(x,u,v;\omega), \\ \dot{v} = -B(\omega)u + g_2(x,u,v;\omega), \end{cases} \tag{1.1}$$

where the variables $x = (x_1,\dots,x_n)^T$, $u = (u_1,\dots,u_p)^T$, and $v = (v_1,\dots,v_q)^T$ are all column vectors and $(x,u,v) \in T^n \times R^p \times R^q$, $(p \le q)$. Note that here and below the superscript "T" always indicates the transpose of matrix. $\omega = (\omega_1,\dots,\omega_n)^T \in O \subset R^n$ is the frequency parameter. $A(\omega)$ and $B(\omega)$ are $p \times q$ and $q \times p$ matrices depending on ω, respectively. The corresponding involution G which characterizes the class of reversible systems is defined by

$$G : (x,u,v) \to (-x,-u,v).$$

Denote the vector field of the dynamical system (1.1) by

$$F = ((\omega + f)^T, (Av + g_1)^T, (-Bu + g_2)^T)^T.$$

*Received by the editors January 27, 2003; accepted for publication (in revised form) October 10, 2003; published electronically June 22, 2004. This work was supported by the National Natural Science Foundation of China (10171012) and the Special Funds for Major State Basic Research Projects (973 projects) and partly supported by the Science Foundation of Southeast University.
http://www.siam.org/journals/sima/36-1/42192.html
†Department of Mathematics, Southeast University, Nanjing 210096, People's Republic of China (xujun@seu.edu.cn).

System (1.1) is called reversible if $DG \cdot F = -F \circ G$. System (1.1) is reversible with respect to G when

$$\begin{cases} f(-x,-u,v;\omega) = f(x,u,v;\omega), \\ g_1(-x,-u,v;\omega) = g_1(x,u,v;\omega), \\ g_2(-x,-u,v;\omega) = -g_2(x,u,v;\omega). \end{cases} \tag{1.2}$$

A mapping $\Phi : (x,u,v) \to (x_+,u_+,v_+)$ is called a compatible transformation with respect to the involution G if $\Phi \circ G = G \circ \Phi$. Under compatible transformations, reversible systems are transformed to reversible systems.

For reversible systems, there are already many well-known results on the persistence of invariant tori. In [15, 16, 17, 18], Sevryuk studied the persistence of n-dimensional invariant tori for reversible systems of the form (1.1) under the following assumptions:

(i) $\det(\Omega) \neq 0$.

(ii) Every eigenvalue of the matrix Ω is simple, where

$$\Omega = \begin{pmatrix} 0 & A \\ -B & 0 \end{pmatrix}.$$

The assumption (i) means that the matrix Ω has no zero-eigenvalue and, (ii) implies that the matrix Ω is diagonalizable.

Recently, in [7] Liu weakened Sevryuk's assumptions and allowed Ω to have the eigenvalue zero or multiple eigenvalues. In the proof, he supposed the following nonresonance conditions: for all $k \neq 0$,

$$|\langle \omega, k\rangle| \geq \alpha |k|^{-\tau}, \tag{1.3}$$

$$|\det(i\langle \omega, k\rangle I_{(p+q)} - \Omega)| \geq \alpha |k|^{-\tau}, \tag{1.4}$$

$$|\det(i\langle \omega, k\rangle I_{(p+q)^2} + I_{(p+q)^2} \otimes \Omega - \Omega \otimes I_{(p+q)^2})| \geq \alpha |k|^{-\tau}, \tag{1.5}$$

where $i = \sqrt{-1}$, I_N indicates the Nth-order unit matrix, and $\otimes$ is the notation of tensor product of matrices (see [7]).

In this paper we want to prove that conditions (1.3) and (1.4) alone are sufficient for Liu's result. To state our results, we first give some definitions and assumptions. Denote a complex neighborhood of $T^n \times \{0\} \times \{0\}$ by

$$D(s,r) = \{(x,u,v) \mid |\mathrm{Im}\ x| \leq s, |u| \leq r, |v| \leq r\},$$

where $|\mathrm{Im}\ x| = \max_{1\leq i\leq n} |\mathrm{Im}\ x_i|$, $|u| = \max_{1\leq i\leq p} |u_i|$, and $|v| = \max_{1\leq i\leq q} |v_i|$.

Let O be a bounded closed simply connected domain of R^n with positive Lebesgue measure. Let $C^L(O)$ be the space of the Lth continuously differentiable function on O in Whitney's sense (see [23]). For $f(\omega) \in C^L(O)$, define a norm by $\|f\|^L = \max_{|\alpha|\leq L} \sup_{\omega\in O} |D^\alpha f(\omega)|$.

If f is analytic in $(x,u,v;\omega) \in D(s,r) \times O$, then we can write

$$f = \sum_{k,l,m} f_{k,l,m}(\omega) e^{\mathrm{i}\langle k,x\rangle} u^l v^m.$$

Define

$$\|f\|^L_{D(s,r)} = \sup_{(x,u,v)\in D(s,r)} \left| \sum_{k,l,m} \|f_{k,l,m}(\omega)\|^L e^{s|k|} u^l v^m \right|.$$

It is a stronger norm than the usual supremum-norm. If f depends only on x and ω, we write $\|f\|_s^L := \|f\|_{D(s,r)}^L$.

Assume that $A(\omega)$ and $B(\omega)$ are analytic with respect to ω on O. Also assume that f, g_1, and g_2 are analytic on $D(s,r) \times O$.

THEOREM 1.1. *Assume that*

(A.1) $p = q$ *and* $\operatorname{rank}(A) = p$;

(A.2) *there are* $\lambda_1(\omega), \ldots, \lambda_q(\omega)$, *which are analytic in* ω, *such that* $\lambda_1^2, \ldots, \lambda_q^2$ *are the eigenvalues of the matrix* AB *and the following nonresonance conditions hold:*

$$\langle \omega, k \rangle \neq 0 \quad \forall k \neq 0, \tag{1.6}$$

$$\langle \omega, k \rangle - \lambda_j \neq 0 \quad \forall k \neq 0, \ \ j = 1, \ldots, q. \tag{1.7}$$

Then, for sufficiently small $\alpha > 0$, *there exists an* $\epsilon > 0$ *depending on* $\lambda_1, \ldots, \lambda_q, \alpha, O$, n, m *such that if*

$$\|f\|_{D(s,r)}^L \leq \epsilon, \quad \frac{1}{r}\|g_1\|_{D(s,r)}^L \leq \epsilon, \quad \frac{1}{r}\|g_2\|_{D(s,r)}^L \leq \epsilon \tag{1.8}$$

with $L \geq q^2$, *the following holds true: There exists a nonempty subset* O_α *of* O *such that, for all* $\omega \in O_\alpha$, *there exists an analytic compatible transformation*

$$\Phi_*(\cdot;\omega) : D(s/2, r/2) \to D(s,r)$$

which transforms the reversible system (1.1) *into the form*

$$\dot{x} = \omega_* + f_*, \quad \dot{u} = A_*(\omega)v + g_{*1}, \quad \dot{v} = -B_*(\omega)u + g_{*2}, \tag{1.9}$$

where f_*, g_{*1}, *and* g_{*2} *satisfy*

$$f_*(x,0,0;\omega) = 0, \quad g_{*1}(x,0,0;\omega) = 0, \quad g_{*2}(x,0,0;\omega) = 0.$$

Hence, for $\omega \in O_\alpha$, $\Phi_*(T^n \times \{0\} \times \{0\};\omega)$ *is an invariant torus of the reversible system* (1.1) *with the frequency* ω_* *satisfying* $\|\omega_*(\omega) - \omega\|^L \leq 2\epsilon$. *Moreover, we have* $\operatorname{meas}(O \setminus O_\alpha) \to 0$ *as* $\alpha \to 0$.

THEOREM 1.2. *Assume that*

(A.3) $p < q$ *and* $\operatorname{rank}(A) = p$;

(A.4) *there are* $\lambda_1, \ldots, \lambda_q$, *which are analytic in* ω, *such that* $\lambda_1^2, \ldots, \lambda_q^2$ *are the eigenvalues of the square matrix* $(Q^{-1}BP, 0_1)$ *and conditions* (1.6) *and* (1.7) *hold. Here* P *is a nonsingular* $p \times p$*-matrix and* Q *is a nonsingular* $q \times q$*-matrix such that* $PAQ = (I_p, 0_2)$, *where* 0_1 *and* 0_2 *are* $p \times (q-p)$ *and* $(q-p) \times q$ *zero matrices, respectively. Then, for sufficiently small* $\alpha > 0$, *there exists an* $\epsilon > 0$ *depending on* $\lambda_j (1 \leq j \leq q), \alpha, O, n, m$ *such that if* (1.8) *holds, then there exists a nonempty subset* O_α *of* O *such that for all* $\omega \in O_\alpha$, *the reversible system* (1.1) *has an invariant torus.*

Below we first consider the case $p = q$. As preparation for the proof of these two theorems, we reduce a linear reversible system to a normal form, which is necessary to our KAM steps, and we construct a special compatible transformation, which is used to deal with certain resonant relations. Then we prove Theorem 1.1 by KAM iteration. The case $p < q$ we reduce to the special case $p = q$ in Theorem 1.1 and then prove Theorem 1.2.

2. Normal forms for linear reversible systems. In this section we consider the reversible system of the form

$$\dot{x} = \omega, \quad \dot{u} = Av, \quad \dot{v} = -Bu, \tag{2.1}$$

where $(x, u, v) \in T^n \times R^p \times R^p$, A and B being p-order square matrices. Moreover, we take A to be nonsingular. We want to transform the reversible system (2.1) to a simple, symmetrical form, which we call normal form.

At first we consider a special class of compatible transformations. Let a mapping $\Phi : (x, u, v) \to (x_+, u_+, v_+)$, defined by

$$x_+ = x, \quad u_+ = \phi_{11}(x)u + \phi_{12}(x)v, \quad v_+ = \phi_{21}(x)u + \phi_{22}(x)v, \tag{2.2}$$

where ϕ_{ij} $(i, j = 1, 2)$ are $p \times p$ matrices. By definition, Φ is compatible if and only if

$$\begin{aligned} &\phi_{11}(x) = \phi_{11}(-x), \quad \phi_{12}(x) = -\phi_{12}(-x), \\ &\phi_{22}(x) = \phi_{22}(-x), \quad \phi_{21}(x) = -\phi_{21}(-x) \end{aligned}$$

and the matrix $(\phi_{ij})_{1\le i,j\le 2}$ is nonsingular. In particular, if $\phi_{12} = \phi_{21} = 0$ and ϕ_{11} and ϕ_{22} are nonsingular constant matrices, then Φ is a linear compatible transformation. Under this compatible transformation, the reversible system (2.1) is changed to

$$\dot{x} = \omega, \quad \dot{u} = A_+v, \quad \dot{v} = -B_+u, \tag{2.3}$$

where $A_+ = \phi_{11}A\phi_{22}^{-1}$ and $B_+ = \phi_{22}B\phi_{11}^{-1}$. Note that for simplicity we always use (x, u, v) instead of the new variables (x_+, u_+, v_+) in the transformed equations. Taking $\phi_{11} = I_p$ and $\phi_{22} = A$, we have $A_+ = I_p$, $B_+ = AB$.

From hypothesis (A.2) there exists a nonsingular matrix S such that

$$SB_+S^{-1} = \operatorname{diag}(B_{+1}, \dots, B_{+d}) \ \text{ with } B_{+j} = \lambda_j^2 I_{p_j} + J_j,$$

where λ_j^2 $(j = 1, \dots, d)$ are all different eigenvalues of B_+, and $J_j = \operatorname{diag}(J_{j1}, \dots, J_{jd_j})$ is a p_j-order matrix, where $J_{jj'}$ is a zero-matrix or a Jordan form (b_{lm}) satisfying $b_{lm} = 1$ for $m = l + 1$ and $b_{lm} = 0$ for $m \ne l + 1$.

It is easy to see that for $\lambda_j \ne 0$, $I_{p_j} + \lambda_j^{-2}J_j$ and $(I_{p_j} + J_j)^2$ are similar . Thus, there exists a nonsingular matrix $\tilde{S}_j$ such that $\tilde{S}_j(\lambda_j^2 I_{p_j} + J_j)\tilde{S}_j^{-1} = [\lambda_j(I_{p_j} + J_j)]^2$. Suppose $\lambda_j \ne 0$ for $j = 1, \dots, d-1$ and $\lambda_d = 0$. Let $\tilde{S} = \operatorname{diag}(\tilde{S}_1, \tilde{S}_2, \dots, \tilde{S}_{d-1}, I_{p_d})$. Under a compatible transformation of the form (2.2) with $\phi_{12} = \phi_{21} = 0$ and $\phi_{11} = \phi_{22} = \tilde{S}S$, the reversible system (2.3) becomes

$$\dot{x} = \omega, \quad \dot{u} = I_pv, \quad \dot{v} = -\tilde{B}u, \tag{2.4}$$

where $\tilde{B} = \operatorname{diag}(\tilde{B}^1, \dots, \tilde{B}^d)$ with $\tilde{B}^j = \lambda_j^2(I_{p_j} + J_j)^2$ for $j \le d - 1$ and $\tilde{B}_d = J_d$.

Taking a compatible transformation of the form (2.2) with

$$\phi_{11} = \operatorname{diag}(\lambda_1 E_1, \dots, \lambda_{d-1}E_{d-1}, I_{p_d}), \quad \phi_{22} = I_p, \quad \phi_{12} = \phi_{21} = 0,$$

where $E_j = I_{p_j} + J_j$, system (2.4) is changed to

$$\dot{x} = \omega, \quad \dot{u} = A_*v, \quad \dot{v} = -B_*u, \tag{2.5}$$

where

$$A_* = \operatorname{diag}(\lambda_1 E_1, \dots, \lambda_{d-1}E_{d-1}, I_{p_d}), \ \ B_* = \operatorname{diag}(\lambda_1 E_1, \dots, \lambda_{d-1}E_{d-1}, J_d).$$

We call the form (2.5) a normal form for the reversible system. This result is stated concisely in the following lemma.

LEMMA 2.1. *The reversible system* (2.1) *can always be changed to the normal form* (2.5) *by compatible transformation. Moreover, if all the eigenvalues of* AB *are not zero, we have* $\lambda_j \neq 0$, $j = 1, 2, \ldots, d$, *and* $A_* = B_* = \operatorname{diag}(\lambda_1 E_1, \ldots, \lambda_d E_d)$.

Now we consider a small perturbation of normal form (2.5), namely the following reversible systems:

$$\dot{x} = \omega, \quad \dot{u} = (A + g_1)v, \quad \dot{v} = -(B + g_2)u, \tag{2.6}$$

where $A = A^*, B = B^*$.

LEMMA 2.2. *There exists a compatible transformation* Φ *satisfying* $\|\Phi - Id\|^L \leq c\epsilon$ *by which* (2.6) *is changed to*

$$\dot{x} = \omega, \quad \dot{u} = A_+ v, \quad \dot{v} = -B_+ u, \tag{2.7}$$

where $A_+ = \operatorname{diag}(A_{+1}, \ldots, A_{+d})$ *and* $B_+ = \operatorname{diag}(B_{+1}, \ldots, B_{+d})$, *with* $A_{+j} = A_j + \hat{A}_j, B_{+j} = A_{+j}$ *for* $j = 1, 2, \ldots, d-1$, *and* $A_{+d} = I_{p_d}$, $B_{+d} = J_d + \hat{B}_d$. *Moreover,* $\|A_+ - A\|^L, \|B_+ - B\|^L \leq c\epsilon$.

Proof. Since A has the eigenvalues $\lambda_1, \ldots, \lambda_{d-1}, 1$ satisfying $|\lambda_j| \geq \delta > 0$ for $1 \leq j \leq d-1$, we have

$$A + g_1 = A(I_p + A^{-1} g_1) \quad \text{with} \quad \|A^{-1} g_1\| \leq c\epsilon,$$

where c depends on δ and p. If ϵ is sufficiently small, $I_p + A^{-1} g_1$ is nonsingular, and

$$(I_p + A^{-1} g_1)^{-1} = I_p - A^{-1} g_1 + (A^{-1} g_1)^2 - \ldots = I_p + P.$$

It follows that $\|P\| \leq c\epsilon$. So $A + g_1$ is also nonsingular. With the compatible transformation $\Phi^1 : (x, u, v) \to (x_+, u_+, v_+)$, defined by

$$x_+ = x, \quad u_+ = (A + g_1)^{-1} u, \quad v_+ = I_p v,$$

the reversible system (2.6) is changed to

$$\dot{x} = \omega, \quad \dot{u} = I_p v, \quad \dot{v} = -\tilde{B} u, \tag{2.8}$$

where $\tilde{B} = AB + P'$, with $P' = g_2 A + B g_1 + g_2 g_1$ and $\|P'\| \leq c\epsilon$. Since

$$AB = \operatorname{diag}\big((\lambda_1 E_1)^2, \ldots, (\lambda_{d-1} E_{d-1})^2, J_d\big),$$

by Lemmas 6.2 and 6.3, if ϵ is sufficiently small, we have a nonsingular matrix S satisfying $\|S - I_p\| \leq |P'\| \leq c\epsilon$ such that

$$S\tilde{B}S^{-1} = AB + \operatorname{diag}(P_1, \ldots, P_d),$$

where the P_j are $p_j \times p_j$-matrices with $\|P_j\| \leq c\|P'\| \leq c\epsilon$.

Define a compatible transformation $\Phi^2 : (x, u, v) \to (x_+, u_+, v_+)$ by

$$x_+ = x, \;\; u_+ = Su, \;\; v_+ = Sv.$$

Then the reversible system (2.8) is changed to

$$\dot{x} = \omega, \quad \dot{u} = I_p v, \quad \dot{v} = -(AB + P'')u, \tag{2.9}$$

where $P'' = \operatorname{diag}(P_1, \ldots, P_d)$.

By Lemma 6.4, for sufficiently small ϵ, there exist $\hat{A}_j$ such that

$$(\lambda_j E_j)^2 + P_j = (\lambda_j E_j + \hat{A}_j)^2,\ 1 \le j \le d-1, \text{ and } \|\hat{A}_j\| \le c\epsilon.$$

Let $\hat{A} = \mathrm{diag}(\hat{A}_1, \dots, \hat{A}_{d-1}, 0)$ and $A_+ = A + \hat{A} = \mathrm{diag}(A_{+1}, \dots, A_{+d})$. We have $A_{+j} = \lambda_j E_j + \hat{A}_j$ $(j \le d-1)$ and $A_{+d} = I_{p_d}$. Let $B_{+j} = A_{+j}$ $(j \le d-1)$ and $B_{+d} = J_d + P_d$. Set $B_+ = \mathrm{diag}(B_{+1}, \dots, B_{+d})$.

Define a compatible transformation

$$\Phi^3 : (x, u, v) \to (x_+, u_+, v_+) \text{ by } x_+ = x,\ u_+ = A_+ u,\ v_+ = I_p v.$$

Thus, the reversible system (2.9) is changed to (2.7). By the compatible transformation $\Phi = \Phi^3\Phi^2\Phi^1$, the reversible system (2.6) is changed to the normal form (2.7). Moreover, Φ is given by $x_+ = x$ and $w_+ = \phi w$, where

$$w = (u, v)^T,\ w_+ = (u_+, v_+)^T, \text{ and } \phi = \mathrm{diag}(A_+ S(A + g_1)^{-1}, S).$$

Write the term

$$A_+ S(A + g_1)^{-1} = (A + \hat{A})S(A + g_1)^{-1} = I_p + P'''.$$

It follows easily that $\|P'''\| \le c\epsilon$. Thus we have $\|\phi - I_{2p}\| \le c\epsilon$. Moreover,

$$\|\phi - I_{2p}\|^L \le c\epsilon, \quad \|A_+ - A\|^L \le c\epsilon, \quad \|B_+ - B\|^L \le c\epsilon,$$

where c indicates constants independent of ϵ. Thus, Lemma 2.2 is proved.

3. A compatible transformation. Define a transformation $\Phi : (x, u, v) \to (x_+, u_+, v_+)$ by

$$\begin{cases} x_+ = x, \\ u_+ = (e^{\mathrm{i}\langle k,x\rangle} + e^{-\mathrm{i}\langle k,x\rangle})u - \mathrm{i}(e^{\mathrm{i}\langle k,x\rangle} - e^{-\mathrm{i}\langle k,x\rangle})v, \\ v_+ = \mathrm{i}(e^{\mathrm{i}\langle k,x\rangle} - e^{-\mathrm{i}\langle k,x\rangle})u + (e^{\mathrm{i}\langle k,x\rangle} + e^{-\mathrm{i}\langle k,x\rangle})v, \end{cases} \tag{3.1}$$

where $k \in Z^n$ is fixed. Obviously, this transformation is compatible.

For simplicity, let

$$E = \begin{pmatrix} I_p & -\mathrm{i}I_p \\ \mathrm{i}I_p & I_p \end{pmatrix}, \quad \bar{E} = \begin{pmatrix} I_p & \mathrm{i}I_p \\ -\mathrm{i}I_p & I_p \end{pmatrix},$$

so that (3.1) can be written in the more compact form

$$x_+ = x, \quad w_+ = \phi w, \tag{3.2}$$

where $\phi = e^{\mathrm{i}\langle k,x\rangle}E + e^{-\mathrm{i}\langle k,x\rangle}\bar{E}$. It follows that $E\bar{E} = 0$, $E^2 = 2E$, $E + \bar{E} = 2I_{2p}$, and so ϕ is invertible with

$$\phi^{-1} = \frac{1}{4}(e^{\mathrm{i}\langle k,x\rangle}\bar{E} + e^{-\mathrm{i}\langle k,x\rangle}E). \tag{3.3}$$

By means of this compatible transformation, we have the following result.

LEMMA 3.1. *Consider the reversible system*

$$\dot{x} = \omega, \quad \dot{u} = \lambda(I + J)v, \quad \dot{v} = -\lambda(I + J)u, \tag{3.4}$$

where for simplicity $I = I_{p_j}$ *and* $J = J_j$ *. If* $\lambda \neq 0$ *and* $\lambda + \langle k, \omega \rangle \neq 0$*, then there is a compatible transformation which carries over the system* (3.4) *to the system*

$$\dot{x} = \omega, \quad \dot{u} = (\lambda + \langle k, \omega \rangle)(I + J)v, \quad \dot{v} = -(\lambda + \langle k, \omega \rangle)(I + J)u. \tag{3.5}$$

Proof. Let

$$\Omega = \begin{pmatrix} 0 & A \\ -B & 0 \end{pmatrix}, \quad A = B = \lambda(I + J).$$

Then (3.4) can be written as $\dot{x} = \omega$, $\dot{w} = \Omega w$. Under the transformation (3.2) the reversible system (3.4) is changed to

$$x_+ = \omega, \quad \dot{w}_+ = (\partial_\omega \phi \phi^{-1} + \phi \Omega \phi^{-1}) w_+ = \Omega_+ w_+. \tag{3.6}$$

By direct calculation, we have

$$\Omega_+ = \begin{pmatrix} 0 & A + \langle k, \omega \rangle I \\ -B - \langle k, \omega \rangle I & 0 \end{pmatrix}.$$

Since $\lambda + \langle k, \omega \rangle \neq 0$, in the same way as in section 2, we have a compatible transformation that changes the reversible system (3.6) to (3.5). □

In the normal form (2.5), if λ_i and λ_j satisfy

$$\lambda_j - \lambda_i = \langle k, \omega \rangle, \qquad k \neq 0 \text{ and } \lambda_i \neq \langle k, \omega \rangle,$$

the normal form can be changed by a compatible transformation to another normal form with $\lambda_j = \lambda_i$. This means that the resonant case $\lambda_j - \lambda_i = \langle k, \omega \rangle$ is equivalent to the multiple case $\lambda_j = \lambda_i$ by a compatible transformation. Thus, in the proof of our results, we can always suppose that λ_i and λ_j satisfy $\lambda_j - \lambda_i \neq \langle k, \omega \rangle$ for all $k \in Z^n$.

4. Proof of Theorem 1.1. Below we use the KAM iteration to prove Theorem 1.1. By Lemma 2.1, without loss of generality, we suppose that $\Omega = \Omega^0$ has the form

$$\Omega^0 = \begin{pmatrix} 0 & A_0 \\ -B_0 & 0 \end{pmatrix},$$

where A_0 and B_0 have the same form as A_* and B_* in (2.5). Let $\lambda_d = 0$. By condition (1.7) and Lemma 3.1 we suppose that on the set O the nonresonance conditions hold:

$$\langle \omega, k \rangle \neq 0 \qquad \forall k \in Z^n \setminus \{0\}, \tag{4.1}$$

$$\langle \omega, k \rangle + \lambda_j(\omega) \neq 0 \qquad \forall k \in Z^n, \forall j = 1, \ldots, d-1, \tag{4.2}$$

$$\langle \omega, k \rangle + \lambda_i(\omega) - \lambda_j(\omega) \neq 0 \qquad \forall k \in Z^n, \ \forall i \neq j. \tag{4.3}$$

Moreover, suppose that there exist the following resonant relations:

$$2\lambda_j(\omega) = \langle \omega, k_j \rangle \ \forall \omega \in O, \qquad k_j \neq 0, \ j = 1, \ldots, d-1. \tag{4.4}$$

Remark. The nonresonant condition (4.3) is the second Melnikov's condition. By Lemma 3.1, the second Melnikov's condition can hold automatically by compatible transformation, so we need not give them as an additional condition.

Remark. Because of the resonant relations (4.4), in the KAM steps we have to retain some x-dependent terms from terms which are linear in u and v; this makes the KAM iteration more complicated.

A. *Outline of the iteration.* At each step we consider the reversible system

$$\begin{cases} \dot{x} = \tilde{\omega} + f(x,w;\omega), \\ \dot{w} = \Omega w + G(x)w + g(x,w;\omega), \end{cases} \tag{4.5}$$

where

$$\Omega = \begin{pmatrix} 0 & A \\ -B & 0 \end{pmatrix},$$

$$A = \mathrm{diag}(A_1,\dots,A_d),\ A_j = \lambda_j E_j + \hat{A}_j,\ 1 \le j \le d-1,\ A_d = I_{p_d},$$
$$B = \mathrm{diag}(B_1,\dots,B_d),\ B_j = A_j,\qquad 1 \le j \le d-1,\ B_d = J_d + \hat{B}_d.$$

Moreover, we have $\|\hat{A}_j\|^L \le c\epsilon_0$ for $1 \le j \le d-1$ and $\|\hat{B}_d\|^L \le c\epsilon_0$. The matrix G depends on x and satisfies $\|G\|^L \le c\epsilon_0$. Write as $G = (G_{lm})_{1\le l,m\le 2}$ in block form, where $G_{ij} = \mathrm{diag}(G_{ij}^1,\dots,G_{ij}^d)$ with $G_{ij}^d = 0$ for $i,\ j = 1,2$, and G_{ij}^l being a $p_l \times p_l$-matrix. Let $G^j = (G_{lm}^j)_{1\le l,m\le 2}$ with

$$G_{11}^j = \frac{\mathrm{i}}{2}\left(G_{k_j}^j e^{\mathrm{i}\langle k_j,x\rangle} + G_{-k_j}^j e^{-\mathrm{i}\langle k_j,x\rangle}\right),\quad G_{22}^j = -G_{11}^j,$$
$$G_{12}^j = -\frac{1}{2}\left(G_{k_j}^j e^{\mathrm{i}\langle k_j,x\rangle} - G_{-k_j}^j e^{-\mathrm{i}\langle k_j,x\rangle}\right),\quad G_{21}^j = G_{12}^j.$$

Let

$$S^j = \begin{pmatrix} I_{p_j} & I_{p_j} \\ \mathrm{i}I_{p_j} & -\mathrm{i}I_{p_j} \end{pmatrix} \quad \text{for } 1 \le j \le d-1,\ S^d = I_{p_d}. \tag{4.6}$$

It is easy to verify that

$$(S^j)^{-1} G^j S^j = \mathrm{i}\begin{pmatrix} 0 & G_{k_j}^j e^{\mathrm{i}\langle k_j,x\rangle} \\ G_{-k_j}^j e^{-\mathrm{i}\langle k_j,x\rangle} & 0 \end{pmatrix}$$

and

$$(S^j)^{-1}\begin{pmatrix} 0 & A_j \\ -B_j & 0 \end{pmatrix} S^j = \mathrm{i}\begin{pmatrix} A_j & 0 \\ 0 & -B_j \end{pmatrix}.$$

Remark. Gw consists of some special x-dependent terms which cannot be killed in KAM steps because of the resonant relations (4.4). These terms are retained through the iteration step.

Considering small perturbations, suppose that f and g are analytic on $D(s,r)$ in (x,w) and belong to $C^L(O)$ in ω with

$$\|f\|^L_{D(s,r)} \le \epsilon,\quad \|g\|^L_{D(s,r)} \le r\epsilon. \tag{4.7}$$

The idea of our KAM iteration is to find a compatible transformation such that the system (4.5) can be changed to

$$\begin{cases} \dot{x} = \tilde{\omega}_+ + f_+(x,w;\omega), \\ \dot{w} = \Omega_+ w + G_+(x)w + g_+(x,w;\omega) \end{cases} \tag{4.8}$$

with f_+ and g_+ being a much smaller perturbation and $\tilde{\omega}_+, \Omega_+, G_+(x)$ being the corrections of $\tilde{\omega}, \Omega, G(x)$, respectively. If this correction can be carried out at each step of the iteration, then, after infinitely many steps, the system converges to the form in Theorem 1.1.

B. *Normalization.* We first normalize the first order constant terms of w. Let $f^0 = f(x,w)|_{w=0}$, $[f^0]$ being the average of f^0 with respect to x on T^n, and $\hat{\Omega} = [\mathcal{D}_w g(x,w)\,|_{w=0}]$. Since the system is reversible, we have

$$\hat{\Omega} = \begin{pmatrix} 0 & \mathcal{D}_v g_1(x,u,v)\,|_{u,v=0} \\ \mathcal{D}_u g_2(x,u,v)\,|_{u,v=0} & 0 \end{pmatrix}.$$

By Lemma 2.2, we have a linear compatible transformation Φ^1 defined by $x_+ = x, w_+ = Sw$ satisfying

$$\|S - I_{2p}\|^L \le c\epsilon \tag{4.9}$$

such that the system (4.5) is changed to

$$\begin{cases} \dot{x} = \tilde{\tilde{\omega}} + \big(f(x,S^{-1}w;\omega) - [f^0]\big), \\ \dot{w} = \tilde{\Omega} w + SG(x)S^{-1}w + Sg(x,S^{-1}w;\omega) - S\hat{\Omega}S^{-1}w, \end{cases} \tag{4.10}$$

where $\tilde{\tilde{\omega}} = \tilde{\omega} + [f^0]$ and $\tilde{\Omega} = \left(\begin{smallmatrix} 0 & \tilde{A} \\ -\tilde{B} & 0 \end{smallmatrix}\right)$ with

$$\begin{aligned} &\tilde{A} = \operatorname{diag}(\tilde{A}_1, \dots, A_d), \quad \tilde{A}_j = \lambda_j E_j + \hat{\tilde{A}}_j,\ 1 \le j \le d-1, \quad \tilde{A}_d = I_{p_d}, \\ &\tilde{B} = \operatorname{diag}(\tilde{B}_1, \dots, \tilde{B}_d), \quad \tilde{B}_j = \tilde{A}_j, \qquad\qquad 1 \le j \le d-1, \quad \tilde{B}_d = J_d + \hat{\tilde{B}}_d\,. \end{aligned}$$

Moreover, we have the estimates

$$\|\hat{\tilde{A}}_j - \hat{A}_j\|^L \le c\epsilon \ \text{ for } \ 1 \le j \le d-1, \ \ \|\hat{\tilde{B}}_d - \hat{B}_d\|^L \le c\epsilon. \tag{4.11}$$

Write $SG(x)S^{-1} = G + (S - I_{2p})GS^{-1} + SG(S^{-1} - I_{2p})$. Let

$$\begin{aligned} &\tilde{f} = \big(f(x,S^{-1}w;\omega) - [f^0]\big), \\ &\bar{g} = S\big(g(x,S^{-1}w;\omega) - \hat{\Omega}S^{-1}w\big) + \big[(S - I_{2p})GS^{-1} + SG(S^{-1} - I_{2p})\big]w. \end{aligned}$$

It is easy to see that

$$\|\tilde{f}\|^L_{D(s,r/2)} \le 2\epsilon, \quad \|\bar{g}\|^L_{D(s,r/2)} \le cr\epsilon. \tag{4.12}$$

Note that in the KAM steps we use c to indicate constants which are independent of the iteration process.

Let

$$\bar{g}^1 = \frac{\partial \bar{g}}{\partial w}\bigg|_{w=0} = \begin{pmatrix} \bar{g}_{11} & \bar{g}_{12} \\ \bar{g}_{21} & \bar{g}_{22} \end{pmatrix}.$$

Write $\bar g_{ij}=(\bar g_{ij}^{lm})_{1\le l,m\le d}$, where $\bar g_{ij}^{lm}$ is a $p_l\times p_m$-matrix. Take

$$\begin{aligned}
\hat G_{11}^j =& -\frac14\Big\{-\bar g_{11k_j}^{jj}e^{\mathrm{i}\langle k_j,x\rangle}-\bar g_{11-k_j}^{jj}e^{-\mathrm{i}\langle k_j,x\rangle}+\bar g_{22k_j}^{jj}e^{\mathrm{i}\langle k_j,x\rangle}+\bar g_{22-k_j}^{jj}e^{-\mathrm{i}\langle k_j,x\rangle}\\
&+\mathrm{i}\big[\bar g_{12k_j}^{jj}e^{\mathrm{i}\langle k_j,x\rangle}-\bar g_{12-k_j}^{jj}e^{-\mathrm{i}\langle k_j,x\rangle}+\bar g_{21k_j}^{jj}e^{\mathrm{i}\langle k_j,x\rangle}-\bar g_{21-k_j}^{jj}e^{-\mathrm{i}\langle k_j,x\rangle}\big]\Big\},\\
\hat G_{21}^j =& -\frac14\Big\{-\mathrm{i}\big[\bar g_{11k_j}^{jj}e^{\mathrm{i}\langle k_j,x\rangle}-\bar g_{11-k_j}^{jj}e^{-\mathrm{i}\langle k_j,x\rangle}\big]\\
&+\mathrm{i}\big[\bar g_{22k_j}^{jj}e^{\mathrm{i}\langle k_j,x\rangle}-\bar g_{22-k_j}^{jj}e^{-\mathrm{i}\langle k_j,x\rangle}\big]\\
&-\bar g_{12k_j}^{jj}e^{\mathrm{i}\langle k_j,x\rangle}-\bar g_{12-k_j}^{jj}e^{-\mathrm{i}\langle k_j,x\rangle}-\bar g_{21k_j}^{jj}e^{\mathrm{i}\langle k_j,x\rangle}-\bar g_{21-k_j}^{jj}e^{-\mathrm{i}\langle k_j,x\rangle}\Big\},\\
\hat G_{22}^j =& -\hat G_{11}^j,\ \ \hat G_{12}^j=\hat G_{21}^j \text{ for } j=1,\dots,d-1,\ \ \hat G_{ij}^d=0,\\
\hat G_{ij} =& \operatorname{diag}\big(\hat G_{ij}^1,\dots,\hat G_{ij}^{d-1},0\big),\ \hat G=\big(\hat G_{ij}\big)_{1\le i,j\le 2},\ \hat G^j=\begin{pmatrix}\hat G_{11}^j & \hat G_{12}^j\\ \hat G_{21}^j & \hat G_{22}^j\end{pmatrix}.
\end{aligned}$$

In addition, set

$$\tilde G=G+\hat G,\ \tilde g=\bar g-\hat G w,\ \text{and } \tilde g^1=\frac{\partial\tilde g}{\partial w}\Big|_{w=0}=\bar g^1-\hat G. \tag{4.13}$$

It follows easily that $\|\tilde g\|^L_{D(s,r/2)}\le cr\epsilon$. The system (4.10) becomes

$$\begin{cases}\dot x &= \tilde{\tilde\omega} \qquad\qquad +\tilde f,\\ \dot w &= \tilde\Omega w+\tilde G(x)w+\tilde g.\end{cases} \tag{4.14}$$

By the above discussion, we have $[\tilde f\ |_{w=0}]=0$ and $[\tilde g_w\ |_{w=0}]=0$. In particular, we take the terms $\tilde g_{ij}^{ll}$ from $\tilde g^1$ in the same way as $\bar g_{ij}^{ll}$ from $\bar g^1$ and let

$$(\tilde g_{ij}^l)_{1\le i,j\le 2}=(S^l)^{-1}(\tilde g_{ij}^{ll})_{1\le i,j\le 2}S^l.$$

Then for the Fourier coefficients of $\tilde g_{ij,k}^l$ we have

$$\tilde g_{12,k_l}^l=0\ \text{ and }\ \tilde g_{21,-k_l}^l=0. \tag{4.15}$$

C. *Constructing compatible transformation.* Define $\Phi^2:(x_+,w_+)\to(x,w)$ by

$$x=x_++h(x_+),\quad w=\big(I_{2p}+a(x_+)\big)w_++b(x_+),$$

where $h(x)$, $b(x)$ are vector functions, respectively, and $a(x)$ is a $2p\times 2p$-matrix function. Write a and b in the block form

$$a=\begin{pmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{pmatrix},\quad b=\begin{pmatrix}b_1\\ b_2\end{pmatrix},$$

where a_{ij} are $p\times p$-matrices and b_1,b_2 are p-dimensional vectors. By definition, it is easy to see that Φ^2 is compatible if and only if

$$\begin{cases}h(-x) &= -h(x),\\ b_1(-x) &= -b_1(x), \quad b_2(-x)=b_2(x),\\ a_{11}(-x)=a_{11}(x), &\quad a_{12}(-x)=-a_{12}(x),\\ a_{21}(-x)=-a_{21}(x), &\quad a_{22}(-x)=\ a_{22}(x).\end{cases} \tag{4.16}$$

Before doing the transformation, we first consider the composition function $\tilde{G}(x+h(x))$. Suppose that $\|h(x)\|^L_{D(s-\rho,r/2)} \le 1$. By series expansion, it follows that

$$e^{\mathrm{i}\langle k_j,h(x)\rangle} = 1 + \mathrm{i}\langle k_j, h(x)\rangle + O_j(h^2),$$

where $O_j(h^2)$ satisfies $\|O_j(h^2)\| \le c\|h\|^2$ as $h \to 0$. Let

$$\tilde{G}^j_*(x) = -\mathrm{i}(S^j)^{-1}\tilde{G}^j S^j = \begin{pmatrix} 0 & \tilde{G}^j_{k_j} e^{\mathrm{i}\langle k_j,x\rangle} \\ \tilde{G}^j_{-k_j} e^{-\mathrm{i}\langle k_j,x\rangle} & 0 \end{pmatrix}.$$

Then we have

$$\tilde{G}^j_*(x+h) = \tilde{G}^j_* + \tilde{G}^j_* \operatorname{diag}\big(-\mathrm{i}\langle k_j,h\rangle, \mathrm{i}\langle k_j,h\rangle\big) + \tilde{G}^j_* \operatorname{diag}\big(O_j(-h^2), O_j(h^2)\big)\big).$$

With

$$\begin{aligned} \tilde{G}^j_h(x) &= \mathrm{i}S^j\tilde{G}^j_*(x)\operatorname{diag}\big(-\mathrm{i}\langle k_j,h\rangle, \mathrm{i}\langle k_j,h\rangle\big)(S^j)^{-1}, \\ \tilde{G}^j_{h^2}(x) &= \mathrm{i}S^j\tilde{G}^j_*(x)\operatorname{diag}\big(O_j(-h^2), O_j(h^2)\big)(S^j)^{-1} \end{aligned}$$

we obtain

$$\tilde{G}^j(x+h(x)) = \tilde{G}^j(x) + \tilde{G}^j_h(x) + \tilde{G}^j_{h^2}(x).$$

In the same way as defining G from G^j, we define $\tilde{G}(x+h(x))$, $\tilde{G}(x)$, $\tilde{G}_h(x)$, and $\tilde{G}_{h^2}(x)$. It is easy to see that

$$\tilde{G}(x+h(x)) = \tilde{G}(x) + \tilde{G}_h(x) + \tilde{G}_{h^2}(x).$$

Moreover, we have

$$\|\tilde{G}_h\|^L_{D(s-\rho,r/2)} \le c\|h\|^L_{D(s-\rho,r/2)}, \quad \|\tilde{G}_{h^2}\|^L_{D(s-\rho,r/2)} \le c(\|h\|^L_{D(s-\rho,r/2)})^2.$$

In particular, if $[h(x)] = 0$, $\tilde{G}_h$ has the same property as (4.15) for $\tilde{g}^1$.

Under the transformation Φ_2 the system (4.14) is changed to

$$\begin{cases} \dot{x} = \tilde{\tilde{\omega}} + (I_n + \mathcal{D}_x h)^{-1}(\tilde{f}^0 - \partial_{\tilde{\tilde{\omega}}} h) + R_0 = \tilde{\tilde{\omega}} + f_+, \\ \dot{w} = (\tilde{\Omega} + \tilde{G})w + (I_{2p} + a)^{-1}[(\tilde{\Omega} + \tilde{G})b - \partial_{\tilde{\tilde{\omega}}} b + \tilde{g}^0 \\ \qquad + (\tilde{\Omega} + \tilde{G})aw - a(\tilde{\Omega} + \tilde{G})w - \partial_{\tilde{\tilde{\omega}}} aw + (\tilde{g}^1 + \tilde{G}_h)w] + R \\ \qquad = (\tilde{\Omega} + \tilde{G})w + g_+, \end{cases} \tag{4.17}$$

where

$$\tilde{f}^0 = \tilde{f}(x,0), \;\; \tilde{g}^0 = \tilde{g}(x,0), \;\; \tilde{g}^1 = \tilde{g}_w(x,0), \;\; \partial_{\tilde{\tilde{\omega}}} h = \sum_k \mathrm{i}\langle k, \tilde{\tilde{\omega}}\rangle h_k e^{\mathrm{i}\langle k,x\rangle},$$

$$R_0 = (I_n + \mathcal{D}_x h)^{-1}\left(\tilde{f} - \tilde{f}^0 + \int_0^1 < \nabla_z \tilde{f}(z + t\delta z), \delta z > \, dt\right) \tag{4.18}$$

with $\mathcal{D}_x h$ the Jacobian matrix of h with respect to x, $z = (x, w)$, and $\delta z = (h, aw + b)$,

$$\begin{aligned} R = {} & (I_{2p} + a)^{-1}\big[(\tilde{G}_h + \tilde{G}_{h^2})(aw + b) + G_{h^2} w - (\partial_{f_+} aw + \partial_{f_+} b)\big] \\ & + \; (I_{2p} + a)^{-1}\left[\tilde{g} - \tilde{g}^0 - \tilde{g}^1 w + \int_0^1 < \nabla_z \tilde{g}(z + t\delta z), \delta z > \, dt\right]. \end{aligned} \tag{4.19}$$

Here $\partial_f a = (\partial_f a_{ij})_{ij}$ is also a matrix.

We want to find h, a, b such that

$$-\partial_{\tilde{\omega}} h + \tilde{f}^0 = 0, \tag{4.20}$$

$$(\tilde{\Omega} + \tilde{G})b - \partial_{\tilde{\omega}} b + \tilde{g}^0 = 0, \tag{4.21}$$

$$(\tilde{\Omega} + \tilde{G})a - a(\tilde{\Omega} + \tilde{G}) - \partial_{\tilde{\omega}} a + \tilde{g}^1 + \tilde{G}_h = 0. \tag{4.22}$$

Thus, the system (4.17) becomes

$$\begin{cases} \dot{x} = \tilde{\omega}_+ \qquad\qquad + f_+, \\ \dot{w} = (\Omega_+ + G_+)w + g_+, \end{cases} \tag{4.23}$$

where $\tilde{\omega}_+ = \tilde{\tilde{\omega}},\ \Omega_+ = \tilde{\Omega},\ G_+ = \tilde{G},\ f_+ = R_0,\ g_+ = R.$

D. *Solving linear homological equations.* As usual, the first equation (4.20) is easy to solve. By condition (4.1) and Lemma 6.5, there exists a subset O_+ of O such that for $\omega \in O_+$, we have $|\langle\tilde{\tilde{\omega}}, k\rangle| \geq \alpha/|k|^\tau$ for all $0 \neq k \in Z^n$, and $\tau > n-1$. By $[\tilde{f}^0] = 0$, for $\omega \in O_+$ we obtain $h_k = \tilde{f}^0_k / \mathrm{i}\langle\tilde{\tilde{\omega}}, k\rangle$ for all $k \neq 0$, where h_k and $\tilde{f}^0_k$ are the Fourier coefficients of h and $\tilde{f}^0$. Since $\|\tilde{f}^0\|^L_{D(s,r/2)} \leq \|\tilde{f}\|^L_{D(s,r/2)} \leq \epsilon$, it follows that

$$\|h\|^L_{D(s-\rho,r/2)} \leq c\epsilon\, \alpha^{-L-1}\rho^{-\kappa}, \quad \kappa = 2\tau + n + 1. \tag{4.24}$$

Note that the measure of the set $O \setminus O_+$ will be estimated later.

Remark. In this paper, we can also use Rüssmann's nondegeneracy condition and Bruno's small divisor condition; for this case we refer the reader to [12, 13].

To solve (4.21), let $b = (b_1^T, \ldots, b_d^T, b_{d+1}^T, \ldots, b_{2d}^T)^T$, where b_j and b_{d+j} are p_j-dimensional column vectors. Let $y = \tilde{g}^0 = \tilde{g}(x, 0)$. Similarly, write $y = (y_1^T, \ldots, y_d^T, y_{d+1}^T, \ldots, y_{2d}^T)^T$. It follows easily that

$$\|y\|^L_{D(s,r/2)} \leq \|\tilde{g}\|^L_{D(s,r/2)} \leq cr\epsilon.$$

Let

$$\tilde{\Omega}^j = \begin{pmatrix} 0 & \tilde{A}_j \\ -\tilde{B}_j & 0 \end{pmatrix}.$$

By the special form of $\tilde{\Omega}$ and $\tilde{G}$, (4.21) is equivalent to

$$\begin{pmatrix} \partial_{\tilde{\omega}} b_j \\ \partial_{\tilde{\omega}} b_{d+j} \end{pmatrix} - \left(\tilde{\Omega}^j + \tilde{G}^j\right) \begin{pmatrix} b_j \\ b_{d+j} \end{pmatrix} = \begin{pmatrix} y_j \\ y_{d+j} \end{pmatrix},\ j = 1, 2, \ldots, d.$$

Let

$$(\tilde{b}_j^T, \tilde{b}_{d+j}^T)^T = (S^j)^{-1}(b_j^T, b_{d+j}^T)^T S^j, \quad (\tilde{y}_j^T, \tilde{y}_{d+j}^T)^T = (S^j)^{-1}(y_j^T, y_{d+j}^T)^T S^j,$$

where S^j is defined in (4.6). Set $\tilde{\Omega}^j_* = -\mathrm{i}(\mathrm{S}^\mathrm{j})^{-1}\tilde{\Omega}^\mathrm{j}\mathrm{S}^\mathrm{j}$. Then we have

$$\mathrm{i}\begin{pmatrix} \partial_{\tilde{\tilde{\omega}}} \tilde{b}_j \\ \partial_{\tilde{\tilde{\omega}}} \tilde{b}_{d+j} \end{pmatrix} + \tilde{\Omega}^j_* \begin{pmatrix} \tilde{b}_j \\ \tilde{b}_{d+j} \end{pmatrix} + \tilde{G}^j_* \begin{pmatrix} \tilde{b}_j \\ \tilde{b}_{d+j} \end{pmatrix} = \mathrm{i}\begin{pmatrix} \tilde{y}_j \\ \tilde{y}_{d+j} \end{pmatrix};$$

that is,

$$\begin{aligned} -\mathrm{i}\partial_{\tilde{\tilde{\omega}}} \tilde{b}_j - \tilde{A}_j \tilde{b}_j - \tilde{G}^j_{k_j} e^{\mathrm{i}\langle k_j, x\rangle} \tilde{b}_{d+j} &= -\mathrm{i}\tilde{y}_j, \\ -\mathrm{i}\partial_{\tilde{\tilde{\omega}}} \tilde{b}_{d+j} + \tilde{B}_j \tilde{b}_{d+j} - \tilde{G}^j_{-k_j} e^{-\mathrm{i}\langle k_j, x\rangle} \tilde{b}_j &= -\mathrm{i}\tilde{y}_{d+j}. \end{aligned}$$

Let $\tilde{b}_j = \sum_{k\in Z^n} \tilde{b}_{j,k} e^{\mathrm{i}\langle k,x\rangle}$. Then we obtain the following linear equations for the Fourier coefficients $\{\tilde{b}_{j,k}\}$:

$$[\langle k,\tilde{\tilde{\omega}}\rangle I_{p_j} - \tilde{A}_j]\tilde{b}_{j,k} - \tilde{G}^j_{k_j}\tilde{b}_{d+j,k-k_j} = -\mathrm{i}\tilde{y}_{j,k},$$
$$[\langle k-k_j,\tilde{\tilde{\omega}}\rangle I_{p_j} + \tilde{B}_j]\tilde{b}_{d+j,k-k_j} - \tilde{G}^j_{-k_j}\tilde{b}_{j,k} = -\mathrm{i}\tilde{y}_{d+j,k-k_j}.$$

Using (4.4), the coefficient matrix for $\tilde{b}_{j,k}$, and $\tilde{b}_{d+j,k-k_j}$,

$$\tilde{M} = \langle k,\tilde{\tilde{\omega}}\rangle I_{2p} - \mathrm{diag}(\lambda_j(I_{p_j}+J_j),\lambda_j(I_{p_j}-J_j)) + O(\epsilon_0) = M + O(\epsilon_0).$$

Obviously, M has only the eigenvalue $\langle k,\tilde{\tilde{\omega}}\rangle - \lambda_j$, and $O(\epsilon_0)$ is a small matrix. Suppose that

$$\|\tilde{\tilde{\omega}} - \tilde{\omega}\|^L,\ \|\tilde{A}_j - A_j\|^L,\ \|\tilde{B}_j - B_j\|^L,\ \|\tilde{G}^j_{\pm k_j}\|^L \le c\epsilon_0. \tag{4.25}$$

Then we have $\|O(\epsilon_0)\|^L \le c\epsilon_0$. By condition (4.2) and Lemmas 6.6 and 6.7, there exists a subset O_+ of O such that, for $\omega \in O_+$, $\tilde{M}$ is nonsingular and, for its inverse $\tilde{M}^{-1}$, we have

$$\|\tilde{M}^{-1}\|^L \le c(|k|+1)^\tau \alpha^{-L-1}.$$

Thus, we solve $\tilde{b}_{j,k}$ and $\tilde{b}_{d+j,k-k_j}$ and have the following estimates:

$$\|(\tilde{b}^T_{j,k},\tilde{b}^T_{d+j,k-k_j})^T\|^L \le c(|k|+1)^\tau\alpha^{-L-1}\|(\tilde{y}^T_{j,k},\tilde{y}^T_{d+j,k-k_j})^T\|^L \tag{4.26}$$

for $j = 1,\dots,d-1$ and for all $k \in Z^n$.

For $j = d$, we need to solve the following equations:

$$\begin{cases} \mathrm{i}\langle k,\tilde{\tilde{\omega}}\rangle\tilde{b}_{d,k} - \tilde{b}_{2d,k} = \tilde{y}_{d,k}, \\ \mathrm{i}\langle k,\tilde{\tilde{\omega}}\rangle\tilde{b}_{2d,k} + (J_d + \hat{\tilde{B}}_d)\tilde{b}_{d,k} = \tilde{y}_{2d,k}. \end{cases} \tag{4.27}$$

If $k \neq 0$, similarly to the above, we can find $\tilde{b}_{d,k}$ and $\tilde{b}_{2d,k}$ for $\omega \in O_+$ and have

$$\|(\tilde{b}^T_{d,k},\tilde{b}^T_{2d,k})^T\|^L_{D(s-\rho,r/2)} \le c(1+|k|)^\tau\alpha^{-L-1}\|(\tilde{y}^T_{d,k},\tilde{y}^T_{2d,k})^T\|^L. \tag{4.28}$$

For $k = 0$, (4.27) becomes $\tilde{b}_{2d,0} = -\tilde{y}_{d,0}$ and $(J_d + \hat{J}_d)\tilde{b}_{d,0} = \tilde{y}_{2d,0}$. Since $\tilde{y}_{2d,0} = 0$ by the special structure of reversible systems, we take $\tilde{b}_{d,0} = 0$. Thus, (4.28) still holds for $k = 0$. Combining the above estimates, for all $k \in Z^n$, we have

$$\|b_k\|^L \le c\|\tilde{b}_k\|^L \le c(|k|+1)^\tau\alpha^{-L-1}\|y_k\|^L.$$

Hence,

$$\|b(x)\|^L_{D(s-\rho,r/2)} \le cr\epsilon\alpha^{-L-1}\rho^{-\kappa}. \tag{4.29}$$

Now we consider the last equation, (4.22), which is the most difficult one. Let $a = (a_{ij})_{1\le i,j\le 2}$ and $y = \tilde{g}^1 + \tilde{G}_h = (y_{ij})_{1\le i,j\le 2}$, where a^{ij} and y^{ij} are $p\times p$-matrices. By (4.24), we have

$$\|y(x)\|^L_{D(s-\rho,r/2)} \le c\epsilon\alpha^{-L-1}\rho^{-\kappa}. \tag{4.30}$$

Write the matrices $a_{ij} = (a_{ij}^{lm})_{1\le l,m\le d}$ and $y_{ij} = (y_{ij}^{lm})_{1\le l,m\le d}$ in block form with a_{ij}^{lm} and y_{ij}^{lm} being $p_l \times p_m$-matrices. Define two $2p_l \times 2p_m$-matrices x^{lm} and y^{lm} by $x^{lm} = (a_{ij}^{lm})_{1\le i,j\le 2}$ and $y_{lm} = (y_{ij}^{lm})_{1\le i,j\le 2}$. Let

$$\tilde\Omega^j = \begin{pmatrix} 0 & \tilde A_j \\ -\tilde B_j & 0 \end{pmatrix} \text{ for } 1 \le j \le d-1 \text{ and } \tilde\Omega^d = \begin{pmatrix} 0 & I_{p_d} \\ J_d + \hat{\tilde B}_d & 0 \end{pmatrix}.$$

Then (4.22) becomes

$$\partial_{\tilde\omega} x^{ij} + x^{ij}(\tilde\Omega^j + \tilde G^j) - (\tilde\Omega^i + \tilde G^i)x^{ij} = y^{ij}. \tag{4.31}$$

Let $X = (S^i)^{-1}x^{ij}S^j = (X^{lm})_{1\le l,m\le 2}$, $Y = (S^i)^{-1}y^{ij}S^j = (Y^{lm})_{1\le l,m\le 2}$, with X^{lm}, Y^{lm} being $p_i \times p_j$-matrices. Then (4.31) changes to

$$\partial_{\tilde\omega} X + X(S^j)^{-1}(\tilde\Omega^j + \tilde G^j)S^j - (S^i)^{-1}(\tilde\Omega^i + \tilde G^i)S^i X = Y. \tag{4.32}$$

In the case of $i = j = d$, $\tilde G^d = 0$, so (4.32) becomes

$$\partial_{\tilde\omega} X - \tilde\Omega^d X + X\tilde\Omega^d = Y. \tag{4.33}$$

Hence, using Lemma 6.1, the equation for the Fourier coefficient $\{X_k\}$ is

$$\tilde M\{X_k\} = \big(\langle k, \tilde\omega\rangle I_{2p} + M + O(\epsilon_0)\big)\{X_k\} = -\mathrm{i}\{Y_k\},$$

where M has the only eigenvalue zero and $O(\epsilon_0)$ depends on $\hat{\tilde B}_d$ with the same estimates as before. For the notation $\{X_k\}$, see Lemma 6.1. In the same way as above, by Lemmas 6.6 and 6.7, there exists a subset O_+ of O such that, for $\omega \in O_+$, we can find $\{X_k\}$ for the above equation and obtain $\|X_k\|^L \le c|k|^\tau \alpha^{-L-1}\|Y_k\|^L$ for all $k \ne 0$. If $k = 0$, we take $X_0 = 0$ because of $Y_0 = 0$.

If $i = d$ and $1 \le j < d$, (4.32) becomes

$$\partial_{\tilde\omega} X + X(S^j)^{-1}(\tilde\Omega^j + \tilde G^j)S^j - \tilde\Omega^d X = Y. \tag{4.34}$$

Then

$$\begin{aligned}
&-\mathrm{i}\partial_{\tilde\omega} X^{11} + X^{11}\tilde A_j + e^{-\mathrm{i}\langle k_j,x\rangle}X^{12}\tilde G^j_{-k_j} + \mathrm{i}X^{21} = -\mathrm{i}Y^{11},\\
&-\mathrm{i}\partial_{\tilde\omega} X^{12} - X^{12}\tilde B_j + e^{\mathrm{i}\langle k_j,x\rangle}X^{11}\tilde G^j_{k_j} + \mathrm{i}X^{22} = -\mathrm{i}Y^{12},\\
&-\mathrm{i}\partial_{\tilde\omega} X^{21} + X^{21}\tilde A_j + e^{-\mathrm{i}\langle k_j,x\rangle}X^{22}\tilde G^j_{-k_j} - \mathrm{i}\tilde B_d X^{11} = -\mathrm{i}Y^{21},\\
&-\mathrm{i}\partial_{\tilde\omega} X^{22} - X^{22}\tilde B_j + e^{\mathrm{i}\langle k_j,x\rangle}X^{21}\tilde G^j_{k_j} - \mathrm{i}\tilde B_d X^{12} = -\mathrm{i}Y^{22}.
\end{aligned}$$

Comparing the Fourier coefficients on both sides of these equations and replacing k by $k|_j = k + k_j$ in the second and the fourth equation, we obtain

$$\begin{aligned}
&\langle k,\tilde\omega\rangle X^{11}_k + X^{11}_k\tilde A_j + X^{12}_{k|_j}\tilde G^j_{-k_j} + \mathrm{i}X^{21}_k = -\mathrm{i}Y^{11}_k,\\
&\langle k|_j,\tilde\omega\rangle X^{12}_{k|_j} - X^{12}_{k|_j}\tilde B_j + X^{11}_k\tilde G^j_{k_j} + \mathrm{i}X^{22}_{k_j} = -\mathrm{i}Y^{12}_{k|_j},\\
&\langle k,\tilde\omega\rangle X^{21}_k + X^{21}_k\tilde A_j + X^{22}_{k|_j}\tilde G^j_{-k_j} - \mathrm{i}\tilde B_d X^{11}_k = -\mathrm{i}Y^{21}_k,\\
&\langle k|_j,\tilde\omega\rangle X^{22}_{k|_j} - X^{22}_{k|_j}\tilde B_j + X^{21}_k\tilde G^j_{k_j} - \mathrm{i}\tilde B_d X^{12}_{k|_j} = -\mathrm{i}Y^{22}_{k|_j}.
\end{aligned}$$

Set

$$\tilde{X}_k = \begin{pmatrix} X_k^{11} & X_{k|_j}^{12} \\ X_k^{21} & X_{k|_j}^{22} \end{pmatrix}, \quad \tilde{Y}_k = \begin{pmatrix} Y_k^{11} & Y_{k|_j}^{12} \\ Y_k^{21} & Y_{k|_j}^{22} \end{pmatrix}, \quad \tilde{G}_{**}^j = \begin{pmatrix} 0 & \tilde{G}_{k_j}^j \\ \tilde{G}_{-k_j}^j & 0 \end{pmatrix}.$$

Then we find

$$\langle k, \tilde{\tilde{\omega}}\rangle \tilde{X}_k + \tilde{X}_k \Big[\mathrm{diag}\big(\tilde{A}_j, -\tilde{B}_j + \langle k_j, \tilde{\tilde{\omega}}\rangle I_{p_j}\big) + \tilde{G}_{**}^j \Big] - \mathrm{i}\tilde{\Omega}^{\mathrm{d}}\tilde{\mathrm{X}}_{\mathrm{k}} = -\mathrm{i}\tilde{\mathrm{Y}}_{\mathrm{k}}.$$

By Lemma 6.1, we consider the above matrix equation for $\tilde{X}_k$ as a linear equation for $\{\tilde{X}_k\}$. Denote the coefficient matrix for $\{\tilde{X}_k\}$ by

$$\tilde{M} = \langle k, \tilde{\tilde{\omega}}\rangle I_{4p_j p_d} + M + O(\epsilon_0),$$

where M has the only eigenvalue λ_j and $O(\epsilon_0)$ is a small matrix with the same estimates as before. By Lemma 6.6, there exists a subset O_+ of O such that, for $\omega \in O_+$, we can find $\tilde{X}_k$ with $\|\tilde{X}_k\|^L \le c(|k|+1)^\tau \alpha^{-L-1}\|\tilde{Y}_k\|^L$.

If $1 \le i, j < d$, (4.32) becomes

$$-\mathrm{i}\partial_{\tilde{\tilde{\omega}}}\mathrm{X} - (\tilde{\Omega}_*^{\mathrm{i}} + \tilde{\mathrm{G}}_{**}^{\mathrm{i}})\mathrm{X} + \mathrm{X}(\tilde{\Omega}_*^{\mathrm{j}} + \tilde{\mathrm{G}}_{**}^{\mathrm{j}}) = -\mathrm{i}\mathrm{Y}.$$

Using the abbreviation $\tilde{G}_{k_\alpha}^{x,\alpha} = e^{\mathrm{i}\langle k_\alpha, x\rangle}\tilde{G}_{k_\alpha}^{\alpha}$, we have

$$\begin{aligned}
&-\mathrm{i}\partial_{\tilde{\tilde{\omega}}}X^{11} - \tilde{A}_i X^{11} - \tilde{G}_{k_i}^{x,i} X^{21} + X^{11}\tilde{A}_j + X^{12}\tilde{G}_{-k_j}^{x,j} = -\mathrm{i}Y^{11},\\
&-\mathrm{i}\partial_{\tilde{\tilde{\omega}}}X^{12} - \tilde{A}_i X^{12} - \tilde{G}_{k_i}^{x,i} X^{22} - X^{12}\tilde{B}_j + X^{11}\tilde{G}_{k_j}^{x,j} = -\mathrm{i}Y^{12},\\
&-\mathrm{i}\partial_{\tilde{\tilde{\omega}}}X^{21} + \tilde{B}_i X^{21} - \tilde{G}_{-k_i}^{x,i} X^{11} + X^{21}\tilde{A}_j + X^{22}\tilde{G}_{-k_j}^{x,j} = -\mathrm{i}Y^{21},\\
&-\mathrm{i}\partial_{\tilde{\tilde{\omega}}}X^{22} + \tilde{B}_i X^{22} - \tilde{G}_{-k_i}^{x,i} X^{12} - X^{22}\tilde{B}_j + X^{21}\tilde{G}_{k_j}^{x,j} = -\mathrm{i}Y^{22}.
\end{aligned}$$

Comparing the Fourier coefficients and replacing k by $k|_j = k + k_j$, $k|_i = k - k_i$, and $k|_{ji} = k + k_j - k_i$ in the last three equations of the above system, respectively, we find

$$\begin{aligned}
&\langle k, \tilde{\tilde{\omega}}\rangle X_k^{11} - \tilde{A}_i X_k^{11} - \tilde{G}_{k_i}^i X_{k|_i}^{21} + X_k^{11}\tilde{A}_j + X_{k|_j}^{12}\tilde{G}_{-k_j}^j = -\mathrm{i}Y_k^{11},\\
&\langle k|_j, \tilde{\tilde{\omega}}\rangle X_{k|_j}^{12} - \tilde{A}_i X_{k|_j}^{12} - \tilde{G}_{k_i}^i X_{k|_{ji}}^{22} - X_{k|_j}^{12}\tilde{B}_j + X_k^{11}\tilde{G}_{k_j}^j = -\mathrm{i}Y_{k|_j}^{12},\\
&\langle k|_i, \tilde{\tilde{\omega}}\rangle X_{k|_i}^{21} + \tilde{B}_i X_{k|_i}^{21} - \tilde{G}_{-k_i}^i X_k^{11} + X_{k|_i}^{21}\tilde{A}_j + X_{k|_{ji}}^{22}\tilde{G}_{-k_j}^j = -\mathrm{i}Y_{k|_i}^{21},\\
&\langle k|_{ji}, \tilde{\tilde{\omega}}\rangle X_{k|_{ji}}^{22} + \tilde{B}_i X_{k|_{ji}}^{22} - \tilde{G}_{-k_i}^i X_{k|_j}^{12} - X_{k|_{ji}}^{22}\tilde{B}_j + X_{k|_i}^{21}\tilde{G}_{k_j}^j = -\mathrm{i}Y_{k|_{ji}}^{22}.
\end{aligned}$$

Set

$$\tilde{X}_k = \begin{pmatrix} X_k^{11} & X_{k|_j}^{12} \\ X_{k|_i}^{21} & X_{k|_{ij}}^{22} \end{pmatrix}, \quad \tilde{Y}_k = \begin{pmatrix} Y_k^{11} & Y_{k|_j}^{12} \\ Y_{k|_i}^{21} & Y_{k|_{ij}}^{22} \end{pmatrix}.$$

Then we have

$$\begin{aligned}
\langle k, \tilde{\tilde{\omega}}\rangle \tilde{X}_k + \tilde{X}_k[\mathrm{diag}(\tilde{A}_j, -\tilde{B}_j + \langle k_j, \tilde{\tilde{\omega}}\rangle I_{p_j}) + \tilde{G}_{**}^j]\\
-[\mathrm{diag}(\tilde{A}_i, -\tilde{B}_i + \langle k_i, \tilde{\tilde{\omega}}\rangle I_{p_i}) + \tilde{G}_{**}^i]\tilde{X}_k = -\mathrm{i}\tilde{Y}_k.
\end{aligned}$$

Using the resonant relations (4.4) in the same way as before, the above matrix equation is equivalent to a linear equation for $\{\tilde{X}_k\}$:

$$(\langle k, \tilde{\tilde{\omega}}\rangle I_{4p_i p_j} + M + O(\epsilon_0))\{\tilde{X}_k\} = -\mathrm{i}\{\tilde{Y}_k\},$$

where M has the eigenvalue $\lambda_j - \lambda_i$ and $O(\epsilon_0)$ has the same estimates as stated previously.

If $i \neq j$ or $i = j, k \neq 0$, by Lemmas 6.6 and 6.7, there exists a subset O_+ of O such that, for $\omega \in O_+$, the above linear equation for $\{\tilde{X}_k\}$ is solvable, and we have

$$\|\tilde{X}_k\|^L \leq c(|k|+1)^\tau \alpha^{-L-1} \|\tilde{Y}_k\|^L.$$

If $i = j$ and $k = 0$, by our choice of $\hat{G}$ and (4.15), it follows that $\tilde{Y}_0 = 0$, and, consequently, we have $\tilde{X}_0 = 0$.

Thus, for $\omega \in O_+$ we can find all X_k satisfying

$$\|X_k\|^L \leq c(|k|+1)^\tau \alpha^{-L-1} \|Y_k\|^L.$$

So

$$\|a_k\|^L \leq c(|k|+1)^\tau \alpha^{-L-1} \|y_k\|^L. \tag{4.35}$$

By (4.30) it follows that

$$\|a(x)\|^L_{D(s-2\rho,r/2)} \leq c\epsilon \alpha^{-2L-2} \rho^{-2\kappa}. \tag{4.36}$$

By (4.24), (4.29), and (4.36), there exists a subset O_+ of O such that, for $\omega \in O_+$, we can solve (4.20), (4.21), and (4.22) for h, b, a with the estimates

$$\|h\|^L_{D(s-\rho,r/2)},\ \frac{1}{r}\|b\|^L_{D(s-\rho,r/2)} \leq \frac{c\epsilon}{\alpha^{L+1}\rho^\kappa},\ \ \|a\|^L_{D(s-2\rho,r/2)} \leq \frac{c\epsilon}{\alpha^{2(L+1)}\rho^{2\kappa}}.$$

By Cauchy's estimate it follows that

$$\|\mathcal{D}_x h\|^L_{D(s-2\rho,r/2)},\ \frac{1}{r}\|\mathcal{D}_x b\|^L_{D(s-2\rho,r/2)} \leq \frac{c\epsilon}{\alpha^{L+1}\rho^{\kappa+1}},$$
$$\|\mathcal{D}_x a\|^L_{D(s-3\rho,r/2)} \leq \frac{c\epsilon}{\alpha^{2(L+1)}\rho^{2\kappa+1}}.$$

Now we consider the Lebesgue measure of O_+. For simplicity, denote by $\tilde{M}_k$ the matrices of coefficients for the linear equations in the above discussion. Thus

$$O_+ = \{\omega \in O \mid \forall k, \|\tilde{M}_k^{-1}\| \leq (1+|k|)^\tau/\alpha\},$$

where $\tilde{M}_k^{-1}$ is the inverse of $\tilde{M}_k$. We divide the set O_+ into two parts, O_+^1 and O_+^2, where

$$O_+^1 = \{\omega \in O \mid \forall |k| \leq K, \|\tilde{M}_k^{-1}\| \leq (1+|k|)^\tau/\alpha\}$$

and $O_+^2 = O_+ \setminus O_+^1$. By Lemma 6.6, if $\tau \geq nL+1$ and K is sufficiently large, we have

$$\operatorname{meas}(O - O_+^2) \leq c\alpha^{\frac{1}{L}} \sum_{k\neq 0} (|k|+1)^{\frac{L-\tau}{L}} \leq c\alpha^{\frac{1}{L}}. \tag{4.37}$$

For $|k| \leq K$, by the remark for Lemma 6.6, we need only consider $M_k = \tilde{M}_k|_{\epsilon_0=0}$. By the assumption of analyticity and conditions (4.1), (4.2), and (4.3), the set $\{\omega \,|\, \det(M_k) = 0\}$ has zero-measure. Let

$$O^1 = \{\omega \in O \mid \forall |k| \leq K, \|M_k^{-1}\| \leq (1+|k|)^\tau/(2\alpha_0)\},$$

where α_0 is the α at the first step with $\alpha \le \alpha_0$. We have $\mathrm{meas}(O \setminus O^1)$ tends to zero as $\alpha_0 \to 0$. Since $\tilde{M}_k$ is a small perturbation of M_k with $\|\tilde{M}_k - M_k\| \le c\epsilon_0$, if ϵ_0 is sufficiently small, we always have $O^1 \subseteq O^1_+$. Thus, we can take $O^1_+ = O^1$ and $O_+ = O^1 \cup O^2_+$, where O^1 is independent of the KAM step.

Below we verify the symmetry of (4.16). The symmetry of h holds by its definition. So we need only to consider b and a. Let $Q = \mathrm{diag}(-I_p, I_p)$. Then it follows easily that

$$Q\tilde{g}^0(-x) = -\tilde{g}^0(x),\ Q\tilde{G}(-x)Q = -\tilde{G}(x),\ \text{and}\ \ Q\tilde{g}^1(-x)Q = -\tilde{g}^1(x).$$

By (4.21), we have

$$(\tilde{\Omega} + \tilde{G}(-x))b(-x) + \partial_{\tilde{\omega}} b(-x) + \tilde{g}^0(-x) = 0.$$

Multiplying the above equation by Q from the left and using the properties of $\tilde{g}^0$ and $\tilde{G}$, we have

$$-(\tilde{\Omega} + \tilde{G}(x))Qb(-x) + \partial_{\tilde{\omega}} Qb(-x) + Q\tilde{g}^0(-x) = 0.$$

By (4.21) it follows that

$$(\tilde{\Omega} + \tilde{G}(x))(b(x) - Qb(-x)) - \partial_{\tilde{\omega}}(b(x) - Qb(-x)) = 0.$$

Because the solution of b in the above equation is unique, we have $b(x) = Qb(-x)$, which exhibits the symmetry of b in (4.16). Let $X = Qa(-x)Q - a$. By (4.22) and in the same way, we obtain

$$(\tilde{\Omega} + \tilde{G}(x))X - X(\tilde{\Omega} + \tilde{G}(x)) - \partial_{\tilde{\omega}} X = 0.$$

So $X = 0$ and $a = Qa(-x)Q$, which is equivalent to the symmetry of a.

E. *Estimates for the transformation.* By means of our construction, if

$$\omega \in O_+ \quad \text{and} \quad \frac{c\epsilon}{\alpha^{2(L+1)}\rho^{2\kappa+1}} \le \eta < \rho < \frac{1}{8},$$

we have a compatible transformation

$$\Phi^2 : (x_+, w_+) \in D(s - 3\rho, \eta r) \to (x, w) \in D(s - \rho, 2\eta r) \subset D(s, r/2)$$

such that

$$\|x - x_+\|^L_{D_+}, \qquad \frac{1}{r}\|w - w_+\|^L_{D_+} \le \frac{c\epsilon}{\alpha^{2(L+1)}\rho^{2\kappa}}, \tag{4.38}$$

$$\|\mathcal{D}_{x_+} x - I_n\|^L_{D_+}, \qquad \frac{1}{r}\|\mathcal{D}_{x_+} w\|^L_{D_+} \le \frac{c\epsilon}{\alpha^{2(L+1)}\rho^{2\kappa+1}}, \tag{4.39}$$

$$\|\mathcal{D}_{w_+} w - I_{2p}\|^L_{D_+} \le \frac{c\epsilon}{\alpha^{2(L+1)}\rho^{2\kappa}}, \tag{4.40}$$

where $D_+ = D(s - 3\rho, \eta r)$.

Define the compatible transformation $\Phi = \Phi^2 \circ \Phi^1 : (x_+, w_+) \to (x, w)$. By (4.9) it follows easily that Φ has the same structure and satisfies the same estimates as Φ^2, with possibly different constants. So we may regard Φ^2 as Φ for simplicity.

F. *Estimates of the perturbation after transformation.* Let

$$\eta = \left(\frac{\epsilon}{\alpha^{2(L+1)}\rho^{2\kappa+1}}\right)^{\frac{1}{2}}, \quad r_+ = r\eta, \quad s_+ = s - 3\rho.$$

By (4.18) and (4.19), using the procedure of [7] and [21], it follows that

$$\|f_+\|^L_{D(s_+,r_+)} \le c\epsilon\eta \quad = \epsilon_+, \tag{4.41}$$

$$\|g_+\|^L_{D(s_+,r_+)} \le c\epsilon\eta r_+ = \epsilon_+ r_+ \tag{4.42}$$

with $\epsilon_+ = c\epsilon\eta$. Here we omit the details of estimating, as they are very similar to those of [7] and [21].

G. *Convergence of the iteration.* For given s, ϵ, r in the theorems, we define several sequences, which depend inductively on s, ϵ, r:

$$\epsilon_0 = \epsilon, \quad r_0 = r, \quad s_0 = s, \quad \alpha_0 = \alpha, \quad \rho_0 = \frac{s_0}{12}, \quad \eta_0 = \frac{\epsilon_0^{\frac{1}{2}}}{\alpha_0^{L+1}\rho_0^{\kappa+\frac{1}{2}}},$$

$$A_0 = A, \quad B_0 = B, \quad G^0 = 0, \quad \Omega^0 = \begin{pmatrix} 0 & A_0 \\ -B_0 & 0 \end{pmatrix} = \text{normal form (2.5)},$$

$$\epsilon_{\nu+1} = c\eta_\nu\epsilon_\nu, \quad r_{\nu+1} = \eta_\nu r_\nu, \quad s_{\nu+1} = s_\nu - 3\rho_\nu, \quad \rho_{\nu+1} = \frac{1}{2}\rho_\nu,$$

$$\eta_\nu = \frac{\epsilon_\nu^{\frac{1}{2}}}{\alpha_\nu^{L+1}\rho_\nu^{\kappa+\frac{1}{2}}}, \quad D_\nu = D(s_\nu, r_\nu),\ \alpha_{\nu+1} = \alpha_\nu/2, \quad \nu = 0, 1, \ldots.$$

At the νth step, there exists a subset $O_\nu = O^1 \cap O^2_\nu$ of O. For $\omega \in O_\nu$, there exists a compatible transformation $\Phi_\nu : (x_{\nu+1}, w_{\nu+1}) \to (x_\nu, w_\nu)$ satisfying

$$\|x_{\nu+1} - x_\nu\|^L_{D_\nu} \le c\eta_\nu^2, \quad \|w_{\nu+1} - w_\nu\|^L_{D_\nu} \le cr_\nu\eta_\nu^2. \tag{4.43}$$

Moreover, we have $\text{meas}(O - O^2_\nu) \le c\alpha_\nu^{\frac{1}{L}}$.

Let $\Phi^{\nu+1} = \Phi^\nu \circ \Phi_{\nu+1}$ with $\Phi^0 = Id$ and $f_0 = f,\ g_0 = g,\ \tilde{\omega}^0 = \omega$. By the compatible transformation Φ^ν, system (1.1) is changed to

$$\begin{cases} \dot{x} = \tilde{\omega}^\nu \qquad\qquad + f^\nu(x, w; \omega), \\ \dot{w} = (\Omega^\nu + G^\nu)w + g^\nu(x, u, v; \omega). \end{cases} \tag{4.44}$$

Moreover,

$$\|\tilde{\omega}^{\nu+1} - \tilde{\omega}^\nu\|^L \le \epsilon_\nu, \quad \|\Omega^{\nu+1} - \Omega^\nu\|^L \le c\epsilon_\nu, \tag{4.45}$$

$$\|G^{\nu+1} - G^\nu\|^L \le c\epsilon_\nu, \quad \|f^\nu\|^L_{D_\nu} \le \epsilon_\nu, \quad \|g^\nu\|^L_{D_\nu} \le r_\nu\epsilon_\nu. \tag{4.46}$$

By definition, we have $\eta_{\nu+1} \le c^{\frac{1}{2}}(\eta_\nu)^{\frac{3}{2}}$; hence, $c\eta_{\nu+1} \le (c\eta_\nu)^{\frac{3}{2}}$. If $c\eta_0 = c\epsilon_0{}^{\frac{1}{2}}\alpha_0^{-L-1}\rho_0^{-\kappa-\frac{1}{2}} \le 2^{-1}$, then $\eta_{\nu+1} \le c^{-1}2^{-(\frac{3}{2})^\nu}$. Thus, we have $\epsilon_{\nu+1} \le 2^{-1}\epsilon_\nu$, and so $\epsilon_\nu \le 2^{-\nu}\epsilon_0$. By the above estimates, assumption (4.25) obviously holds.

Let $O_\alpha = \cap_{\nu\ge1}O_\nu$. For $(x, w; \omega) \in D(s/2, r/2) \times O_\alpha$, we can prove that the transformation Φ^ν is convergent to Φ_*. The proof is the same as in the case of Hamiltonian systems (in fact simpler), so we omit the details and refer the reader to [10]. For $\nu \to \infty$ let $\tilde{\omega}^\nu \to \omega_*,\ \Omega^\nu \to \Omega_*,\ G^\nu \to G_*,\ f^\nu \to f_*,\ g^\nu \to g^*$. By (4.45)

and (4.46), it follows that $\|\omega_* - \omega\|^L \le c\epsilon$, $\|\Omega_* - \Omega\|^L \le c\epsilon$, $\|G_*\|^L \le c\epsilon$. Moreover, by (4.46) we have

$$f_*(x,w)\mid_{w=0}= 0, \quad g^*(x,w)\mid_{w=0}= 0, \quad \text{and} \quad g^*_w(x,w)\mid_{w=0}= 0.$$

H. *Measure estimate.* It remains to estimate the measure of O_α. By the previous discussion, $O_\alpha = O^1 \cap O^2$, where $O^2 = \cap_{\nu=1}^{\infty} O^2_\nu$. It follows that

$$\mathrm{meas}(O \setminus O^2) \le \sum_{\nu\ge 1} \mathrm{meas}(O \setminus O^2_\nu) \le c\alpha^{\frac{1}{L}}.$$

Thus,

$$\mathrm{meas}(O \setminus O_\alpha) \le \mathrm{meas}(O \setminus O^1) + \mathrm{meas}(O \setminus O^2) \to 0 \ (\alpha \to 0).$$

For $\omega \in O_\alpha$, by the compatible transformation Φ_*, the reversible system (1.1) is changed to (1.9). Thus, the proof of Theorem 1.1 is complete .

5. Proof of Theorem 1.2. Take the matrices P and Q as in Theorem 1.2. By the compatible map $\Phi : (x,w) \to (x_+, w_+)$ defined by $x_+ = x, w_+ = \mathrm{diag}(P,Q)w$, the reversible system (1.1) is changed to

$$\begin{cases} \dot x = \omega & + f(x,u,v;\omega), \\ \dot u = (I_p, 0_2)v & + g_1(x,u,v;\omega), \\ \dot v = PBQu & + g_2(x,u,v;\omega). \end{cases} \tag{5.1}$$

Note that here f, g_1, g_2 may be different from those in (1.1). Let $\tilde u = (u_1, \dots, u_p, u_{p+1}, \dots, u_q)^T = (u^T, u_*^T)^T$ and $v_* = (v_{p+1}, \dots, v_q)^T$, where $u_* = (u_{p+1}, \dots, u_q)^T$. We consider the reversible system

$$\begin{cases} \dot x = \omega & + f(x,u,v;\omega), \\ \dot u = (I_p, 0_2)v & + g_1(x,u,v;\omega), \\ \dot u_* = I_{q-p} v_*, & \\ \dot v = (PBQ, 0_1)\tilde u & + g_2(x,u,v;\omega), \end{cases} \tag{5.2}$$

where $(PBQ, 0_1)$ is a $q \times q$-matrix. Using Theorem 1.1 we obtain a nonempty subset O_α of O. For $\omega \in O_\alpha$, we have a compatible transformation $\Phi(x, \tilde u, v; \omega)$ such that $\Phi(T^n, 0, 0; \omega)$ is an invariant torus of the reversible system (5.2). Taking projection maps $P_{u_*} : (x,u,u_*,v) \to (x,u,0,v)$ and $P_0 : (x,u,0,v) \to (x,u,v)$, then $P_0 \circ P_{u_*}(\Phi(T^n, 0, 0; \omega))$ is an invariant torus of the reversible system (5.1).

Remark. Here we show that condition (A.1) and (1.3) are necessary for the result. If $\mathrm{rank}(A) = \gamma < p$, there exists a compatible transformation such that the reversible system (1.1) is changed to

$$\begin{cases} \dot x = \omega & + f(x,u,v;\omega), \\ \dot u = \tilde I v & + g_1(x,u,v;\omega), \\ \dot v = PBQu & + g_2(x,u,v;\omega), \end{cases} \qquad \tilde I = \begin{pmatrix} I_\gamma & 0 \\ 0 & 0 \end{pmatrix}. \tag{5.3}$$

Let $g_1 = (0, 0, \dots, \epsilon)$. Then for all $\epsilon > 0$ the reversible system (5.3) has no invariant torus. From section 3, we know that if there are $\lambda_j \ne 0$ and $\lambda_j = \langle \omega, k \rangle$, then, by a compatible transformation, we arrive at the case $\lambda_j = 0$ and $\mathrm{rank}(A) < p$. This means that condition (1.7) is necessary.

6. Appendix.

LEMMA 6.1. *Let A and B be an $m \times m$ and $n \times n$-matrix, respectively, and let C and $X = (x_{ij})$ be $m \times n$-matrices. Then the matrix equation $AX + XB = C$ is equivalent to the following linear equation $D\{X\} = \{C\}$, where*

$$\{X\} = (x_{11}, x_{12}, \dots, x_{1n}, x_{21}, x_{22}, \dots, x_{2n}, \dots, x_{m1}, x_{m2}, \dots, x_{mn})^T,$$
$$\{C\} = (c_{11}, c_{12}, \dots, c_{1n}, c_{21}, c_{22}, \dots, c_{2n}, \dots, c_{m1}, c_{m2}, \dots, c_{mn})^T$$

are mn-columns and the coefficient matrix

$$D = \begin{pmatrix} a_{11}I_n + B^T & a_{12}I_n & \cdots & a_{1m}I_n \\ a_{21}I_n & a_{22}I_n + B^T & \cdots & a_{2m}I_n \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1}I_n & a_{m2}I_n & \cdots & a_{mn}I_n + B^T \end{pmatrix}$$

is an $mn \times mn$-matrix with the eigenvalues $\{\lambda_i + \mu_j \mid i = 1, \dots, m, j = 1, \dots, n\}$, if A has the eigenvalues $\lambda_1, \dots, \lambda_m$ and B has the eigenvalues $\mu_1, \dots, \mu_n$.

This lemma has a straightforward proof, which we will omit.

LEMMA 6.2. *Let A and B be an $m \times m$ and an $n \times n$-matrix, respectively, let $\lambda_1, \dots, \lambda_m$ be the eigenvalues of A and $\mu_1, \dots, \mu_n$ the eigenvalues of B, and let $P = (P_{ij})$ be an $(m+n) \times (m+n)$-matrix. If*

$$|\lambda_i - \mu_j| \geq \alpha > 0 \ \text{ for } \ i = 1, \dots, m, \ \ j = 1, \dots, n,$$

where α is constant, then there exists an $\epsilon > 0$ such that, if $\|P\| \leq \epsilon$, $\operatorname{diag}(A, B) + P$ is similar to a diagonal block form $\operatorname{diag}(A', B')$. Moreover, there is a nonsingular matrix $S = I_{m+n} + \hat{S}$ satisfying $\|\hat{S}\| \leq c\|P\|$ such that

$$S^{-1}\big(\operatorname{diag}(A, B) + P\big)S = \operatorname{diag}(A', B') = \operatorname{diag}(A + \hat{A}, B + \hat{B})$$

with $\|\hat{A}\| \leq c\|P\|$ and $\|\hat{B}\| \leq c\|P\|$.

Proof. Write

$$P = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix}, \quad S = \begin{pmatrix} I_m & S_{12} \\ S_{21} & I_n \end{pmatrix}, \quad \operatorname{diag}(A', B') = \begin{pmatrix} A + Q_{11} & 0 \\ 0 & B + Q_{22} \end{pmatrix}.$$

We solve the following equation for S_{12}, S_{21}, Q_{11}, and Q_{22}:

$$\begin{pmatrix} A + P_{11} & P_{12} \\ P_{21} & B + P_{22} \end{pmatrix} \times \begin{pmatrix} I_m & S_{12} \\ S_{21} & I_n \end{pmatrix} = \begin{pmatrix} I_m & S_{12} \\ S_{21} & I_n \end{pmatrix} \times \begin{pmatrix} A + Q_{11} & 0 \\ 0 & B + Q_{22} \end{pmatrix}.$$

This equation is equivalent to

$$\begin{cases} (A + P_{11})S_{12} - S_{12}B - S_{12}Q_{22} = -P_{12}, \\ (B + P_{22})S_{21} - S_{21}A - S_{21}Q_{11} = -P_{21}, \\ Q_{11} - P_{12}S_{21} = \ \ P_{11}, \\ Q_{22} - P_{21}S_{12} = \ \ P_{22}. \end{cases} \tag{6.1}$$

Using the last two equations in (6.1), the first two equations become

$$(A + P_{11})S_{12} - S_{12}(B + P_{22}) + S_{12}P_{21}S_{12} = -P_{12}, \tag{6.2}$$
$$(B + P_{22})S_{21} - S_{21}(A + P_{11}) + S_{21}P_{12}S_{21} = -P_{21}. \tag{6.3}$$

Instead of (6.2) for S_{12}, we now consider the equivalent equation for the column $\{S_{12}\}$; see Lemma 6.1. Let $\lambda'_1, \dots, \lambda'_m$ be the eigenvalues of $A + P_{11}$ and $\mu'_1, \dots, \mu'_n$ be the eigenvalues of $B + P_{22}$. At first, we take ϵ sufficiently small such that

$$|\lambda'_i - \mu'_j| \geq \frac{\alpha}{2} > 0 \text{ for } i = 1, \dots, m, \ \ j = 1, \dots, n.$$

By Lemma 6.1 the coefficient matrix for the linear part of $\{S_{12}\}$ has eigenvalues $\lambda'_i - \mu'_j$ for $i = 1, \dots, m$, $j = 1, \dots, n$. Thus, the coefficient matrix is nonsingular. The equation for $\{S_{12}\}$ is nonlinear. But by the implicit function theorem we can solve this equation near $\{P_{12}\} = 0$. Thus if $\{P_{12}\}$ is sufficiently small, there exists a unique solution $\{S_{12}\}$. Moreover, we have $\|S_{12}\| \leq c\|P\|$. Similarly, we can solve (6.3) with the same estimate $\|S_{21}\| \leq c\|P\|$. From the last two equations in (6.1) we can get Q_{11} and Q_{22} with $\|Q_{11}\| \leq c\|P\|$ and $\|Q_{22}\| \leq c\|P\|$. □

By induction we can easily obtain the following result.

LEMMA 6.3. *Let $\tilde{A}_j$ be an $n_j \times n_j$-matrix and $\lambda^j_1, \dots, \lambda^j_{m_j}$ be the eigenvalues of $\tilde{A}_j$. Let $P = (P_{ij})$ be a $\sum_j n_j \times \sum_j n_j$-matrix. If*

$$|\lambda^l_i - \lambda^m_j| \geq \alpha > 0 \text{ for } i = 1, \dots, n_l, \ \ j = 1, \dots, n_m, \ \ l \neq m,$$

where α is a constant, then there exists an $\epsilon > 0$ such that, if $\|P\| \leq \epsilon$, there is a nonsingular matrix $S = I_{\sum_j n_j} + \hat{S}$, satisfying $\|\hat{S}\| \leq c\|P\|$, with

$$S^{-1}\big(\operatorname{diag}(A_1, \dots, A_d) + P\big)S = \operatorname{diag}(A_1 + \hat{A}_1, \dots, A_d + \hat{A}_d).$$

Moreover, we have $\|\hat{A}_j\| \leq c\|P\|$ for $j = 1, \dots, d$.

LEMMA 6.4. *Let A and P be as in Lemma 6.2. Let $\{\lambda_j\}$ be the eigenvalues of the matrix A. Suppose that $\lambda_j = \lambda + \epsilon_j$, where $\lambda \neq 0$ and $|\epsilon_j| \leq \frac{1}{2}|\lambda|$ for all j. Suppose $\|P\| \leq \epsilon$. If ϵ is sufficiently small, then there exists a matrix X such that $(A + X)^2 = A^2 + P$ with $\|X\| \leq c\epsilon$.*

Proof. We consider the following equivalent equation $AX + XA + X^2 = P$. The coefficient matrix of the linear part $AX + XA$ for $\{X\}$ has all eigenvalues of $\{\lambda_i + \lambda_j = 2\lambda + \epsilon_i + \epsilon_j\}$. Because $|\lambda_i + \lambda_j| \geq 2|\lambda| - |\epsilon_i| - |\epsilon_j| \geq |\lambda|$, it is nonsingular. Using the same method as in the proof of Lemma 6.2 and by the implicit function theorem, we can obtain the solution of the above equation for $\{X\}$ with an estimate for small ϵ. □

LEMMA 6.5. *Let D be an $N \times N$-matrix depending on ω and $\|D\|^L \leq M$. Let P be an $N \times N$-matrix with $\|P\| \leq 1$. Then $\det(D + \epsilon P) = \det(D) + \epsilon F$ with $\|F\|^L \leq cM^{N-1}$, where c is a constant depending on N.*

Proof. Let $f(\epsilon) = \det(D + \epsilon P)$. Then $f(\epsilon) = f_0 + f_1\epsilon + f_2\epsilon^2 + \cdots + f_n\epsilon^N$, where $f_j = \frac{1}{j!}\frac{d^j f}{d\epsilon^j}\,|_{\epsilon=0}$. Obviously, $f_0 = \det(D)$. By the properties of determinants with respect to differentiation, it is easy to obtain $\|f_j\|^L \leq cM^{N-j}$, where c depends only on N.

LEMMA 6.6. *Suppose that $\|\tilde{\omega} - \omega\|^L \leq \epsilon$, where $f_j(\omega)$ is an Lth continuously differentiable function with $\|f_j(\omega)\|^L \leq M$ $(1 \leq j \leq L)$. Let $P(\omega) = (p_{ij}(\omega))$ be an $L \times L$-matrix, Lth continuously differentiable with respect to $\omega \in O$, with $\|P\|^L = \max_{|\beta| \leq L} \max_{1 \leq i,j \leq L} \sup_{\omega \in O} |\frac{\partial^\beta p_{ij}}{\partial \omega^\beta}| \leq \epsilon$. Let $\mathcal{R}_k(\alpha)$ be the subset of O such that $\|D^{-1}\| > \frac{c|k|^\tau}{\alpha}$, where $D = \langle \tilde{\omega}, k\rangle I + \operatorname{diag}(f_1(\omega), f_2(\omega), \dots, f_L(\omega)) + P(\omega)$, $\tau > nL$. Then, if $\epsilon > 0$ is sufficiently small, there is a constant $K > 0$ depending on M such*

that, for $\alpha > 0$ and $|k| > K$,

$$\mathrm{meas}(\mathcal{R}_k(\alpha)) \le \left(\frac{c\alpha}{|k|^{\tau-L}}\right)^{\frac{1}{L}},$$

where c is a constant depending on ϵ, M.

Proof. Since $\|D\| = O(|k|)$, we know that the norm of the inverse of a matrix is controlled by $|k|^L$ times of the lower bound of its determinant. In fact,

$$\mathcal{R}_k(\alpha) \subseteq \left\{\omega \in O \;\middle|\; \det D| < c\frac{\alpha}{|k|^{\tau-L}}\right\}.$$

Note that

$$g(k,\omega) = \det(D) = \prod_{j=1}^{L}(\langle\tilde{\omega}, k\rangle + f_j(\omega)) + \sum_{l=0}^{L-1} a_l \prod_{j=1}^{L}(\langle\tilde{\omega}, k\rangle + f_j(\omega))_j^l,$$

where $\|a_l\|^L \le c\epsilon$ and $l = (l_1, l_2, \ldots, l_L), l_j = 0$ or 1 and $\sum_{j=1}^{L} l_j \le L-1$. Also note that there exists a sufficiently large $K > 0$ such that, if $|k| \ge K$ and ϵ is sufficiently small,

$$\left|\frac{\partial^L}{\partial\nu^L} g(k,\omega)\right| \ge \left|\prod_{j=1}^{L}\left(\left\langle\frac{\partial}{\partial\nu}\tilde{\omega}, k\right\rangle + \frac{\partial}{\partial\nu} f_j(\omega)\right)\right| - c\epsilon|k|^L \ge [(1-c\epsilon)|k| - M]^L - c\epsilon|k|^L,$$

where $\frac{\partial^L}{\partial\nu^L}$ is the Lth direction derivative along the direction $\nu = \frac{k}{|k|}$ at ω. Thus, if ϵ is sufficiently small, it follows that $|\frac{\partial^L g(\omega)}{\partial\nu^L}| \ge \frac{1}{4}|k|^L$, which implies

$$\mathrm{meas}(\mathcal{R}_k(\alpha) \le c\left(\frac{\alpha}{|k|^{\tau-L}}\right)^{\frac{1}{L}} \mathrm{diam}(O)^{n-1},$$

where c is a constant depending only on M, ϵ, n, K, and, in particular, it is independent of α, k. ∎

Remark. If $f_j(\omega)$ is analytic and $\langle\omega, k\rangle + f_j(\omega) \ne 0$, then $\{\omega \mid \langle\omega, k\rangle + f_j(\omega) = 0\} \subset O$ has zero-measure. Let $O^1 = \{\omega \in O \mid |\langle\omega, k\rangle + f_j(\omega)| \ge \alpha \text{ for all } |k| \le K\}$. We have $\mathrm{meas}(\mathrm{O} \setminus \mathrm{O}^1) \to 0$ as $\alpha > 0 \to 0$. Therefore, $|\langle\omega, k\rangle + f_j(\omega)| \ge \alpha > 0$ holds on a nonempty subset O^1 of O. Then, for sufficiently small $\epsilon > 0$ depending on α, if $\|P\| \le \epsilon$, the matrix D is invertible for all $\omega \in O^1 \subset O$ and $|k| \le K$ with $\|D^{-1}\| \le \frac{c}{\alpha}$.

LEMMA 6.7. *Let D be a matrix depending on ω and $\|D\|^L \le M$. If D is invertible with $\|D^{-1}\| \le N$, then $\|D^{-1}\|^L \le cM^L N^{L+1}$, where c is a constant depending on L.*

Proof. By differentiating the two sides of the equation $D^{-1}D = I$, we have $(D^{-1})' = D^{-1}D'D^{-1}$. So

$$\|(D^{-1})'\| \le \|D'\| \cdot \|D^{-1}\|^2 \le MN^2.$$

Inductively it follows that $\|D^{-1}\|^L \le cM^L N^{L+1}$. ∎

Acknowledgments. This paper was finished under the guidance of Professor Helmut Rüssmann when the author was visiting Mainz University in Germany with a scholarship from the People's Republic of China. The author thanks him for many heuristic suggestions for this paper and is also very grateful to him for much help

during the author's stay in Mainz. The author would also like to thank Professor Jiangong You and Mr. Joachim Albrecht for much help. The author would like to thank the Department of Mathematics and Computer Science of Mainz University for their hospitality. The author is also very grateful to the referees for their suggestions for this revised version.

REFERENCES

[1] J. Bourgain, *On Melnikov's persistency problem*, Math. Res. Lett., 4 (1997), pp. 445–458.
[2] J. Bourgain, *Construction of quasi-periodic solutions for Hamiltonian perturbations of linear equations and applications to nonlinear PDE*, Internat. Math. Res. Notices, (1994), pp. 475–497.
[3] C.-Q. Cheng, *Birkhoff-Kolmogorov-Arnold-Moser tori in convex Hamiltonian systems*, Comm. Math. Phys., 177 (1996), pp. 529–559.
[4] L. H. Eliasson, *Perturbations of stable invariant tori for Hamiltonian systems*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 15 (1988), pp. 115–147.
[5] S. M. Graff, *On the continuation of hyperbolic invariant tori for Hamiltonian systems*, J. Differential Equations, 15 (1974), pp. 1–69.
[6] S. B. Kuksin, *Nearly Integrable Infinite Dimensional Hamiltonian Systems*, Lecture Notes in Math. 1556, Springer-Verlag, Berlin, 1993.
[7] B. Liu, *On lower dimensional invariant tori in reversible systems*, J. Differential Equations, 176 (2001), pp. 158–194.
[8] V. K. Melnikov, *On some cases of conservation of conditionally periodic motions under a small change of the Hamiltonian function*, Soviet Math. Dokl., 6 (1965), pp. 1592–1596.
[9] V. K. Melnikov, *A family of conditionally periodic solutions of a Hamiltonian system*, Soviet Math. Dokl., 9 (1968), pp. 882–886.
[10] J. Pöschel, *On elliptic lower dimensional tori in Hamiltonian systems*, Math. Z., 202 (1989), pp. 559–608.
[11] H. Rüssmann, *On Twist Hamiltonian*, in Colloque International: Mécanique Céleste et Systémes Hamiltoniens, Marseille, 1990.
[12] H. Rüssmann, *Invariant tori in non-degenerate nearly integrable Hamiltonian systems*, Regul. Chaotic Dyn., 6 (2001), pp. 119–204.
[13] H. Rüssmann, *Stability of elliptic fixed points of analytic area-preserving mappings under the Bruno condition*, Ergodic. Theory Dynam. Systems, 10 (2001), pp. 1–22.
[14] M. B. Sevryuk, *KAM-stable Hamiltonians*, J. Dynam. Control Systems, 1 (1995), pp. 351–366.
[15] M. B. Sevryuk, *Reversible Systems*, Lecture Notes in Math. 1211, Springer-Verlag, New York, Berlin, 1986.
[16] M. B. Sevryuk, *Invariant m-dimensional tori of reversible systems with phase space of dimension greater than* $2m$, J. Soviet. Math., 51 (1990), pp. 2374–2386.
[17] M. B. Sevryuk, *The iteration-approximation decoupling in the reversible KAM theorey*, Chaos, 5 (1995), pp. 552–565.
[18] M. B. Sevryuk, *Quasi-Periodic Motions in Families of Dynamical*, Lecture Notes in Math. 1645, Springer-Verlag, New York, Berlin, 1996.
[19] J. Xu, J. You, and Q. Qiu, *Invariant tori of nearly integrable Hamiltonian systems with degeneracy*, Math. Z., 226 (1997), pp. 375–386.
[20] J. You, *Perturbations of lower dimensional tori for Hamiltonian systems*, J. Differential Equations, 152 (1999), pp. 1–29.
[21] J. Xu and J. You, *Persistence of lower dimensional tori under the first Melnikov's Nonresonance condition*, J. Math. Pures Appl. (9), 80 (2001), pp. 1045–1067.
[22] J. Xu and J. You, *A symplectic map and its application to persistence of lower dimensional tori*, Sci. China Ser. A., 45 (2002), pp. 598–603.
[23] H. Whitney, *Analytical extensions of differentiable functions defined in closed sets*, Trans. Amer. Math. Soc., 36 (1934), pp. 63–89.

SIAM J. MATH. ANAL.
Vol. 36, No. 1, pp. 256–271

A WELL-POSED FREE BOUNDARY VALUE PROBLEM FOR A HYPERBOLIC EQUATION WITH DIRICHLET BOUNDARY CONDITIONS*

JOHN V. MATTHEWS† AND DAVID G. SCHAEFFER‡

Abstract. We construct solutions of a free boundary value problem for a hyperbolic equation with Dirichlet boundary data. This problem arises from a model of deformation of granular media.

Key words. hyperbolic, PDE, free boundary, Dirichlet

AMS subject classifications. 35L20, 35R35, 76T25

DOI. 10.1137/S0036141002408708

1. Introduction.

In this work we will study the partial differential equation (PDE) for a scalar function $v(x, y)$,

$$\operatorname{div}\left(R_\alpha \frac{\nabla v}{|\nabla v|}\right) = 0\,, \tag{1.1}$$

where R_α is a rotation counterclockwise by an angle $0 < \alpha < \pi/4$,

$$R_\alpha = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix}.$$

Throughout the paper we consider only functions v such that $\nabla v \neq 0$, which avoids the singularity of (1.1). As will be discussed more fully in the appendix, this equation represents the steady state of a kind of granular flow problem studied in [12], [11].

The domain Ω on which we consider this equation is approximately rectangular, with one free boundary along the topmost edge (see Figure 1.1(a)). Even though, as we will show below, (1.1) is hyperbolic, we will impose Dirichlet-type boundary conditions: specifically,

$$\begin{aligned} v(0,y) &= \phi_0(y), & 0 < y < L_0, \\ v(1,y) &= \phi_1(y), & 0 < y < L_1, \\ v(x,0) &= 0, & 0 < x < 1, \\ v(x,s(x)) &= V, & 0 < x < 1, \end{aligned} \tag{1.2}$$

where the function $s = s(x)$ defined on $0 < x < 1$ describes the free boundary and V is a constant. Naturally, s must satisfy

$$s(0) = L_0, \quad s(1) = L_1.$$

*Received by the editors May 30, 2002; accepted for publication (in revised form) September 12, 2003; published electronically June 22, 2004. This research was supported under NSF grants DMS-9803305 and DMS-9983320.
http://www.siam.org/journals/sima/36-1/40870.html
†Department of Mathematics, Duke University, Box 90320, Durham, NC 27708-0320 (jvmatthe@math.duke.edu).
‡Department of Mathematics and Center for Nonlinear and Complex Systems, Duke University, Box 90320, Durham, NC 27708-0320 (dgs@math.duke.edu).

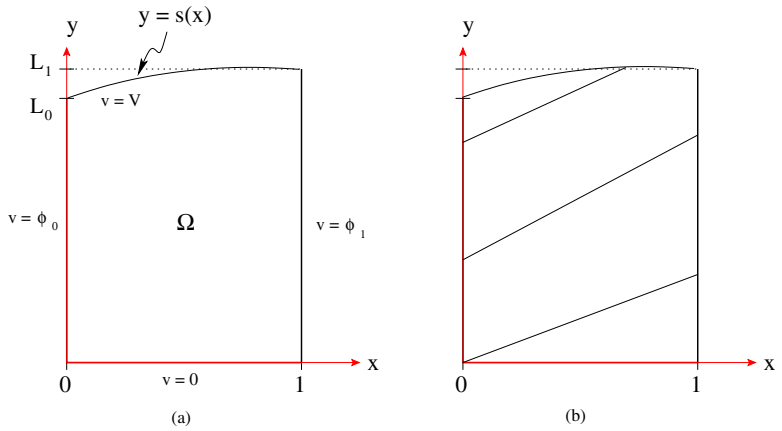

FIG. 1.1. *Two portraits of the domain* Ω, (a) *showing the boundary data and the free boundary and* (b) *illustrating straight-line characteristics along which* v *may have a discontinuity in its third derivative.*

We also impose the obvious compatibility conditions at the corners of the domain,

$$\begin{aligned} \phi_0(0) &= 0, \quad \phi_1(0) = 0, \\ \phi_0(L_0) &= V, \quad \phi_1(L_1) = V. \end{aligned}$$

Finally, to ensure that a solution v is differentiable on Ω, we will need to impose the nonlocal compatibility conditions

$$\phi_0'(0) = \phi_1'(0) \qquad \text{and} \tag{1.3a}$$

$$\phi_0''(0) = \phi_1''(0). \tag{1.3b}$$

Nonlocal compatibility conditions for the boundary data arise in this problem because (1.1) is hyperbolic. Even when (1.3a), (1.3b) are satisfied, a possible discontinuity in third-order derivatives of v propagates along a sequence of straight-line characteristics starting at the lower left corner of Ω (see Figure 1.1(b)). (Note that, even though Dirichlet boundary conditions are imposed, singularity information flows from the fixed boundary to the free boundary.) By imposing additional compatibility conditions like (1.3a), (1.3b), one could increase the order of the derivative of v that may suffer a discontinuity. If one or both conditions (1.3a), (1.3b) were omitted, one could study weak solutions of (1.1), (1.2). We do not pursue this idea here.

In a rectangular domain (as opposed to the quasi-rectangular domain Ω) with boundary data given by $\phi_0 = \phi_1 = y$, the equation (1.1) admits the simple solution $v(x, y) = y$. The problem we consider may be regarded as perturbing the data on the left and right, necessitating the incorporation of the free surface along the top boundary. For the perturbed problem, we will show the following.

THEOREM 1.1. *There exists an* $\epsilon > 0$ *such that if* $\phi_0, \phi_1 \in C^2$ *satisfy* (1.3a), (1.3b), *and* $\|\phi_0 - y\|_2 < \epsilon$, $\|\phi_1 - y\|_2 < \epsilon$, *then there exists a unique* C^2 *solution* v *in* Ω *to* (1.1) *and* (1.2).

Here and throughout, $\|g\|_k$ denotes the C^k-norm of the function g:

$$\|g\|_k = \max_{x \in D(g)} \left\{ |g(x)|, |g'(x)|, \ldots, |g^{(k)}(x)| \right\}.$$

If $0 < \epsilon < 1$, then both of the ϕ_i are monotone. Below we shall in fact assume that

$$\epsilon < \frac{1}{2}. \tag{1.4}$$

Thus, for such ϵ, monotonicity is an implicit assumption. Incidentally, the hypothesis that $\|\phi_i - y\|_2 < \epsilon$ implies that $|L_1 - L_0| = O(\epsilon)$; however, we shall not make explicit use of this fact. The ϵ that we derive below depends on L_0, L_1; indeed, it decreases geometrically as $L_0, L_1 \to \infty$. We are not sure whether this limitation could be avoided through a different argument.

As we shall see below, the free boundary is a characteristic curve. The well-posedness of (1.1), (1.2) depends crucially on this fact: because of it, (1.1), (1.2) resemble a Goursat problem [5].

In section 2 we analyze the characteristics of (1.1) and without loss of generality simplify the boundary data. In section 3, the original BVP for v is reduced to solving a functional equation along the left boundary of the domain; also, limitations on the regularity of the solution are explained. The existence of solutions to the functional equation is proven in section 4, thereby establishing the existence of solutions to the original BVP. Uniqueness is obtained in section 5. Finally, in the appendix we discuss the physical interpretation of (1.1) and describe connections with similar equations from mathematical modeling of granular flow.

2. Preliminaries.

2.1. Analysis by characteristics. We begin our analysis of (1.1) by converting it into an equivalent first-order system. We define

$$\tau = R_\alpha \frac{\nabla v}{|\nabla v|}$$

such that (1.1) becomes

$$\begin{aligned} \operatorname{div} \tau &= 0, \\ |\tau| &= 1, \\ R_\alpha^{-1}\tau \times \nabla v &= 0, \end{aligned} \tag{2.1}$$

where $\times$ denotes the cross product in two dimensions, interpreted as a scalar. This equation implies that $R_\alpha^{-1}\tau$ and ∇v are parallel vectors. Note that the middle equation involves no derivatives; thus (2.1) is a differential-algebraic system. However, (2.1) is easily reduced to a purely differential system. Since τ is a unit vector, we may represent it as

$$\tau = \begin{bmatrix} -\sin(\theta + \alpha) \\ \cos(\theta + \alpha) \end{bmatrix},$$

for some angle $\theta = \theta(x, y)$. This representation of τ, while unusual, simplifies the analysis of characteristics below. Then the three equations in (2.1) are equivalent to the 2×2 quasi-linear system of differential equations,

$$\partial_x \begin{bmatrix} \theta \\ v \end{bmatrix} + \begin{bmatrix} \tan(\theta + \alpha) & 0 \\ 0 & \tan\theta \end{bmatrix} \partial_y \begin{bmatrix} \theta \\ v \end{bmatrix} = 0, \tag{2.2}$$

which is clearly hyperbolic, but not in conservation form.

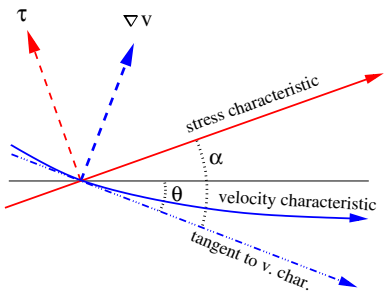

FIG. 2.1. *The orientation of τ, ∇v, and the stress and velocity characteristics. Velocity characteristics are inclined at an angle θ relative to the x-axis, and stress characteristics are inclined at an angle $\theta + \alpha$. (The angle θ in the figure is negative.)*

Observe that the first equation in (2.2) completely decouples from the second. The characteristics of this equation, along which θ is constant, are straight lines of slope

$$\frac{dy}{dx} = \tan(\theta + \alpha). \tag{2.3}$$

Although this equation is decoupled from the other, we cannot solve it separately because no explicit data for θ is given on the boundary of the domain.[1]

The equation for v is not independent of θ, but if we regard θ as known, then the v equation is linear. Characteristics for this equation, along which v is constant, are curves $y = y(x)$ that satisfy the differential equation

$$\frac{dy}{dx} = \tan\theta. \tag{2.4}$$

Boundary data for v—Dirichlet data—is given by (1.2), but we cannot solve the v equation without knowing θ. Based on the interpretation of (2.1) discussed in the appendix, we shall refer to the θ-characteristics and v-characteristics, described by (2.3) and (2.4), as *stress* and *velocity* characteristics, respectively.

Comparing (2.3) and (2.4), we see that the velocity characteristics intersect the stress characteristics at an angle α. To be more precise, the unit tangent along a stress characteristic at a point (x, y) equals the unit tangent at the same point along the velocity characteristic, rotated by an angle α counterclockwise (see Figure 2.1).

Recall our assumption that both ϕ_0, ϕ_1 are strictly monotonic and onto a common range. Therefore, given a point y_r in $[0, L_1]$ along the right boundary of the domain, $\{x = 1\}$, there exists a unique point y_ℓ in $[0, L_0]$ along the left boundary such that

$$\phi_0(y_\ell) = \phi_1(y_r). \tag{2.5}$$

(See Figure 2.2.) Velocity characteristics are level curves for the velocity, and we shall see that y_ℓ and y_r are connected by a velocity characteristic. However, this association of boundary points can be made without knowing the function v on the interior of Ω.

[1] While it is not obvious from the analysis thus far, data for θ actually is determined along the bottom edge of Ω, $\{(x, y) : 0 < x < 1, y = 0\}$. Indeed, it will be shown in the next section that $\theta \equiv 0$ there.

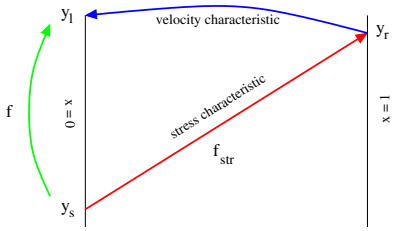

FIG. 2.2. *Connecting boundaries with characteristics: θ is constant along stress characteristics and v is constant along velocity characteristics.*

2.2. Simplification of the boundary data. Suppose that Φ is a C^2 function on the real line such that $\Phi' > 0$. If v satisfies the PDE (1.1) and boundary conditions (1.2) then the function $\Phi(v)$ also satisfies the same PDE,

$$\operatorname{div}\left(R_\alpha \frac{\nabla\Phi(v)}{|\nabla\Phi(v)|}\right) = \operatorname{div}\left(R_\alpha \frac{\Phi'(v)\nabla v}{|\Phi'(v)\nabla v|}\right) = \operatorname{div}\left(R_\alpha \frac{\nabla v}{|\nabla v|}\right) = 0,$$

with the modified boundary data

$$\begin{aligned}\tilde{\phi}_0 &= \Phi \circ \phi_0,\\ \tilde{\phi}_1 &= \Phi \circ \phi_1.\end{aligned}$$

In the following, we use $\Phi = \phi_1^{-1}$ to simplify the boundary data and the subsequent analysis. Because of this transformation, it suffices to solve (1.1) with the boundary data on the sides

$$\begin{aligned}\tilde{\phi}_0(y_\ell) &= \phi_1^{-1} \circ \phi_0(y_\ell), \qquad 0 < y_\ell < L_0,\\ \tilde{\phi}_1(y_r) &= y_r, \qquad 0 < y_r < L_1,\end{aligned} \tag{2.6}$$

provided that $\|\tilde{\phi}_0 - y\|_2$ is sufficiently small. As regards estimates, note that given (1.4) there is a constant C such that

$$\|\tilde{\phi}_0 - y\|_2 \le C \max\left\{\|\phi_0 - y\|_2, \|\phi_1 - y\|_2\right\}.$$

Henceforth, we drop the tilde notation but assume that the boundary data is given as in (2.6). On the left boundary, let us introduce the notation

$$\phi_0(y_\ell) = y_\ell + \beta(y_\ell), \quad 0 < y_\ell < L_0. \tag{2.7}$$

Note that after the reduction in (2.6), equations (1.3a), (1.3b) become

$$\phi_0'(0) = 1, \text{ or equivalently, } \beta'(0) = 0 \text{ and} \tag{2.8a}$$

$$\phi_0''(0) = 0, \text{ or equivalently, } \beta''(0) = 0. \tag{2.8b}$$

For convenient reference below, we list the boundary data on the top and bottom of Ω:

$$\begin{aligned}v(x,0) &= 0, \qquad 0 < x < 1,\\ v(x,s(x)) &= L_1, \qquad 0 < x < 1,\end{aligned} \tag{2.9}$$

where the transformation ϕ_1^{-1} in (2.6) has modified the data on the top of Ω.

3. Reduction of the BVP to a functional equation on the boundary. For a solution v of (1.1), $\theta(x,y)$ represents the angle of inclination of the velocity characteristic at the point (x,y) (see Figure 2.1). The θ equation in (2.2) can be rewritten as

$$\partial_x\theta + \partial_y \ln\left(\frac{1}{\cos(\theta+\alpha)}\right) = 0, \tag{3.1}$$

a scalar hyperbolic equation in conservation form. In the following we will consider x as a time-like variable, as its placement in this equation suggests.

We shall derive an initial condition for (3.1) along $\{(x,y) : x = 0\}$. To do so, however, we will need the following result about certain special solutions of (1.1) and their properties.[2]

LEMMA 3.1. *If $\theta(x,y)$ is a C^1 solution of* (3.1) *such that $|\theta| \le \alpha/2$, then there exists a C^2 function $T(x,y)$ such that*

$$\nabla T = e^{(\cot\alpha)\theta}\begin{bmatrix} -\sin\theta \\ \cos\theta \end{bmatrix}, \tag{3.2}$$

and T is a solution of (1.1). *If θ is derived from a solution v of* (1.1), *then T is a function of v, say $T = \Phi(v)$, and Φ is invertible; in particular, T is constant along velocity characteristics.*

Proof. Given a solution $\theta(x,y)$ of (3.1), define a vector field

$$H = \begin{bmatrix} H_1 \\ H_2 \end{bmatrix} = e^{(\cot\alpha)\theta}\begin{bmatrix} -\sin\theta \\ \cos\theta \end{bmatrix}. \tag{3.3}$$

By (3.1), it is easy to verify that

$$\partial_y H_1 = \partial_x H_2.$$

Since the domain is simply connected, H is conservative. That is, there exists a C^2 function T such that $\nabla T = H$.

Recall that τ is given by

$$\tau = \begin{bmatrix} -\sin(\theta+\alpha) \\ \cos(\theta+\alpha) \end{bmatrix}.$$

Then by (3.1), $\operatorname{div}\tau = 0$, and of course $|\tau| = 1$. Also,

$$R_\alpha^{-1}\tau \times \nabla T = \cos\theta\,\partial_x T + \sin\theta\,\partial_y T = 0.$$

Thus the pair τ, T satisfy (2.1), which means that T satisfies (1.1).

Suppose $\theta(x,y)$ is the angle of inclination of the velocity characteristics of a solution v of (1.1). The directional derivative of T along a velocity characteristic equals $\cos\theta\,\partial_x T + \sin\theta\,\partial_y T$, and by (3.3) this vanishes. Thus, T is constant along velocity characteristics, and it follows that $T = \Phi(v)$. Moreover, since neither ∇T nor ∇v vanishes, Φ is invertible. □

[2]The idea for this step in the proof is due to Robert Bryant. Using Darboux's method [1], [2], he obtained this result as an affirmative answer to the question, Are there coordinates in which the general solution of (1.1) be can represented explicitly in terms of two arbitrary functions (as in d'Alembert's solution of the wave equation)?

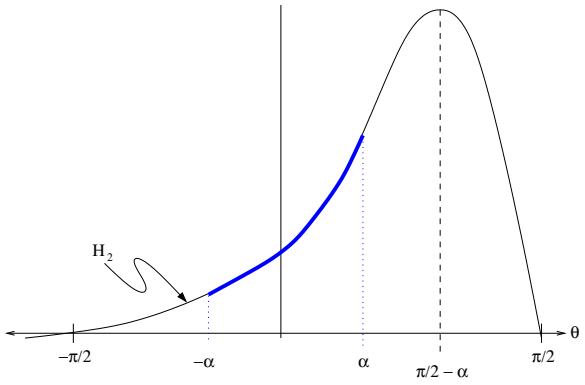

FIG. 3.1. *Graph of the function H_2. This function is invertible on the interval $[-\alpha, \alpha]$, for $0 < \alpha < \pi/4$.*

Only the second component of $\nabla T = H$ will be used below. Slightly changing the notation from (3.3), we shall let H_2 denote the function of one real variable,

$$H_2(\theta) = e^{(\cot \alpha)\theta} \cos \theta. \tag{3.4}$$

The profile of a typical H_2 is shown in Figure 3.1.

Remark 1. We observe that H_2 is strictly positive on $[-\alpha, \alpha]$, a fact which follows from $\alpha < \pi/4$ and the definition of H_2.

Let $v(x, y)$ be a solution of (1.1) with boundary data on the sides of Ω as in (2.6), and let $\theta(x, y)$ be the angle of inclination of the corresponding velocity characteristics. As illustrated in Figure 2.2, given a starting point y_s on the y-axis, define

$$y_r = y_s + \tan\left[\theta(0, y_s) + \alpha\right] \tag{3.5}$$

such that $(1, y_r)$ is connected to $(0, y_s)$ by a stress characteristic, and define

$$y_\ell = \phi_0^{-1}(y_r) \tag{3.6}$$

such that $(0, y_\ell)$ is connected to $(1, y_r)$ by a velocity characteristic. We claim that

$$H_2 \circ \theta(0, y_\ell) = \phi_0'(y_\ell)\, H_2 \circ \theta(0, y_s). \tag{3.7}$$

With this relation we will be able to determine boundary conditions for θ from the given boundary conditions for v.

Proof of (3.7). By the above lemma, the function T is constant along velocity level lines. Therefore,

$$T\left(0, \phi_0^{-1}(y_r)\right) = T\left(1, y_r\right).$$

Temporarily treating y_r as an independent variable, we differentiate T with respect to its second argument on both sides of the above equation to obtain

$$\partial_2 T\left(0, \phi_0^{-1}(y_r)\right)\left(\phi_0^{-1}\right)'(y_r) = \partial_2 T\left(1, y_r\right). \tag{3.8}$$

Recalling (3.2), (3.4), and (3.6), we can rewrite (3.8) as

$$H_2 \circ \theta\left(0, y_\ell\right) \phi_0^{-1}(y_r) = H_2 \circ \theta\left(1, y_r\right). \tag{3.9}$$

As noted above, stress characteristics are straight lines along which θ is constant. Since $(0, y_s)$ and $(1, y_r)$ lie on the same stress characteristic,

$$\theta(0, y_s) = \theta(1, y_r). \tag{3.10}$$

Finally, from (2.5) and (2.6) one can derive

$$\left(\phi_0^{-1}\right)'(y_r) = \frac{1}{\phi_0'\left(\phi_0^{-1}(y_r)\right)} = \frac{1}{\phi_0'(y_\ell)}. \tag{3.11}$$

Then equation (3.7) follows immediately from the substitution of (3.10) and (3.11) into (3.9). □

Below it will be useful to rewrite (3.7) in a more systematic notation. First, we abbreviate $\theta(0, y)$ to $\theta_0(y)$. Let f be the mapping

$$y_s \to y_r \to y_\ell$$

as given by (3.5), (3.6) (see Figure 2.2). In this notation, (3.7) becomes

$$H_2 \circ \theta_0(f(y)) = \phi_0'(f(y))\, H_2 \circ \theta_0(y). \tag{3.12}$$

The above derivation, in which θ is determined from a solution v of (1.1), (2.6), shows that (3.12) holds for all y such that

$$0 \le y < f(y) \le L_0.$$

We now show that if v also satisfies (2.9), then in fact (3.12) holds for a slightly larger range of y.

If follows from (2.9) that the bottom edge of Ω is a velocity characteristic and, since it has zero slope, we conclude that

$$\theta(x, 0) = 0, \quad 0 < x < 1. \tag{3.13}$$

If we regard (3.13) as initial data, then (3.1) has the unique solution $\theta \equiv 0$ in the parallelogram B indicated in Figure 3.2, the parallelogram with vertices $(0, 0)$, $(1, \tan\alpha)$, $(1, 0)$, $(-\tan\alpha, 0)$. This construction shows that if θ is the slope of the velocity characteristics of a solution v of (1.1), (2.6), (2.9), then

(i) $\theta \equiv 0$ on the triangle $B \cap \Omega$, and

(ii) θ has a natural extension from its original domain Ω to $B \cup \Omega$, where $\theta \equiv 0$ on B.

Let v be a solution of (1.1), (2.6), (2.9), and let θ be the angle of inclination of the corresponding velocity characteristics, extended as in (ii). We claim that (3.12) holds for all y such that

$$-\tan\alpha \le y < f(y) \le L_0. \tag{3.14}$$

Indeed, this claim follows from the above proof of (3.7).

With these ideas, we can now explain why θ and v may fail to have higher-order derivatives. Let us consider the continuity of θ_0 at $y = 0$. Of course, for the limit from below, $\lim_{y\to 0^-} \theta_0(y) = 0$. For the limit $y \to 0^+$, we let $y \to -\tan\alpha$ from above in (3.12), observing that $f(-\tan\alpha) = 0$; thus

$$H_2\left(\lim_{y\to 0^+} \theta_0(y)\right) = \phi_0'(0)\, H_2(0).$$

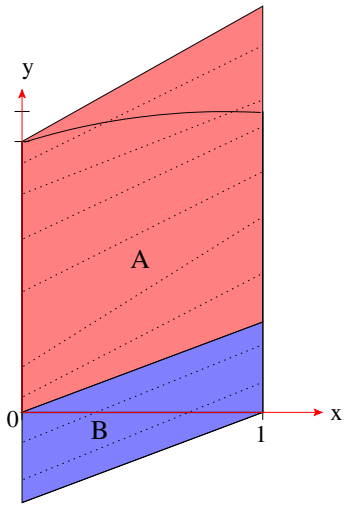

FIG. 3.2. *How θ is determined in Ω from its boundary data θ_0. The region A along with the function θ in A are determined by θ_0 on $\{y \geq 0\}$. Note that the stress characteristics pass out of the top of the domain, through the free boundary. Likewise θ in region B is determined by $\theta \equiv 0$ along $y = 0$.*

Since $H_2(0) = 1$, we conclude that $\lim_{y\to 0^+} \theta_0(y) = 0$ if and only if $\phi_0'(0) = 1$, as required in (2.8a). Thus θ_0 is continuous at $y = 0$ if and only if (2.8a) holds. Similarly, θ' is continuous at $y = 0$ if and only if (2.8b) holds. However, without further restrictions on ϕ_0, second- or higher-order derivatives of θ_0 may jump at $y = 0$.

According to (3.1), a jump in θ_0'' propagates into the interior of Ω along a stress characteristic at the origin. Also, according to (3.12), the jump in θ_0'' at the origin creates other jumps at points $f(0), f \circ f(0), \ldots$ higher on the y-axis, which in turn also propagate into Ω. As in (3.3), jumps in the second-order derivatives of $\theta(x, y)$ lead to jumps in the third derivatives of v. In this way, v may exhibit singularities along a set of straight-line characteristics, as illustrated in Figure 1.1(b).

We derived (3.12) for θ_0 by assuming that a solution v existed in the domain Ω. In the following theorem, we assume that (3.12) has a solution θ_0 along the boundary and construct the related solution v in Ω.

THEOREM 3.2. *Suppose θ_0 is a C^1 function on $[-\tan\alpha, L_0]$ such that*

(i) *$\theta_0 \equiv 0$ on $[-\tan\alpha, 0]$,*

(ii) *θ_0 satisfies* (3.12) *for all y such that* (3.14) *holds, and*

(iii) *$\|\theta_0\|_0 < \frac{\alpha}{2}$ and $\|\theta_0'\|_0 < \frac{1}{2}\cos^2(\frac{3}{2}\alpha)$.*

Then there exists a C^2 solution v to (1.1) *that satisfies the boundary conditions* (2.6), (2.9).

Proof. Consider the IVP for (3.1) with the initial condition θ_0 given in the hypothesis of the theorem: in symbols,

$$\theta(0, y) = \theta_0(y), \quad -\tan\alpha \leq y \leq L_0. \tag{3.15}$$

The theory of scalar conservation laws [4] indicates that this problem has a C^1 solution in a portion of some narrow strip

$$\{(x, y) : 0 < x < \zeta\}$$

that is bounded above and below by characteristics. This solution extends to larger values of x provided the characteristics of (3.1) do not cross. If the initial condition satisfies

$$1 + \sec^2(\theta_0(y) + \alpha)\,\theta_0'(y) > 0 \tag{3.16}$$

for all $y \in [-\tan\alpha, L_0]$, then the characteristics do not cross in the larger strip $\{(x,y) : 0 < x < 1\}$. For convenience below, we have assumed a stronger estimate on θ_0 that bounds the left-hand side of (3.16) from below by $1/2$. Therefore, this initial value problem has a C^1 solution θ in a quadrilateral domain $A \cup B$ as sketched in Figure 3.2. Further, by the theory of hyperbolic conservation laws, that solution on $A \cup B$ satisfies $|\theta| < \alpha/2$ since the boundary data satisfies (iii).

Having constructed $\theta(x,y)$, we now define $T(x,y)$ as in Lemma 3.1. We construct Ω such that its top boundary is a level curve of T:

$$\Omega = \{(x,y) : 0 < x < 1,\ 0 < y,\ T(x,y) < T(0,L_0)\}.$$

Since the stress characteristics arising in the solution of (3.1) are inclined at the positive angle α with respect to level curves of T, we have $\Omega \subset A \cup B$.

To construct v from T, we define

$$\Phi(y) = T(1,y), \quad y > 0. \tag{3.17}$$

We may compute a first derivative of Φ using (3.8), (3.9),

$$\Phi'(y) = H_2 \circ \theta(1,y) \neq 0, \tag{3.18}$$

since $|\theta| < \alpha/2$ on Ω and H_2 is strictly positive for such θ, as we noted in Remark 1; therefore, Φ is invertible. Let us define

$$v = \Phi^{-1} \circ T. \tag{3.19}$$

By the discussion in section 2, we know that v satisfies (1.1). By construction, $v(1,y) = y$ along the right boundary of Ω as required in (2.6). Examining the derivation of (3.7), we see that $v = \phi_0$ along the left boundary. Finally, also by the construction, the remaining boundary conditions (2.9) are satisfied. □

Remark 2. Let f_{str} be the mapping $y_s \to y_r$ as given by (3.5). The derivative of this map is given by the left-hand side of (3.16). Therefore, condition (iii) of the theorem guarantees that

$$f_{str}'(y) \geq \frac{1}{2}; \tag{3.20}$$

in particular f_{str} is one-to-one.

4. Solution of the functional equation. In the next three lemmas, we suppose that θ_0 is a C^1 function defined for y in a subinterval of the y-axis

$$I = [-\tan\alpha, y^*], \tag{4.1}$$

where

$$f(y^*) \leq L_0. \tag{4.2}$$

We assume that $\theta_0 \equiv 0$ for $y \leq 0$ and that θ_0 satisfies the functional equation (3.12) for all y such that both y and $f(y)$ are in I. We propose to use (3.12) to extend θ_0 to the larger interval

$$I_+ = [-\tan\alpha, f(y^*)]\,. \tag{4.3}$$

Formally, we may apply H_2^{-1} to (3.12) and obtain

$$\theta_0\left(f(y)\right) = H_2^{-1}\left[\phi_0'\left(f(y)\right)\, H_2 \circ \theta_0(y)\right], \tag{4.4}$$

a formula which relates θ_0 at $f(y)$ to the boundary data ϕ_0 at $f(y)$ and θ_0 at y. However, care must be taken to ensure that (4.4) represents a meaningful definition. Below, by iterating this basic step, we shall obtain a function θ_0 defined on the entire interval $[-\tan\alpha, L_0]$ as needed in Theorem 3.2.

In these lemmas we explicitly indicate the precise domain over which certain norms are to be evaluated: e.g.,

$$\|\theta_0 : I\|_0 = \max_{x\in I} |\theta_0(x)|.$$

The norms $\|\beta\|_k$ are to be evaluated over $[0, L_0]$; although $I_+ \cap [0, L_0]$ would suffice, enlarging to $[0, L_0]$ simplifies the notation.

LEMMA 4.1. *There exists $\delta > 0$ such that if*

(i) $\|\beta\|_1 < \delta$ *and*

(ii) $\|\theta_0 : I\|_0 < \frac{\alpha}{2}$ *and* $\|\theta_0' : I\|_0 < \frac{1}{2}\cos^2(\frac{3}{2}\alpha)$,

then (4.4) *supplies a well-defined extension of θ_0 from I to I_+ that satisfies the functional equation* (3.12) *for all y such that y and $f(y)$ belong to I_+.*

Proof. For (4.4) to be a valid definition, we need the argument of H_2^{-1} on the right-hand side of (4.4) to belong to the domain of H_2^{-1} and we need f to be one-to-one. Regarding the latter point, we recorded in Remark 2 that if θ_0 satisfies condition (ii), then the map f_{str}, and hence $f = \phi_0^{-1} \circ f_{str}$, is one-to-one. Regarding the former point, observe that by (i), $H_2 \circ \theta_0(y)$ belongs to the set $H_2([-\alpha/2, \alpha/2])$. However, since $\alpha < \pi/4$, the function H_2 is monotone on the larger interval $[-\alpha, \alpha]$. Thus H_2^{-1} is defined and smooth on $H_2([-\alpha, \alpha])$, and $H_2([-\alpha, \alpha])$ contains $H_2([-\alpha/2, \alpha/2])$ in its interior. Provided δ is sufficiently small, if (i) and (ii) hold, then for any $y \in I$

$$\phi_0'(f(y))\, H_2 \circ \theta_0(y) = [1 + \beta'(f(y))]\; H_2 \circ \theta_0(y) \in H_2([-\alpha, \alpha]). \tag{4.5}$$

Then (4.4) gives an unambiguous extension of θ_0 on I_+, and by construction (3.12) is satisfied on the extended interval. □

For subsequent arguments we need to be more quantitative about the monotonicity of H_2 in this proof: let

$$m = \min_{\theta\in[-\alpha,\alpha]} H_2'(\theta) > 0, \tag{4.6}$$

the positivity following from the fact that $[-\alpha, \alpha]$ is compact.

LEMMA 4.2. *If β and θ_0 satisfy conditions* (i) *and* (ii) *of Lemma* 4.1*, then there are positive constants C_1 and C_2 such that*

$$\|\theta_0 : I_+\|_0 \leq \|\theta_0 : I\|_0 + C_1\, \|\beta\|_1 \tag{4.7}$$

and

$$\|\theta_0' : I_+\|_0 \leq C_2\, \|\theta_0' : I\|_0 + C_1\, \|\beta\|_2\,. \tag{4.8}$$

Proof. In proving this result, it is convenient to introduce the inverse function $g = f^{-1}$ and rewrite (4.4) as

$$\theta_0(y) = H_2^{-1}\left[\phi_0'(y)\, H_2 \circ \theta_0\left(g(y)\right)\right]. \tag{4.9}$$

Applying H_2 to both sides of (4.9), subtracting $H_2 \circ \theta_0(g(y))$ from both sides, and recalling the definition (2.7) of β, we see that

$$H_2 \circ \theta_0(y) - H_2 \circ \theta_0(g(y)) = \beta'(y)\, H_2 \circ \theta_0(g(y)). \tag{4.10}$$

Regarding the left-hand side of (4.10), since H_2 is monotone and satisfies (4.6) on $[-\alpha, \alpha]$, we deduce by the mean-value theorem that

$$m\left|\theta_0(y) - \theta_0(g(y))\right| \le \left|H_2 \circ \theta_0(y) - H_2 \circ \theta_0(g(y))\right|. \tag{4.11}$$

Combining (4.10) and (4.11) with the triangle inequality

$$\left|\theta_0\left(y\right)\right| \le \left|\theta_0(g(y))\right| + \left|\theta_0\left(y\right) - \theta_0(g(y))\right|,$$

we obtain (4.7) with

$$C_1 = m^{-1} \max_{\theta \in [-\alpha, \alpha]} |H_2(\theta)|. \tag{4.12}$$

Turning to (4.8), by differentiating (4.9), we obtain

$$\theta_0'(y) = \frac{\left[\phi_0'(y)\right]^2 H_2' \circ \theta_0(g(y))}{H_2' \circ \theta_0(y) f_{str}'(g(y))}\theta_0'(g(y)) + \frac{H_2 \circ \theta_0(g(y))}{H_2' \circ \theta_0(y)}\phi_0''(y). \tag{4.13}$$

The second term in (4.13) is no greater than $C_1\|\beta\|_2$, where C_1 is given by (4.12). The first term is no greater than $C_2\|\theta_0' : I\|_0$, where to obtain C_2 we invoke condition (i) of Lemma 4.1 to estimate ϕ_0', (4.6) to estimate H_2', and (3.20) to estimate f_{str}'. This proves (4.8). □

In the next lemma, we show that a single application of Lemma 4.1 extends the domain of θ_0 by a distance of at least h, where

$$h = \frac{1}{2}\tan\left(\frac{\alpha}{2}\right). \tag{4.14}$$

LEMMA 4.3. *If β and θ_0 satisfy conditions* (i) *and* (ii) *of Lemma* 4.1, *where in condition* (ii) *we have $\delta \le h$, then*

$$f(y^*) > y^* + h.$$

Proof. Recall that $f(y^*) = \phi_0^{-1} \circ f_{str}(y^*)$ and that

$$f_{str}(y^*) = y^* + \tan\left[\theta_0(y^*) + \alpha\right].$$

By condition (ii), $\theta_0(y^*) \ge -\alpha/2$, so

$$f_{str}(y^*) \ge y^* + \tan\left(\frac{\alpha}{2}\right) = y^* + 2h. \tag{4.15}$$

Regarding ϕ_0^{-1}, the other factor in f, first observe that for any y

$$|\phi_0(y) - y| = |\beta(y)| < \delta \le h.$$

Substituting $z = \phi_0(y)$ we deduce

$$\left|\phi_0^{-1}(z) - z\right| < h,$$

from which it follows that

$$\phi_0^{-1}(z) > z - h.$$

Taking $z = f_{str}(y^*)$ and recalling (4.15), we see that

$$f(y^*) > f_{str}(y^*) - h \geq y^* + h,$$

as claimed. □

Given the boundary data (2.6), (2.9) for (1.1), we want to construct a function θ_0 on $[-\tan\alpha, L_0]$ as in Theorem 3.2. Starting from $\theta_0 \equiv 0$ on

$$I_0 = [-\tan\alpha, 0],$$

we propose to apply Lemma 4.1 iteratively on an increasing sequence of intervals

$$I_1 = [-\tan\alpha, y_1],\ I_2 = [-\tan\alpha, y_2], \ldots$$

where $y_0 = 0$ and for $k = 1, 2, \ldots$

$$y_k = f(y_{k-1}).$$

Provided $\|\beta\|_1 < \delta$, it follows from Lemma 4.1 that extension from I_0 to I_1 is possible. By Lemma 4.2, provided

$$C_1\|\beta\|_1 < \frac{\alpha}{2} \qquad \text{and} \qquad C_1\|\beta\|_2 < \frac{1}{2}\cos^2\left(\frac{3}{2}\alpha\right),$$

extension from I_1 to I_2 is also possible. More generally, provided

$$N \cdot C_1\|\beta\|_1 < \frac{\alpha}{2} \qquad \text{and} \qquad \frac{C_2^N - 1}{C_2 - 1}\, C_1\, \|\beta\|_2 < \frac{1}{2}\cos^2\left(\frac{3}{2}\alpha\right),$$

the estimates of Lemma 4.1 will remain valid for N iterations. Thus for any positive integer N, if $\|\beta\|_2$ is sufficiently small, we can extend θ_0 to I_N, provided $y_k \leq L_0$ for $k = 1, 2, \ldots, N$.

According to Lemma 4.3, after some number N iterations, where $N \leq L_0/h$, we shall arrive at a point where

$$y_N \leq L_0 \qquad \text{but} \qquad f_{str}(y_N) > L_1$$

so that $f(y_N)$ is undefined. In this case, there is some $\tilde{y} < y_N$ such that $f(\tilde{y}) = L_0$, and Lemma 4.1 may be applied one more time to extend θ_0 from $[-\tan\alpha, \tilde{y}]$ to the entire interval $[-\tan\alpha, L_0]$. (This modification of the process also handles the "short domain" case in which $f_{str}(0) > L_1$.) In this way, provided $\|\beta\|_2$ is sufficiently small, we may obtain θ_0 as in Theorem 3.2 and then invoke that theorem to solve (1.1), (2.6), and (2.9).

5. Proof of uniqueness. Briefly, the uniqueness proof proceeds by checking that each step of the existence proof, which is constructive, has a unique outcome. Specifically, let v and $\widehat{v}$ be two solutions of (1.1) on domains Ω and $\widehat{\Omega}$ and satisfying the reduced boundary conditions (2.6), (2.9). We assume that (i) neither ∇v nor $\nabla\widehat{v}$ vanishes and (ii) the boundary-data function $\phi_0(y) = y+\beta(y)$ satisfies $\|\beta\|_2 < \epsilon$ where ϵ is the constant of Theorem 1.1: i.e., ϵ is sufficiently small to guarantee existence through the above construction.

- Define $\theta = \arg(\nabla v) - \pi/2$ to be the angle of inclination of the level curves, or characteristics, of this solution, and define $\widehat{\theta}$ likewise. Both θ and $\widehat{\theta}$ satisfy (3.1), vanish identically on the triangle $B \cap \Omega$ of Figure 3.2, and extend as solutions of (3.1) to be identically zero on all of B. In particular for the extended functions, $\theta_0(y) = \widehat{\theta}_0(y) = 0$ for $-\tan\alpha \le y \le 0$.
- Both θ and $\widehat{\theta}$ satisfy the functional equation (3.12). It follows from the estimates in section 3 that f is one-to-one and that $\theta_0(f(y))$ belongs to the domain of H_2^{-1}, and likewise for $\widehat{f}$ and $\widehat{\theta}_0(\widehat{f}(y))$. Therefore, (3.12) may be rewritten in the form (4.9), which shows that $\theta_0(y) = \widehat{\theta}_0(y)$ for the entire interval $-\tan\alpha \le y \le L_0$.
- By solving (3.1) with the initial condition (3.15), we obtain identical extensions of θ and $\widehat{\theta}$ on the quadrilateral domain $A \cup B$ of Figure 3.2 (the same domain for both functions).
- Thus the characteristic equation (2.4) for the two solutions, v and $\widehat{v}$, is the same. Integrating this equation with the initial condition $y(0) = L_0$ to obtain the upper boundary of Ω we conclude that $\Omega = \widehat{\Omega}$. More generally, all the level curves or characteristics of v and $\widehat{v}$ in Ω coincide. Since v and $\widehat{v}$ are equal on the left boundary of Ω, they are equal throughout Ω.

Appendix: Connection to granular flow. Equation (1.1), or the equivalent first-order system (2.1), arises from a model [12] for steady-state antiplane shearing of a granular medium. The term *antiplane shear* refers to a special class of deformation of a three-dimensional solid in which: (i) all motion is in the z-direction,

$$\overrightarrow{v} = (0, 0, v),$$

where v is the scalar velocity in (2.1); (ii) the stress tensor has the reduced form

$$T = \begin{pmatrix} \sigma & 0 & \tau_1 \\ 0 & \sigma & \tau_2 \\ \tau_1 & \tau_2 & \sigma \end{pmatrix},$$

where σ is a uniform confining pressure and τ is the vector in (2.1); and (iii) the velocity and the stress depend on x and y but are independent of z. (The confining pressure σ is independent of all three coordinates.) The three equations in (2.1) represent the following, respectively: (a) Force balance or Newton's second law of motion with inertia neglected (appropriate for slow flow). (b) Coulomb's law of friction—for motion to occur, the shearing stress must equal a threshold. In general, this threshold depends on the confining pressure σ and on the internal friction of the material, but we have nondimensionalized the equations. (c) The nonassociative constitutive law proposed in [12]. More accurately, the constitutive law should read as follows: there exists a nonnegative function $\lambda(x, y)$ such that

$$\nabla v = \lambda R_\alpha^{-1} \tau.$$

In (2.1) the function λ has been eliminated, but the condition $\lambda \geq 0$—that friction acts dissipatively—needs to be checked a posteriori in order to verify that a solution of (2.1) is physical.[3]

Physically, (2.1) may be viewed as describing the continuum limit of a collection of infinitely long, thin rods. These rods are parallel to the z-axis, their cross sections fill the domain Ω, and they slide over one another along their axes, subject to the constitutive law proposed in [12]. On physical grounds, one would expect to be able to control the velocity of all the rods at the boundary of Ω—in mathematical terms Dirichlet boundary conditions are suggested. Similarly, it would seem impossible to control *both* velocity and stress on any portion of $\partial\Omega$—again in mathematical terms, Cauchy data are excluded.

The model (2.1) was proposed as a technically simpler analogue of the equations of slow, two-dimensional flow of an incompressible Coulomb material [3], [8], [7]. The physical unknowns for such flow consist of a 2-component velocity v and a 2×2 stress tensor T (three scalars, since T is symmetric). These unknowns satisfy

$$
\begin{aligned}
\sum_{j=1}^{2} \partial_j T_{ij} &= \rho g_i, && i = 1, 2, \\
\sum_{i,j=1}^{2} \left(T_{ij} - \sigma\delta_{ij}\right)^2 &= k\sigma, && \\
\lambda \left(T_{ij} - \sigma\delta_{ij}\right) &= -\tfrac{1}{2}\left(\partial_i v_j + \partial_j v_i\right), && i, j = 1, 2
\end{aligned}
\tag{5.1}
$$

where ρ is the (constant) density, g is the acceleration of gravity, k is a constant describing internal friction, σ is the mean stress

$$\sigma = \frac{1}{2}\sum_{i=1}^{2} T_{ii},$$

and $\lambda(x_1, x_2) \geq 0$ is an auxiliary function that arises in the mathematical formulation of plasticity.

The close analogy in form between (5.1) and (2.1) is readily apparent, even though the number of equations is different. However, the analogy is far closer than a superficial appearance. Specifically, as with (2.1), we have the following: (i) The algebraic constraint may be eliminated by an appropriate reparameterization of the stress (the Sokolovskii variables [13]). (ii) The two remaining stress variables satisfy a strictly hyperbolic system that is uncoupled from the velocity. (iii) If λ is eliminated from the third equation in (5.1), the two velocity equations—with T regarded as known—are a linear, strictly hyperbolic system in v.

The present work originated from an attempt to use the three-dimensional analogue of (5.1) to model slow, steady flow in a hopper [6], [9]. Although boundary conditions along the walls of the hopper are natural and easy to formulate, the situation at the top and bottom is very unclear: e.g., how many conditions should be imposed at the top and how many at the bottom? Exactly where should they be imposed? Since the equations are hyperbolic, one might seek a Cauchy problem. The stress and velocity equations decouple, and therefore there are various ways of posing

[3] This positivity is readily checked for the problem in the present paper: the coefficient λ is identically $+1$ for the unperturbed solution $v(x, y) = y$, and the perturbation is too small to change the sign of λ.

different Cauchy problems. However, in [9] all ways of prescribing Cauchy data led to unphysical solutions in which the constraint $\lambda \geq 0$ was violated. In such a solution, friction is *adding* energy to the flow.

This paper identifies a well-posed, Dirichlet-type boundary value problem for the analogous, but technically simpler system (2.1). The appearance of a free boundary, which is crucial for the result, was suggested by hopper flow: there is some evidence that exit boundary conditions should be posed along a velocity characteristic, and for a nonlinear equation the location of characteristics is unknown a priori.

While answering one question, this paper raises many others: e.g., finding larger classes of well-posed boundary problems for (1.1) and extending understanding of the model problem to (5.1) and its three-dimensional analogue.

Acknowledgment. We are grateful to Robert Bryant for helpful discussions of the geometry of (1.1) that culminated in the crucial result of Lemma 3.1.

REFERENCES

[1] R. L. Bryant, P. A. Griffiths, and L. Hsu, *Hyperbolic exterior differential systems and their conservation laws, Part* I, Selecta Math. (N.S.), 1 (1995), pp. 21–112.

[2] R. L. Bryant, P. A. Griffiths, and L. Hsu, *Hyperbolic exterior differential systems and their conservation laws, Part* II, Selecta Math. (N.S.), 1 (1995), pp. 265–323.

[3] C. Coulomb, *Essai sur une application des regles des maximis et minimis á l'architecture*, Mem. Math. Acad. R. Sci. Paris, 7 (1776), pp. 343–382.

[4] L. C. Evans, *Partial Differential Equations*, AMS, Providence, RI, 1998.

[5] P. R. Garabedian, *Partial Differential Equations*, Chelsea, New York, 1986.

[6] P. A. Gremaud and J. V. Matthews, *On the computation of steady hopper flows:* I*, Stress determination for Coulomb materials*, J. Comput. Phys., 166 (2001), pp. 63–83.

[7] R. Jackson, *Some mathematical and physical aspects of continuum models for the motion of granular materials*, in Theory of Dispersed Multiphase Flow, R. E. Meyer, ed., Academic Press, New York, 1983, pp. 291–337.

[8] A. Jenike, *Gravity Flows of Bulk Solids*, Bulletin 108, vol. 52, Utah Eng. Expt. Station, University of Utah, Salt Lake City, UT, 1961.

[9] J. V. Matthews, *An Analytical and Numerical Study of Granular Flows in Hoppers*, Ph.D. thesis, North Carolina State University, Raleigh, NC, 2000.

[10] E. B. Pitman, *The stability of granular flow in converging hoppers*, SIAM J. Appl. Math., 48 (1988), pp. 1033–1052.

[11] D. G. Schaeffer, M. Shearer, and T. P. Witelski, *A discrete model for an ill-posed nonlinear parabolic PDE*, Phys. D, 160 (2001), pp. 189–221.

[12] D. G. Schaeffer, *A mathematical model for localization in granular flow*, Proc. Roy. Soc. London Ser. A, 436 (1992), pp. 217–250.

[13] V. V. Sokolovskii, *Statics of Granular Media*, Pergamon Press, Oxford, 1965.

SIAM J. MATH. ANAL.
Vol. 36, No. 1, pp. 272–283

ON THE STRUCTURE OF SOLUTIONS TO THE PERIODIC HUNTER–SAXTON EQUATION*

ZHAOYANG YIN†

Abstract. We prove the local existence of strong solutions of the periodic Hunter–Saxton equation, and we show that all strong solutions except space-independent solutions blow up in finite time.

Key words. local existence of strong solutions, Hunter–Saxton equation, blow-up of strong solutions

AMS subject classifications. 35G25, 35L05

DOI. 10.1137/S0036141003425672

1. Introduction. In this paper, we study the periodic Hunter–Saxton equation [9]

$$\begin{cases} u_{txx} = -2u_x u_{xx} - u u_{xxx}, & t > 0,\ x \in \mathbb{R}, \\ u(0,x) = u_0(x), & x \in \mathbb{R}, \\ u(t,x+1) = u(t,x), & t \geq 0,\ x \in \mathbb{R}, \end{cases} \tag{1.1}$$

which describes the propagation of weakly nonlinear orientation waves in a massive nematic liquid crystal director field. Here, $u(t,x)$ describes the director field of a nematic liquid crystal, x is a space variable in a reference frame moving with the linearized wave velocity, and t is a slow time variable. Nematic liquid crystals are fluids consisting of long rigid molecules. The orientation of the molecules is described by the field of unit vectors

$$(\cos(u(t,x)), \sin(u(t,x))),$$

where $u(t,x)$ is a perturbation about some constant value. Equation (1.1) describes the weakly nonlinear dynamics of the director field of a nematic liquid crystal in the simplest possible setting which includes the effects of the inertia of the director field; cf. [9].

Equation (1.1) also arises in a different physical context as the high-frequency limit [7, 10] of the Camassa–Holm equation—a model equation for shallow water waves [2, 11] and a re-expression of the geodesic flow on the diffeomorphism group of the circle [5] with a bi-Hamiltonian structure [8] which is completely integrable [6]. The Hunter–Saxton equation also has a bi-Hamiltonian structure [9, 16] and is completely integrable [1, 10].

The initial value problem for the Hunter–Saxton equation on the line (nonperiodic case) has been studied by Hunter and Saxton in [9]. Using the method of characteristics, they show that smooth solutions exist locally and break down in finite time;

*Received by the editors April 3, 2003; accepted for publication (in revised form) September 12, 2003; published electronically June 22, 2004. This research was supported by the DFG-Graduiertenkolleg 615 and was also partially supported by the NNSF of China, the SRF for ROCS, SEM, the NSF of Guangdong Province, and the Foundation of Zhongshan University Advanced Research Center.

http://www.siam.org/journals/sima/36-1/42567.html

†Department of Mathematics, Zhongshan University, 510275 Guangzhou, China (mcsyzy@zsu.edu.cn).

cf. [9]. The occurrence of blow-up can be interpreted physically as the phenomenon by which waves that propagate away from the perturbation "knock" the director field out of its unperturbed state (see [9]).

However, the Cauchy problem of the periodic Hunter–Saxton seems not yet to have been discussed. The aim of this paper is to prove the local existence of strong solutions to (1.1) for a large class of initial data and to show that all strong solutions except space-independent solutions to (1.1) blow-up in finite time. Our methods are different from the ones used in [9] and the behavior of the solutions exhibits different features, for example regarding uniqueness (see Theorem 2.12).

Our paper is organized as follows. In section 2, we prove the local existence of the initial value problem associated with (1.1). In section 3, we investigate the blow-up phenomenon of strong solutions to (1.1).

Let us conclude the introduction with a short summary of the mathematical methodology that will be used in our approach. The existence of solutions to the nonlinear partial differential equation (1.1) is established by investigating an equivalent problem. We eliminate two spatial derivatives in (1.1) at the cost of obtaining a nonlocal nonlinear partial differential equation of order 1. Methods of functional analysis are then used to show the local existence of solutions and to address regularity issues. Finally, by looking at the time evolution of the minimum of the slope of a solution, we prove that all solutions of (1.1) with initial data that are not constant functions develop singularities in finite time.

2. Local well-posedness. In this section, we will apply Kato's theory to establish local existence for strong solutions to (1.1) in $H^r(\mathbb{S})$, $r > \frac{3}{2}$ with $\mathbb{S} = \mathbb{R}/\mathbb{Z}$ (the circle of unit length).

Let us first introduce some notation. Let A denote an unbounded operator and let $D(A)$ denote the domain of the operator A. $[A, B]$ denotes the commutator of the linear operators A and B. $\|\cdot\|_X$ denotes the norm of the Banach space X. In particular, $\|\cdot\|_r$ and $(\cdot,\cdot)_r$ denote the norm and the inner product of $H^r(\mathbb{S})$, $r \geq 0$, respectively.

For convenience, we state here Kato's theorem in the form suitable for our purpose.

Consider the abstract quasi-linear evolution equation:

$$\frac{dv}{dt} + A(v)v = f(t, v), \quad t \geq 0,\ v(0) = v_0. \tag{2.1}$$

Let X and Y be Hilbert spaces such that Y is continuously and densely embedded in X and let $Q : Y \to X$ be a topological isomorphism. $L(Y, X)$ denotes the space of all bounded linear operators from Y to X ($L(X)$, if $X = Y$). Assume the following.

(i) $A(y) \in L(Y, X)$ for $y \in X$ with

$$\|(A(y) - A(z))w\|_X \leq \mu_1 \|y - z\|_X \|w\|_Y, \quad y, z, w \in Y,$$

and $A(y) \in G(X, 1, \beta)$ (i.e., $A(y)$ is quasi-m-accretive), uniformly on bounded sets in Y.

(ii) $QA(y)Q^{-1} = A(y) + B(y)$, where $B(y) \in L(X)$ is bounded, uniformly on bounded sets in Y. Moreover,

$$\|(B(y) - B(z))w\|_X \leq \mu_2 \|y - z\|_Y \|w\|_X, \quad y, z \in Y,\ w \in X.$$

(iii) For each $y \in Y$, $t \to f(t,y)$ is continuous on $[0,\infty)$ to X. For each $t \in [0,\infty)$, $f(t,y) : Y \to Y$ and extends also to a map from X into X. For all $t \in [0,\infty)$, f is uniformly bounded on bounded sets in Y, and

$$\|f(t,y) - f(t,z)\|_Y \le \mu_3 \|y - z\|_Y, \quad t \in [0,\infty),\ y, z \in Y,$$
$$\|f(t,y) - f(t,z)\|_X \le \mu_4 \|y - z\|_X, \quad t \in [0,\infty),\ y, z \in X.$$

Here μ_1, μ_2, μ_3, and μ_4 depend only on $\max\{\|y\|_Y, \|z\|_Y\}$.

THEOREM 2.1 (Kato [12]). *Assume that* (i), (ii), *and* (iii) *hold. Given $v_0 \in Y$, there is a maximal $T > 0$ depending only on $\|v_0\|_Y$ and a unique solution v to* (2.1) *such that*

$$v = v(\cdot, v_0) \in C([0,T); Y) \bigcap C^1([0,T); X).$$

Moreover, the map $v_0 \mapsto v(\cdot, v_0)$ is continuous from Y to $C([0,T); Y) \bigcap C^1([0,T); X)$.

We provide now the framework in which we shall reformulate problem (1.1). In order to obtain an equation describing the evolution of u rather than that of u_{xx}, we observe that

$$-2u_x u_{xx} - u u_{xxx} = -\left(u u_{xx} + \frac{1}{2} u_x^2\right)_x.$$

Integrating both sides of (1.1) with respect to x, we obtain

$$\begin{cases} u_{tx} = -u u_{xx} - \dfrac{1}{2} u_x^2 + a, & t > 0,\ x \in \mathbb{R}, \\ u(0,x) = u_0(x), & x \in \mathbb{R}, \\ u(t, x+1) = u(t,x), & t \ge 0,\ x \in \mathbb{R}, \end{cases} \tag{2.2}$$

where $a = -\frac{1}{2}\int_{\mathbb{S}} u_x^2\, dx = -\frac{1}{2}\int_{\mathbb{S}} u_{0,x}^2\, dx$ is a constant (see Lemma 3.2 later in the paper). Then integrating both sides of (2.2) with respect to x, we have

$$\begin{cases} u_t + u u_x = \partial_x^{-1}\left(\dfrac{1}{2} u_x^2 + a\right) + h(t), & t > 0,\ x \in \mathbb{R}, \\ u(0,x) = u_0(x), & x \in \mathbb{R}, \\ u(t, x+1) = u(t,x), & t \ge 0,\ x \in \mathbb{R}, \end{cases} \tag{2.3}$$

where $a = -\frac{1}{2}\int_{\mathbb{S}} u_x^2\, dx$, $\partial_x^{-1} f(x) = \int_0^x f(x)\, dx$ and $h(t) : [0, +\infty) \to \mathbb{R}$ is an arbitrary continuous function.

THEOREM 2.2. *Given $h(t) \in C([0,+\infty); \mathbb{R})$ and $u_0 \in H^r(\mathbb{S})$, $r > \frac{3}{2}$. Then there exists a maximal $T = T(a, h(t), u_0) > 0$, and a unique solution u to* (2.3), *such that*

$$u = u(\cdot, u_0) \in C([0,T); H^r(\mathbb{S})) \bigcap C^1([0,T); H^{r-1}(\mathbb{S})).$$

Moreover, the solution depends continuously on the initial data, i.e., the mapping $u_0 \to u(\cdot, u_0) : H^r(\mathbb{S}) \to C([0,T); H^r(\mathbb{S})) \bigcap C^1([0,T); H^{r-1}(\mathbb{S}))$ is continuous.

Set $A(u) = u\partial_x$, $f(t,u) = \partial_x^{-1}(\frac{1}{2}u_x^2 + a) + h(t)$, $Y = H^r(\mathbb{S})$, $X = H^{r-1}(\mathbb{S})$, and $Q = \Lambda = (1 - \partial_x^2)^{\frac{1}{2}}$. Obviously, Q is an isomorphism of $H^r(\mathbb{S})$ onto $H^{r-1}(\mathbb{S})$. In order to prove Theorem 2.2, by applying Theorem 2.1, we only need to verify that $A(u)$ and $f(t,u)$ satisfy the conditions (i)–(iii).

The following three lemmas are useful for our approach.

LEMMA 2.3 (see [12]). *Let s, t be real numbers such that $-s < t \le s$. Then*

$$\|fg\|_t \le c\|f\|_s\|g\|_t \quad \text{if } s > \frac{1}{2},$$
$$\|fg\|_{s+t-\frac{1}{2}} \le c\|f\|_s\|g\|_t \quad \text{if } s < \frac{1}{2},$$

where c is a positive constant depending on s, t.

LEMMA 2.4 (see [13]). *Let $f \in H^r$, $r > \frac{3}{2}$. Then,*

$$\|\Lambda^{-s}[\Lambda^{s+t+1}, M_f]\Lambda^{-t}\|_{L(L^2(\mathbb{S}))} \le c\|f\|_r, \quad |s|, |t| \le r-1,$$

where M_f is the operator of multiplication by f, c is a constant depending only on r, t.

LEMMA 2.5 (see [17]). *Let X and Y be two Banach spaces such that Y is continuously and densely embedded in X. Let $-A$ be the infinitesimal generator of the C_0-semigroup $T(t)$ on X and let S be an isomorphism from Y onto X. Then Y is $-A$-admissible (i.e., $T(t)Y \subset Y$ for all $t \ge 0$, and the restriction of $T(t)$ to Y is a C_0-semigroup on Y) if and only if $-A_1 = -SAS^{-1}$ is the infinitesimal generator of the C_0-semigroup $T_1(t) = ST(t)S^{-1}$ on X. Moreover, if Y is $-A$-admissible, then the part of $-A$ in Y is the infinitesimal generator of the restriction of $T(t)$ to Y.*

The proof of Lemma 2.5 is given in section 4.5 (Theorems 5.5 and 5.8) in [17].

Next, we prove the following lemma.

LEMMA 2.6. *The operator $A(u) = u\partial_x$, with $u \in H^r(\mathbb{S})$, $r > \frac{3}{2}$, belongs to $G(L^2(\mathbb{S}), 1, \beta)$.*

Proof. Due to $L^2(\mathbb{S})$ being a Hilbert space, $A(u) \in G(L^2(\mathbb{S}), 1, \beta)$ [14] if and only if there is a real number β such that

(1) $(A(u)y, y)_0 \ge -\beta\|y\|_0^2$,

(2) the range of $A + \lambda$ is all of X for some (or all) $\lambda > \beta$.

First, let us prove (1). Due to $u \in H^r(\mathbb{S})$, $r > \frac{3}{2}$, it follows that u and u_x belong to $L^\infty(\mathbb{S})$. Note that $\|u_x\|_{L^\infty(\mathbb{S})} \le \|u\|_r$. Then we have

$$\begin{aligned}(A(u)y, y)_0 = (u\partial_x y, y)_0 &= -\frac{1}{2}(u_x y, y)_0 \\ &\le \frac{1}{2}\|u_x\|_{L^\infty(\mathbb{S})}\|y\|_0^2 \le c\|u\|_r\|y\|_0^2.\end{aligned}$$

Setting $\beta = c\|u\|_s$, we have $(A(u)y, y)_0 \ge -\beta\|y\|_0^2$.

Next, we prove (2). Because $A(u)$ is a closed operator and satisfies (1), it follows that $(\lambda I + A)$ has closed range in $L^2(\mathbb{S})$ for all $\lambda > \beta$. Thus, it suffices to prove that $(\lambda I + A)$ has dense range in $L^2(\mathbb{S})$ for all $\lambda > \beta$.

Given $u \in H^r(\mathbb{S})$, $r > \frac{3}{2}$, $y \in L^2(\mathbb{S})$. Then we have the generalized Leibnitz formula,

$$\partial_x(uy) = u_x y + u\partial_x y \quad \text{in } H^{-1}(\mathbb{S}).$$

Due to $u_x \in L^\infty(\mathbb{S})$, we obtain

$$\begin{aligned}D(A) = D(u\partial_x) &= \{y \in L^2(\mathbb{S}),\ u\partial_x y \in L^2(\mathbb{S})\} \\ &= \{z \in L^2(\mathbb{S}),\ -\partial_x(uz) \in L^2(\mathbb{S})\} = D((u\partial_x)^*) = D(A^*).\end{aligned}$$

Assume that the range of $(A+\lambda)$ is not all of $L^2(\mathbb{S})$. Then there exists $z \in L^2(\mathbb{S})$, $z \ne 0$, such that $((\lambda I + A)y, z)_0 = 0$ for all $y \in D(A)$. Since $H^1(\mathbb{S}) \subset D(A)$, we have

that $D(A)$ is dense in $L^2(\mathbb{S})$. So, it follows that $z \in D(A^*)$ and $\lambda z + A^* = 0$ in $L^2(\mathbb{S})$. Note that $D(A) = D(A^*)$. Multiplying by z and then integrating by parts, we obtain

$$0 = ((\lambda I + A^*)z, z)_0 = (\lambda z, z) + (z, Az) \geq (\lambda - \beta)\|z\|_0^2 \quad \forall \lambda > \beta.$$

Thus, we obtain $z = 0$. This contradicts the previous assumption $z \neq 0$ and completes the proof of Lemma 2.6. □

LEMMA 2.7. *The operator $A(u) = u\partial_x$, with $u \in H^r(\mathbb{S})$, $r > \frac{3}{2}$, belongs to $G(H^{r-1}(\mathbb{S}), 1, \beta)$.*

Proof. Due to $H^{r-1}(\mathbb{S})$ being a Hilbert space, $A(u)$ belongs to $G(H^{r-1}(\mathbb{S}), 1, \beta)$ [14] if and only if there is a real number β such that

(1) $(A(u)y, y)_{r-1} \geq -\beta\|y\|_{r-1}^2$,

(2) $-A(u)$ is the infinitesimal generator of a C_0-semigroup on $H^{r-1}(\mathbb{S})$, for some (or all) $\lambda > \beta$.

First, let us prove (1). Due to $u \in H^r(\mathbb{S})$, $r > \frac{3}{2}$, it follows that u and u_x belong to $L^\infty(\mathbb{S})$ and $\|u_x\|_{L^\infty(\mathbb{S})} \leq \|u\|_r$. Note that

$$\Lambda^{r-1}(u\partial_x y) = [\Lambda^{r-1}, u]\partial_x y + u\Lambda^{r-1}(\partial_x y) = [\Lambda^{r-1}, u]\partial_x y + u\partial_x \Lambda^{r-1} y.$$

Then we have

$$\begin{aligned}
(A(u)y, y)_{r-1} &= (\Lambda^{r-1}(u\partial_x y), \Lambda^{r-1} y)_0 \\
&= ([\Lambda^{r-1}, u]\partial_x y, \Lambda^{r-1} y)_0 - \frac{1}{2}(u_x \Lambda^{r-1} y, \Lambda^{r-1} y)_0 \\
&\leq \|[\Lambda^{r-1}, u]\Lambda^{2-r}\|_{L(L^2(\mathbb{S}))}\|\Lambda^{r-1} y\|_0^2 + \|u_x\|_{L^\infty(\mathbb{S})}\|\Lambda^{r-1} y\|_0^2 \\
&\leq c\|u\|_r\|y\|_{r-1}^2,
\end{aligned}$$

where we applied Lemma 2.4 with $s = 0$, $t = r - 2$. Setting $\beta = c\|u\|_r$, we have $(A(u)y, y)_{r-1} \geq -\beta\|y\|_{r-1}^2$.

Next, we prove (2). Let $S = \Lambda^{r-1}$. Note that S is an isomorphism of $H^{r-1}(\mathbb{S})$ onto $L^2(\mathbb{S})$ and that $H^{r-1}(\mathbb{S})$ is continuously and densely embedded in $L^2(\mathbb{S})$ as $r > \frac{3}{2}$. Define

$$A_1(u) := SA(u)S^{-1} = \Lambda^{r-1}A(u)\Lambda^{1-r}, \qquad B_1(u) = A_1(u) - A(u).$$

Let $y \in L^2(\mathbb{S})$ and $u \in H^r(\mathbb{S})$, $r > \frac{3}{2}$. Then we have

$$\begin{aligned}
\|B_1(u)y\|_0 &= \|[\Lambda^{r-1}, u\partial_x]\Lambda^{1-r} y\|_0 \\
&\leq \|[\Lambda^{r-1}, u]\Lambda^{2-r}\|_{L(L^2(\mathbb{S}))}\|\Lambda^{-1}\partial_x y\|_0 \\
&\leq c\|u\|_r\|y\|_0,
\end{aligned}$$

where we applied Lemma 2.4 with $s = 0$, $t = r - 2$. Therefore, we obtain $B_1(u) \in L(L^2(\mathbb{S}))$.

Note that $A_1(u) = A(u) + B_1(u)$ and $A(u) \in G(L^2(\mathbb{S}), 1, \beta)$ in Lemma 2.6. By a perturbation theorem for semigroups (cf. Section 5.2, Theorem 2.3 in [17]), we obtain $A_1(u) \in G(L^2(\mathbb{S}), 1, \beta')$. Applying Lemma 2.5 with $Y = H^{r-1}(\mathbb{S})$, $X = L^2(\mathbb{S})$, and $S = \Lambda^{r-1}$, we conclude that $H^{r-1}(\mathbb{S})$ is A-admissible. Therefore, $-A(u)$ is the infinitesimal generator of a C_0-semigroup on $H^{r-1}(\mathbb{S})$. This completes the proof of Lemma 2.7. □

LEMMA 2.8. *Let the operator $A(u) = u\partial_x$ with $u \in H^r(\mathbb{S})$, $r > \frac{3}{2}$. Then $A(u) \in L(H^r(\mathbb{S}), H^{r-1}(\mathbb{S}))$ for $u \in H^r(\mathbb{S})$. Moreover,*

$$\|(A(u) - A(z))w\|_{r-1} \le \mu_1 \|u - z\|_{r-1} \|w\|_r, \quad u, z, w \in H^r(\mathbb{S}).$$

Proof. Let $u, z, w \in H^r(\mathbb{S})$, $r > \frac{3}{2}$. Note that $H^{r-1}(\mathbb{S})$ is a Banach algebra. Then we have

$$\begin{aligned}\|(A(u) - A(z))w\|_{r-1} &\le c\|u - z\|_{r-1} \|\partial_x w\|_{r-1} \\ &\le \mu_1 \|u - z\|_{r-1} \|w\|_r.\end{aligned}$$

Taking $z = 0$ in the above inequality, we obtain $A(u) \in L(H^r(\mathbb{S}), H^{r-1}(\mathbb{S}))$. This completes the proof of Lemma 2.8. □

LEMMA 2.9. *$B(u) = [\Lambda^1, u\partial_x]\Lambda^{-1} \in L(H^{r-1}(\mathbb{S}))$ for $u \in H^r(\mathbb{S})$. Moreover,*

$$\|(B(u) - B(z))w\|_{r-1} \le \mu_2 \|u - z\|_r \|w\|_{r-1}.$$

Proof. Let $u, z \in H^r(\mathbb{S})$, $r > \frac{3}{2}$, $w \in H^{r-1}(\mathbb{S})$. Then

$$\begin{aligned}\|(B(u) - B(z))w\|_{r-1} &= \|\Lambda^{r-1}[\Lambda^1, (u - v)\partial_x]\Lambda^{-1} w\|_0 \\ &\le \|\Lambda^{r-1}[\Lambda, (u - v)]\Lambda^{1-r}\|_{L(L^2(\mathbb{S}))} \|\Lambda^{r-2}\partial_x w\|_0 \\ &\le \mu_2 \|y - z\|_r \|w\|_{r-1},\end{aligned}$$

where we applied Lemma 2.4 with $s = 1 - r,\ t = r - 1$. Taking $z = 0$ in the above inequality, we obtain $B(u) \in L(H^{r-1}(\mathbb{S}))$. This completes the proof of Lemma 2.9. □

LEMMA 2.10. *Let $f(t, u) = \partial_x^{-1}(\frac{1}{2}u_x^2 + a) + h(t)$, where $a = -\frac{1}{2}\int_{\mathbb{S}} u_x^2\, dx$. Then for each $y \in Y$, $t \to f(t, y)$ is continuous on $[0, \infty)$ to $H^{r-1}(\mathbb{S})$, $f(t, u)$ is uniformly bounded on bounded sets in $H^r(\mathbb{S})$ for all $t \in [0, \infty)$, and satisfies*

(1) $\|f(t, y) - f(t, z)\|_r \le \mu_3 \|y - z\|_s, \quad t \in [0, \infty),\ y, z \in H^r(\mathbb{S}),$

(2) $\|f(t, y) - f(t, z)\|_{r-1} \le \mu_4 \|y - z\|_{r-1}, \quad t \in [0, \infty),\ y, z \in H^r(\mathbb{S}).$

Proof. Due to $h(t) \in C([0, \infty); \mathbb{R})$, it follows that for each $y \in Y$, $t \to f(t, y)$ is continuous on $[0, \infty)$ to $H^{r-1}(\mathbb{S})$. Therefore, we only need to prove $f(t, u)$ satisfies (1) and (2). Let $y, z \in H^r(\mathbb{S})$, $r > \frac{3}{2}$. Note that $H^{r-1}(\mathbb{S})$ is a Banach algebra. Then we have

$$\begin{aligned}\|f(t, y) - f(t, z)\|_r &= \left\| \partial_x^{-1} \left(\frac{1}{2}(y_x^2 - z_x^2) \right) \right\|_r \\ &\le \frac{1}{2} \|(y_x - z_x)(y_x + z_x)\|_{r-1} \\ &\le \frac{1}{2} \|\partial_x (y - z)\|_{r-1} \|y_x + z_x\|_{r-1} \\ &\le \frac{1}{2} (\|y\|_r + \|z\|_r) \|y - z\|_r.\end{aligned}$$

This proves (1). Taking $z = 0$ in the above inequality, we obtain that f is uniformly bounded on bounded set in $H^r(\mathbb{S})$ for all $t \in [0, \infty)$.

Next, we prove (2). Let $y, z \in H^{r-1}(\mathbb{S})$, $r > \frac{3}{2}$. Note that $H^r(\mathbb{S})$ is a Banach algebra. Then we have

$$\begin{aligned}\|f(t,y) - f(t,z)\|_{r-1} &= \left\| \partial_x^{-1}\left(\frac{1}{2}(y_x^2 - z_x^2)\right)\right\|_{r-1} \\ &\le \frac{1}{2}\|(y_x - z_x)(y_x + z_x)\|_{r-2} \\ &\le \frac{1}{2}\|\partial_x(y-z)\|_{r-2}\|y_x + z_x\|_{r-1} \\ &\le \frac{1}{2}(\|y\|_r + \|z\|_r)\|y - z\|_{r-1},\end{aligned}$$

where we applied Lemma 2.3 with $s = r-1$, $t = r-2$. This completes the proof of Lemma 2.10. □

Proof of Theorem 2.2. Combining Theorem 2.1 and Lemmas 2.7–2.10, we can get the statement of Theorem 2.2. □

THEOREM 2.11. *The maximal time T in Theorem* 2.2 *may be chosen independent of r in the following sense. If $u = u(\cdot, u_0) \in C([0,T); H^r(\mathbb{S})) \bigcap C^1([0,T); H^{r-1}(\mathbb{S}))$ to* (2.3)*, and if $u_0 \in H^{r'}(\mathbb{S})$ for some $r' \neq r$, $r' > \frac{3}{2}$, then*

$$u \in C([0,T); H^{r'}(\mathbb{S})) \bigcap C^1([0,T); H^{r'-1}(\mathbb{S}))$$

and with the same T. In particular, $u_0 \in H^\infty(\mathbb{S}) = \bigcap_{r\ge 0} H^r(\mathbb{S})$, then $u \in C([0,T); H^\infty(\mathbb{S}))$.

Proof. It suffices to consider the case $r' > r$, since the case $r' < r$ is obvious from uniqueness which is guaranteed by Theorem 2.2. In order to prove that Theorem 2.11 is true for the case $r' > r$, let us return to (1.1). By setting $y(t) = \partial_x^2 u(t)$, we have

$$\frac{dy}{dt} + A(t)y + B(t)y = 0, \qquad y(0) = \partial_x^2 u(0). \tag{2.4}$$

Here, $A(t)y = \partial_x(u\,y)$ and $B(t)y = u_x y$.

Because $u \in C([0,T); H^r(\mathbb{S}))$ and $u_0 \in H^{r'}(\mathbb{S})$, we have $y \in C([0,T); H^{r-2}(\mathbb{S}))$ and $y(0) = \partial_x^2 u(0) \in C([0,T); H^{r'-2}(\mathbb{S}))$. It is our purpose to deduce $y \in C([0,T); H^{r'-2}(\mathbb{S}))$, which imply $u \in C([0,T); H^{r'}(\mathbb{S}))$, because ∂_x^2 is an isomorphism from $H^{r'}(\mathbb{S})$ to $H^{r'-2}(\mathbb{S})$. This will complete the proof of Theorem 2.11.

Note that $u \in C([0,T); H^r(\mathbb{S}))$, $u_x \in H^{r-1}(\mathbb{S})$, and $H^{r-1}(\mathbb{S})$ is a Banach algebra. Then we obtain $B(t) \in L(H^{r-1}(\mathbb{S}))$.

To this end, (see Lemmas 3.1–3.3 in [13]) we first need to prove that the family $A(t)$ has a unique evolution operator $\{U(t,\tau)\}$ associated with the spaces $X = H^h(\mathbb{S})$ and $Y = H^k(\mathbb{S})$, where $-r \le h \le r-2$, $1-r \le k \le r-1$, and $k \ge h+1$. Therefore, according to the proof of Lemma 3.1 in [13], we need to verify the following three conditions.

(i) $A(t) \in G(H^h(\mathbb{S}), 1, \beta)$ for all $y \in H^r(\mathbb{S})$.

(ii) $\Lambda^h \partial_x [\Lambda^{k-h}, u]\Lambda^{-k}$ is uniformly bounded on $L^2(\mathbb{S})$.

(iii) $A(t) \in L(H^k(\mathbb{S}), H^h(\mathbb{S}))$ is strongly continuous in t.

Let us begin by verifying condition (i). Due to $H^h(\mathbb{S})$ being a Hilbert space, $A(t) \in G(H^h(\mathbb{S}), 1, \beta)$ [14] if and only if there is a real number β such that

(1) $(A(t)y, y)_h \ge -\beta\|y\|_h^2$,

(2) $-A(t)$ is the infinitesimal generator of a C_0-semigroup on $H^h(\mathbb{S})$, for some (or all) $\lambda > \beta$.

First, we prove (1). Take $y \in H^h(\mathbb{S})$. Note that

$$\begin{aligned}\Lambda^h \partial_x(u\,y) &= \Lambda^h \partial_x(-[\Lambda^{-h}, u]\Lambda^h y + \Lambda^{-h}(u\Lambda^h y)) \\ &= -\Lambda^h \partial_x[\Lambda^{-h}, u]\Lambda^h y + \partial_x(u\Lambda^h y).\end{aligned}$$

Then we have

$$\begin{aligned}(A(t)y, y)_h &= (-\Lambda^h \partial_x[\Lambda^{-h}, u]\Lambda^h y + \partial_x(u\Lambda^h y), \Lambda^h y)_0 \\ &= (\Lambda^{h+1}[\Lambda^{-h}, u]\Lambda^h y, \partial_x \Lambda^{h-1} y)_0 + \frac{1}{2}(u_x \Lambda^h y, \Lambda^h y)_0 \\ &\le \|\Lambda^{h+1}[\Lambda^{-h}, u]\|_{L(L^2(\mathbb{S}))}\|\Lambda^h y\|_0^2 + \frac{1}{2}\|u_x\|_{L^\infty(\mathbb{S})}\|\Lambda^h y\|_0^2 \\ &\le c\|u\|_r \|y\|_h^2,\end{aligned}$$

where we applied Lemma 2.4 with $s = -(h+1)$, $t = 0$. Setting $\beta = c\|u\|_r$, we have $(A(t)y, y)_h \ge -\beta\|y\|_h^2$.

Second, we prove (2). Let $S = \Lambda^{r-1-h}$. Note that S is an isomorphism of $H^{r-1}(\mathbb{S})$ onto $H^h(\mathbb{S})$ and that $H^{r-1}(\mathbb{S})$ is continuously and densely embedded in $H^h(\mathbb{S})$ as $-r \le h \le r-2$. Define

$$\begin{aligned}A_1(t) &:= SA(t)S^{-1} = \Lambda^{r-1-h} A(t) \Lambda^{h+1-r}, \\ B_1(t) &:= A_1(t) - A(t) = [S, A(t)]S^{-1}.\end{aligned}$$

Let $y \in H^h(\mathbb{S})$ and $u \in H^r(\mathbb{S})$, $r > \frac{3}{2}$. Then we have

$$\begin{aligned}\|B_1(t)y\|_h &= \|\Lambda^h \partial_x[\Lambda^{r-1-h}, u]\Lambda^{h+1-r} y\|_0 \\ &\le \|\Lambda^h \partial_x[\Lambda^{r-1-h}, u]\Lambda^{1-r}\|_{L(L^2(\mathbb{S}))}\|\Lambda^h y\|_0 \\ &\le c\|u\|_r \|y\|_h,\end{aligned}$$

where we applied Lemma 2.4 with $s = -(h+1)$, $t = r-1$. Therefore, we obtain $B_1(t) \in L(H^h(\mathbb{S}))$. Note that

$$A(t)y = \partial_x(uy) = u_x y + u\partial_x y \quad \text{and} \quad u_x \in L(H^{r-1}(\mathbb{S})).$$

By applying Lemma 2.7 and a perturbation theorem for semigroups, we obtain that $H^{r-1}(\mathbb{S})$ is $A(t)$-admissible. Then, by applying Lemma 2.5 with $Y = H^{r-1}(\mathbb{S})$, $X = H^h(\mathbb{S})$, and $S = \Lambda^{r-1-h}$, we get that $-A_1(t)$ is the infinitesimal generator of a C_0-semigroup on $H^h(\mathbb{S})$. Note that $A_1(t) = A(t) + B_1(t)$ and $B_1(t) \in L(H^h(\mathbb{S}))$. By a perturbation theorem for semigroups, we have that $-A(t)$ is the infinitesimal generator of a C_0-semigroup on $H^h(\mathbb{S})$. This proves (b).

Next, we verify (ii). Take $y \in L^2(\mathbb{S})$. Then

$$\|\Lambda^h \partial_x[\Lambda^{k-h}, u]\Lambda^{-k} y\|_0 \le c\|u\|_s \|y\|_0,$$

where we applied Lemma 2.4 with $s = -(h+1)$, $t = k$.

Finally, we verify (iii). Take $y \in H^k(\mathbb{S})$. Then

$$\begin{aligned}\|(A(t+\tau) - A(t))y\|_h &= \|\partial_x((u(t+\tau) - u(t))y)\|_h \\ &\le \|(u(t+\tau) - u(t))y\|_{h+1} \\ &\le c\|u(t+\tau) - u(t)\|_{s-1}\|y\|_{h+1} \\ &\le c\|u(t+\tau) - u(t)\|_s \|y\|_k,\end{aligned}$$

where we applied Lemma 2.3 with $s = r-1$, $t = h+1$. So, by the continuity of u, we prove (iii). Thus, the above three conditions imply the existence and uniqueness of evolution operator $U(t,\tau)$ for the family $A(t)$. In particular $U(t,\tau)$ maps $H^s(\mathbb{S})$ into itself for $-r \le s \le r-1$.

Next, choosing $Y = H^{r-2}(\mathbb{S})$, $X = H^{r-3}(\mathbb{S})$, note that

$$y \in C([0,T); H^{r-1}(\mathbb{S})) \bigcap C^1([0,T); H^{r-2}(\mathbb{S})),$$

and by the properties of evolution operator $U(t,\tau)$, we can obtain

$$\frac{d}{d\tau}(U(t,\tau)y(\tau)) = -U(t,\tau)B(\tau)y(\tau).$$

An integration in $\tau \in [0,t]$ yields

$$y(t) = U(t,0)y(0) - \int_0^t U(t,\tau)B(\tau)y(\tau)d\tau. \tag{2.5}$$

If $r < r' \le r+1$, then we have that $B(t) = u_x(t) \in L(H^{r'-2}(\mathbb{S}))$ is strongly continuous in $[0,t)$, and

$$H^{r-1}(\mathbb{S})H^{r'-2}(\mathbb{S}) \subset H^{r'-2}(\mathbb{S})$$

by $r-1 > \frac{1}{2}$. Due to $-r < r-2 < r'-2 \le r-1$, the family $\{U(t,\tau)\}$ is strongly continuous on $H^{r'-2}(\mathbb{S})$ to itself. Note that $y(0) \in H^{r'-2}(\mathbb{S})$. Let us regard (2.5) as an integral equation of Volterra type which can be solved for y by successive approximation. Then the result of Theorem 2.11 is obtained.

If $r' > r+1$, then we obtain the result of Theorem 2.11 by repeating the application of the above argument. This completes the proof of Theorem 2.11. □

From (2.3), we see that if $h_1(t) \not\equiv h_2(t)$, then the corresponding solutions to (2.3) with the same initial data satisfy $u_{h_1}(t) \not\equiv u_{h_2}(t)$. Thus, for each initial data $u_0 \in H^r$, $r > \frac{3}{2}$, there exists an entire corresponding family of solutions to (1.1).

As a consequence of Theorems 2.2 and 2.11, we have the following.

THEOREM 2.12. *Given $u_0 \in H^r(\mathbb{S})$, $r > \frac{3}{2}$. Then there exists locally a family of solutions to* (1.1)*. Moreover, the maximal existence time T of each solution in the family can be chosen independent of r.*

Remark 2.13. If $u_0(x)$ is a constant, then the solution to equation (1.1) is of the form $u(t,x) = H(t)$ with $H \in C^1([0,T))$ for some $T > 0$ and $H(0) = u_0$. Because $\int_{\mathbb{S}} u_x^2(t,x)dx = 0$, it follows that $u_x(t,x) \equiv 0$.

3. Blow-up. In this section, we discuss the question of finite time blow-up of solutions to (1.1) with arbitrary initial data.

In our investigation we will use the following result.

LEMMA 3.1 (see [4]). *Let $T > 0$ and $v \in C^1([0,T); H^2(\mathbb{R}))$. Then for every $t \in [0,T)$, there exists at least one point $\xi(t) \in \mathbb{R}$ with*

$$m(t) := \inf_{x \in R}[v_x(t,x)] = v_x(t,\xi(t)).$$

The function $m(t)$ is absolutely continuous on $(0,T)$ with

$$\frac{dm}{dt} = v_{tx}(t,\xi(t)) \quad \textit{a.e. on } (0,T).$$

Let us now prove the following lemma.

LEMMA 3.2. *If $u_0 \in H^r$, $r \geq 3$, then as long as the solution $u(t,x)$ given by Theorem* 2.12 *exists, we have*

$$\int_{\mathbb{S}} u_x^2(t,x)dx = \int_{\mathbb{S}} u_{0,x}^2(x)dx.$$

Proof. Multiplying (1.1) by u and integrating with respect to x, in view of the periodicity of u, we get

$$\begin{aligned} -\frac{1}{2}\frac{d}{dt}\int_{\mathbb{S}} u_x^2 dx &= -\int_{\mathbb{S}} u_{tx}u_x\,dx = \int_{\mathbb{S}} u_{txx}u\,dx \\ &= -\int_{\mathbb{S}} 2u_x u_{xx}u\,dx - \int_{\mathbb{S}} u^2 u_{xxx}dx \\ &= -\int_{\mathbb{S}} 2u_x u_{xx}u\,dx + \int_{\mathbb{S}} 2u_x u_{xx}u\,dx = 0. \end{aligned}$$

Thus, we have

$$\int_{\mathbb{S}} u_x^2(t,x)\,dx = \int_{\mathbb{S}} u_x^2(0,x)\,dx.$$

This completes the proof of Lemma 3.2. □

We now present the following blow-up theorem.

THEOREM 3.3. *Assume that $u_0 \in H^r$, $r \geq 3$, and u_0 is not a constant. Then the corresponding solution to* (2.3) *blows up in finite time.*

Proof. Let $T > 0$ be the maximal existence time of the solution $u(t,\cdot)$ of (2.3) with initial data $u_0 \in H^3(\mathbb{S})$. By (2.2) and Lemma 3.2, we have

$$u_{tx} = -uu_{xx} - \frac{1}{2}u_x^2 - \frac{1}{2}\int_{\mathbb{S}} u_{0,x}^2(x)dx \quad \text{a.e. } t > 0. \tag{3.1}$$

Define $m(t) = u_x(t,\xi(t)) = \inf_{x\in R}\{u_x(t,x)\}$. Since we deal with a minimum, $u_{xx}(t,\xi(t)) = 0$ for all $t \in [0,T)$. We obtain

$$m'(t) = -\frac{1}{2}m^2(t) - \frac{1}{2}\int_{\mathbb{S}} u_{0,x}^2(x)dx \quad \text{a.e. } t > 0. \tag{3.2}$$

By the above equality, we can get

$$m(t) \leq m(0) - t\frac{1}{2}\int_{\mathbb{S}} u_{0,x}^2(x)dx, \quad t \geq 0.$$

Since $u_0(x)$ is not a constant, it follows that $\frac{1}{2}\int_{\mathbb{S}} u_{0,x}^2(x)dx > 0$. Therefore, there is some $t_0 \geq 0$ such that $m(t) < 0$ for $t \geq t_0$. On the other hand, by (3.2) we have

$$m'(t) \leq -\frac{1}{2}m^2(t) \quad \text{a.e. } t \geq t_0.$$

Thus, it follows that

$$0 > \frac{1}{m(t)} \geq \frac{t-t_0}{2} + \frac{1}{m(t_0)}, \quad t \geq t_0.$$

This forces $T < \infty$ and completes the proof of Theorem 3.3. □

To describe the blow-up mechanism, let us observe that the H^1-norm of the solution does not blow up in finite time. Indeed, multiplying (2.3) by $2u$ and integrating over the unit circle, we obtain that for all $t \in (0,T)$, where $T > 0$ is the maximal existence time of a solution to (2.3) with the initial data $u_0 \in H^r$, $r \geq 3$,

$$\begin{aligned}\frac{d}{dt}\int_{\mathbb{S}} u^2\,dx &= 2\int_{\mathbb{S}} u\partial_x^{-1}\left(\frac{1}{2}u_x^2 + a\right)\,dx + 2h(t)\int_{\mathbb{S}} u\,dx\\ &\leq \int_{\mathbb{S}} u^2\,dx + \int_{\mathbb{S}}\left[\partial_x^{-1}\left(\frac{1}{2}u_x^2 + a\right)\right]^2\,dx + |h(t)|\left[1 + \int_{\mathbb{S}} u^2\,dx\right]\\ &\leq |h(t)| + (1 + |h(t)|)\int_{\mathbb{S}} u^2\,dx + \int_0^1\left(\frac{1}{2}u_x^2 + |a|\right)\,dx\\ &= |h(t)| + (1 + |h(t)|)\int_{\mathbb{S}} u^2\,dx + |a| + \frac{1}{2}\int_0^1 u_{0,x}^2\,dx, \quad t \in (0,T),\end{aligned}$$

where we use Lemma 3.2 in the last step. By Gronwall's inequality, we infer that the L^2-norm of the solution does not blow up in finite time. Considering this in combination with Lemma 3.2, we see that the H^1-norm of any solution to (2.3) with initial data $u_0 \in H^r$, $r \geq 3$, does not blow up in finite time. However, if u_0 is not a constant, then the proof of Theorem 3.3 shows that $\inf u_x(t,\cdot) \to -\infty$ in finite time. We have therefore a weak type of singularity, similar to the case of the Camassa–Holm equation (see [3]) of which (1.1) is the high-frequency limit. As pointed out in the introduction, the blow-up can be interpreted as an altering of the director field from its original state.

As a consequence of Theorem 3.3, and in view of the connection between the family of solutions to (1.1) and the solution to (2.3), we have the following result.

THEOREM 3.4. *Assume that $u_0 \in H^r$, $r \geq 3$, and u_0 is not a constant. Then the corresponding solutions to* (1.1) *blow up in finite time.*

Remark 3.5. Although the Hunter–Saxton equation is the high-frequency limit [7, 10] of the Camassa–Holm equation [2], the structure of its solutions is different. While there are global smooth solutions of the periodic Camassa–Holm [3] with a genuine dependence on the spatial variable, Theorem 3.4 and Remark 2.13 show that the only global solutions of (1.1) are those independent of the spatial variable.

Acknowledgments. This work was performed while the author was a Visiting Researcher at the University of Hanover. The author thanks the referees for their valuable comments and helpful suggestions.

REFERENCES

[1] R. BEALS, D. SATTINGER, AND J. SZMIGIELSKI, *Inverse scattering solutions of the Hunter–Saxton equations*, Appl. Anal., 78 (2001), pp. 255–269.

[2] R. CAMASSA AND D. HOLM, *An integrable shallow water equation with peaked solitons*, Phys. Rev. Lett., 71 (1993), pp. 1661–1664.

[3] A. CONSTANTIN AND J. ESCHER, *Well-posedness, global existence, and blowup phenomena for a periodic quasi-linear hyperbolic equation*, Comm. Pure Appl. Math., 51 (1998), pp. 475–504.

[4] A. CONSTANTIN AND J. ESCHER, *Wave breaking for nonlinear nonlocal shallow water equations*, Acta Math., 181 (1998), pp. 229–243.

[5] A. CONSTANTIN AND B. KOLEV, *On the geometric approach to the motion of inertial mechanical systems*, J. Phys. A, 35 (2002), pp. R51–R79.

[6] A. CONSTANTIN AND H. P. MCKEAN, *A shallow water equation on the circle*, Comm. Pure Appl. Math., 52 (1999), pp. 949–982.

[7] H. H. Dai and M. Pavlov, *Transformations for the Camassa-Holm equation, its high-frequency limit and the Sinh-Gordon equation*, J. P. Soc. Japan, 67 (1998), pp. 3655–3657.
[8] A. Fokas and B. Fuchssteiner, *Symplectic structures, their Bäcklund transformations and hereditary symmetries*, Phys. D, 4 (1981/1982), pp. 47–66.
[9] J. K. Hunter and R. Saxton, *Dynamics of director fields*, SIAM J. Appl. Math., 51 (1991), pp. 1498–1521.
[10] J. K. Hunter and Y. Zheng, *On a completely integrable nonlinear hyperbolic variational equation*, Phys. D, 79 (1994), pp. 361–386.
[11] R. S. Johnson, *Camassa–Holm, Korteweg–de Vries and related models for water waves*, J. Fluid. Mech., 455 (2002), pp. 63–82.
[12] T. Kato, *Quasi-linear equations of evolution, with applications to partial differential equations*, in Spectral Theory and Differential Equations, W. N. Everitt, ed., Lecture Notes in Math. 448, Springer-Verlag, Berlin, 1975, pp. 25–70.
[13] T. Kato, *On the Korteweg-de Vries equation*, Manuscripta Math., 28 (1979), pp. 89–99.
[14] T. Kato, *On the Cauchy problem for the (generalized) Korteweg-de Vries equation*, in Studies in Applied Mathematics, V. Guillemin, ed., Adv. Math. Suppl. Stud. 8, Academic Press, New York, 1983, pp. 93–128.
[15] T. Kato and G. Ponce, *Commutator estimates and the Euler and Navier–Stokes equations*, Comm. Pure Appl. Math., 41 (1988), pp. 891–907.
[16] P. Olver and P. Rosenau, *Tri-Hamiltonian duality between solitions and solitary wave solutions having compact support*, Phys. Rev. E, 53 (1996), pp. 1900–1906.
[17] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

SIAM J. MATH. ANAL.
Vol. 36, No. 1, pp. 284–300

SCATTERED-DATA INTERPOLATION ON $\mathbb{R}$: ERROR ESTIMATES FOR RADIAL BASIS AND BAND-LIMITED FUNCTIONS*

FRANCIS J. NARCOWICH† AND JOSEPH D. WARD†

Abstract. Error estimates for scattered-data interpolation via radial basis functions (RBFs) for target functions in the associated reproducing kernel Hilbert space (RKHS) have been known for a long time. However, apart from settings where data is gridded, these estimates do not apply when the target functions generating the data are outside of the associated RKHS, and, in fact, no estimates were known in such situations. In this paper, we deal with these cases, obtaining Sobolev-type error estimates on compact regions of $\mathbb{R}^n$ when the RBFs have Fourier transforms that decay algebraically. In addition, we show that it is possible to construct band-limited interpolants that are also near-best approximants to such functions, with the band size being inversely proportional to the minimal separation of the data sites.

Key words. interpolation, scattered data, radial basis functions, band-limited functions, error estimates

AMS subject classifications. 41A25, 41A05, 41A63, 42B35

DOI. 10.1137/S0036141002413579

1. Introduction.

The problem of effectively representing an underlying function based on its values sampled at finitely many distinct scattered sites $X = \{x_1, \ldots, x_N\}$ lying in a compact region $\Omega \subset \mathbb{R}^n$ is important and arises in many applications—neural networks, computer aided geometric design, and gridless methods for solving partial differential equations, to name a few. A good example of the type of problem we have in mind is addressed in a recent paper by Carr et al. [6]. There, the authors used radial basis function (RBF) interpolation to reconstruct three-dimensional objects from "clouds" of points. Handling the large numbers of points was aided by new, fast evaluation techniques for RBFs [3].

The problem of representing a multivariate function by interpolating at scattered values is a difficult one. RBFs were introduced as a means to attack this problem. An RBF is a radial function $\Phi(x) = \Phi(|x|)$ that is either positive definite or conditionally positive definite on $\mathbb{R}^n$. Interpolants for multivariate functions sampled at scattered sites are constructed from translates of RBFs, with the possible addition of a polynomial term (see section 4 for details).

It was Duchon [7, 8] who introduced a type of RBF, the thin-plate spline, which he constructed via a variational technique similar to those used to obtain ordinary splines. The error analysis he provided for thin-plate splines involved reproducing kernel Hilbert space (RKHS) methods. Later, there were important contributions by Madych and Nelson [13] and Wu and Schaback [24], who also used RKHS methods to obtain scattered-data interpolation error estimates for a wide class of RBFs, including the Hardy multiquadrics and the Gaussians.

Important as these results are, they do suffer from a common difficulty. In all cases, convergence is proved only for functions in an RKHS that depends on Φ: the smoother the function Φ, the smaller the RKHS for which convergence estimates

*Received by the editors August 26, 2002; accepted for publication (in revised form) November 14, 2003; published electronically July 14, 2004. This research was supported by National Science Foundation grant DMS-0204449.

http://www.siam.org/journals/sima/36-1/41357.html

†Department of Mathematics, Texas A&M University, College Station, TX 77843-3368 (fnarc@math.tamu.edu, jward@math.tamu.edu).

apply. This restriction has seemed artificial, especially in light of both the lattice-based least-squares theory, which was completely and satisfactorily solved in [4], and the work of Schaback [19] dealing with pure approximation by RBFs. Indeed, Yoon [25] also noted this problem and introduced scaled RBFs in which a parameter λ is required to depend on the spacing of the data. In effect, the radial function is changing with the data.

In a recent paper [17], we dealt with these issues when the domain of the underlying function is the n-sphere, rather than $\mathbb{R}^n$, and the interpolants were derived from selected "translates" of spherical basis functions (SBFs). In particular, interpolatory error estimates were established for functions lying *outside* the RKHS for a wide variety of SBFs.

In this paper, we turn to the historically more important (and more difficult) $\mathbb{R}^n$ case, with the interpolants being derived from RBFs whose "centers" come from X. We start with samples $f|_X$ of a function $f \in C^k(\mathbb{R}^n) \cap W_2^k(\mathbb{R}^n)$ that decays to 0 at infinity but that may *not* be smooth enough to be in the RKHS associated with the RBF Φ. (If the target function f is defined only on Ω, then it may be necessary to extend it to all of $\mathbb{R}^n$. See the comments at the end of section 4.2.) For algebraic decay, i.e., $\hat{\Phi}(\xi) \sim |\xi|^{-2r}$, the RKHS is essentially $W_2^r(\mathbb{R}^n)$, so we assume that $k \leq r$. Let $I_X f$ be the RBF interpolant for f. We show that

$$\|D^\alpha f - D^\alpha f\|_{L^\infty(\Omega)} \leq C h^{k-|\alpha|-\frac{n}{2}} \|f\|_{C_0^k \cap W_2^k},$$

provided the sites X cover Ω in an approximately uniform manner and the multi-index satisfies $|\alpha| < k - \frac{n}{2}$. (The parameter h is a mesh norm.) The precise results are given in section 4.2. Prior to this, no estimates were known for $k \leq r$. Our results apply to both the thin-plate splines and to Wendland's class of compactly supported RBFs.

The first step to orchestrating this "escape" from the RKHS setting is showing that it is possible to use band-limited functions simultaneously to approximate and to interpolate an unknown, continuous, square-integrable function f that decays to 0 at infinity. This step is in a sense connected with irregular sampling theory [2]; of course, it differs in that samples $f|_X$ of f are taken on a *finite* set rather than an infinite one. In section 3.1, using functional analytic techniques, we show that there *is* a band-limited function f_σ, where $\operatorname{supp} \hat{f}_\sigma$ is contained in a ball of radius σ, that satisfies $f_\sigma|_X = f_X$ and that approximates f nearly optimally. The only requirement is that the size of the band or "Nyquist frequency" be inversely proportional to the minimal separation of points in X. The second step, taken in section 3.2, is obtaining Jackson-type estimates on distances to spaces of band-limited functions for $f \in C_0^k \cap W_2^k$ and then applying them (section 3.3) to bound the difference $f - f_\sigma$ in the appropriate norm. The last step makes use of these facts: both f and f_σ have the same RBF interpolant, because $f|_X = f_\sigma|_X$, and they have a comparable W_2^k norm. This allows us to deal with $\|D^\alpha f_\sigma - D^\alpha I_X f_\sigma\|$ in place of $\|D^\alpha f - D^\alpha I_X f\|$ and thus permits us to employ known RKHS estimates.

The main results of this paper may be viewed in two ways. From the theoretical point of view, they provide a much larger class of functions for which interpolation error estimates apply. From a practical point of view, they allow more flexibility in the choice of RBFs when applied for collocation purposes and faster convergence rates of interpolants away from singularities of the target function. This may even make possible using RBF methods for singularity detection. In addition, with a little more work, our methods should yield error estimates for discrete least-squares approximation by scattered shifts. We remark that in the situation of continuous least

squares over a domain without boundary, the least-squares approximation problem by shifts of Φ can be recast as an interpolation problem involving shifts of $\Phi * \bar{\Phi}$.

The remainder of the paper is organized as follows. In the next section, we provide the notation and set forth conventions we use throughout the paper. In section 3, we construct a band-limited function that both interpolates and is a near-best approximant to a given continuous L^2 function sampled on X. In addition, we use the Calderón formula to obtain Jackson-type estimates for this interpolant/approximant, provided the target function is in $C_0^k \cap W_2^k$. Finally, in the last section, we establish error estimates for interpolation via RBFs whose Fourier transforms decay algebraically.

2. Notation.

The set of centers. Let $X = \{x_j\}_{j=1}^N$ be a finite subset of $\mathbb{R}^n$, with the points all assumed to be distinct. There are two useful lengths associated with X. The first is the diameter of X, $\operatorname{diam}(X) = \max_{j,k} \|x_j - x_k\|_2$, which is the maximum distance between points in X. Throughout the paper, we will assume that $\operatorname{diam}(X) \leq 1$. The other length is the *separation radius*,

$$q = q_X := \frac{1}{2} \min_{j \neq k} \|x_j - x_k\|_2 \,,$$

which is half of the smallest distance between any two distinct points in X. Clearly, $q \leq \frac{1}{2}\operatorname{diam}(X) \leq \frac{1}{2}$. The set X is itself contained in an n-cube with sides of length $\operatorname{diam}(X) \leq 1$. The union of X, together with the closed balls $B(x_j, q)$ having centers $x_j \in X$ and radius q, can be enclosed in an n-cube with sides of length $\operatorname{diam}(X)+2q \leq 2$. The number of points in X, N can be estimated in terms of the volume of this second cube:

$$N \leq \frac{2^n}{\operatorname{vol}(B(0,q))} = \frac{\Gamma(\frac{n+2}{2})2^n}{\pi^{\frac{n}{2}} q^n}. \tag{1}$$

The region Ω. We take Ω to be a compact, connected region in $\mathbb{R}^n$ that satisfies the *uniform interior cone condition* [11]; i.e., there exists a fixed (open) cone $K \subset \mathbb{R}^n$ such that each $x \in \partial\Omega$ is the vertex of a cone $K_x \subset \Omega$ that is congruent to K. The *mesh norm* for X relative to Ω is

$$h = h_{X,\Omega} := \sup_{x \in \Omega} \inf_{x_j \in X} \|x - x_j\|_2 \,;$$

it measures the maximum distance any point in Ω can be from X. It is easy to see that $h_{X,\Omega} \geq q_X$; equality can hold only for a uniform distribution of points on an interval in $\mathbb{R}^1$. The *mesh ratio*

$$\rho = \rho_{X,\Omega} := h/q \geq 1$$

provides a measure of how uniformly points in X are distributed in Ω. When Ω is an interval ($n = 1$), $\rho = 1$ means that the points are uniformly distributed. In all other cases, $\rho > 1$.

Conventions. Our conventions for the Fourier transform and its inverse are

$$\hat{f}(\xi) := \int_{\mathbb{R}^n} f(x) e^{-i\xi\cdot x} d^n x \quad \text{and} \quad \check{f}(x) = \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} f(\xi) e^{i\xi\cdot x} d^n \xi.$$

We will make use of the Sobolev space $W_2^k = W_2^k(\mathbb{R}^n)$, which is defined to be all $f \in L^2$ having distributional derivatives $D^\alpha f$, $|\alpha| \leq k$, in L^2. The norm that we will

use here is

$$\|f\|_{W_2^k} = \left(\int_{\mathbb{R}^n} (1 + |\xi|^2)^k |\hat{f}(\xi)|^2 d^n\xi \right)^{\frac{1}{2}}.$$

We will denote by $C_0^k = C_0^k(\mathbb{R}^n)$ the set of all functions that are continuously differentiable through order k, that vanish at infinity, and that have all derivatives of order k or less bounded. Of course, $C_0^0 = C_0$ and $C_0^k = C_0(\mathbb{R}^n) \cap C_B^k(\mathbb{R}^n)$. We will use $\mathcal{S}$ to denote Schwartz space and $\mathcal{S}'$ to denote the space of tempered distributions.

Frequently, we will deal with the intersection of two function spaces, for example $C_0 \cap L^2$ or $C_0^k \cap W_2^k$. In these cases, the norm on the intersection will be the maximum of the norm on each space: if $\mathcal{X}$ and $\mathcal{Y}$ are normed spaces of functions, then on the intersection space $\mathcal{X} \cap \mathcal{Y}$ we will always use the norm

$$\|f\|_{\mathcal{X}\cap\mathcal{Y}} := \max\left(\|f\|_{\mathcal{X}}, \|f\|_{\mathcal{Y}}\right).$$

Finally, for an integer $k \geq 0$, we will denote the polynomials of total degree k on $\mathbb{R}^n$ by $\pi_k(\mathbb{R}^n)$. For $k = -1$, we will let $\pi_{-1}(\mathbb{R}^n) = \{0\}$.

3. Band-limited functions. We discuss interpolation and approximation in the Paley–Wiener class of band-limited functions. Since we are dealing with a multi-dimensional space, we will interpret the transform variable ξ as a wavenumber having units of reciprocal length. Let $\sigma > 0$. We then define $\mathcal{B}_\sigma$ to be

$$\mathcal{B}_\sigma := \{f \in L^2(\mathbb{R}^n) \colon \operatorname{supp}(\hat{f}) \subseteq B(0,\sigma)\},$$

where $B(0,\sigma)$ is the (closed) ball in $\mathbb{R}^n$ having center 0 and radius σ. (In optics, σ denotes the spectroscopic wavenumber and is the reciprocal of the wavelength.)

Functions in $\mathcal{B}_\sigma$ are, of course, in L^2 and are continuous. Moreover, they decay to 0 as $|x| \to \infty$. A natural class of functions that includes all $\mathcal{B}_\sigma$ is $C_0 \cap L^2 := C_0(\mathbb{R}^n) \cap L^2(\mathbb{R}^n)$. We can make this a Banach space by employing the norm

$$\|f\|_{C_0\cap L^2} = \max(\|f\|_\infty, \|f\|_2).$$

3.1. Approximation and interpolation from $\mathcal{B}$. The main goal of this section is to show that we can both approximate and interpolate functions in $C_0 \cap L^2$ on X by means of band-limited functions in $\mathcal{B}_\sigma$ if we take $\sigma \sim 1/q$. In finite-dimensional settings, similar problems for interpolation and approximation by polynomials go back to the work of Erdös [9], and problems for trigonometric polynomials on the circle and spherical harmonics on the n-sphere were discussed in [14, 17]. The case of $\mathcal{B}_\sigma$, which is infinite dimensional, is dealt with here.

We begin with the proposition below, which is stated in terms of Banach spaces; in it we show that the problem of finding interpolants that are also near-best approximants can be solved if we can uniformly bound the ratio of the norm of a linear functional to the norm of its restriction to a smaller space.

PROPOSITION 3.1. *Let $\mathcal{Y}$ be a (possibly complex) Banach space, $\mathcal{V}$ be a subspace of $\mathcal{Y}$, and Z^* be a finite-dimensional subspace of $\mathcal{Y}^*$, the dual of $\mathcal{Y}$. If, for every $z^* \in Z^*$ and some $\beta > 1$, β independent of z^*,*

$$\|z^*\|_{\mathcal{Y}^*} \leq \beta \|z^*|_{\mathcal{V}}\|_{\mathcal{V}^*}, \tag{2}$$

then for $y \in \mathcal{Y}$ there exists $v \in \mathcal{V}$ such that v interpolates y on Z^*; that is, $z^*(y) = z^*(v)$ for all $z^* \in Z^*$. In addition, v approximates y in the sense that $\|y - v\|_{\mathcal{Y}} \leq (1 + 2\beta)\operatorname{dist}(y, \mathcal{V})$.

Proof. Given $\varepsilon > 0$, we can find $u \in \mathcal{V}$ for which

$$\|y - u\|_{\mathcal{Y}} = (1+\varepsilon)\text{dist}(y, \mathcal{V}).$$

Set $x := y - u$. Let the restriction map $S\colon\ Z^* \to Z^*|_{\mathcal{V}}$ be given by $S(z^*) = z^*|_{\mathcal{V}}$ for every $z^* \in Z^*$. By (2), S is both one-to-one and onto the image space $S(Z^*) \subset \mathcal{V}^*$. Moreover, $\|S^{-1}\| \le \beta$, where $S^{-1}\colon S(Z^*) \to Z^*$. Viewing x as an element of Z^{**} (i.e., as a functional on Z^*), we have

$$\langle x, z^* \rangle = \langle x, S^{-1}Sz^* \rangle = \langle (S^*)^{-1}x, Sz^* \rangle,$$

where we used the fact that $(S^{-1})^* = (S^*)^{-1}$. Note that $(S^*)^{-1}x$ is in $(S(Z^*))^*$, where $S(Z^*)$ is a finite-dimensional subspace of $\mathcal{V}^*$. By the Hahn–Banach theorem, $(S^*)^{-1}x$ extends in a norm preserving manner to $v_x^{**} \in \mathcal{V}^{**}$. Thus $\langle x, z^* \rangle = \langle y - u, z^* \rangle = \langle v_x^{**}, Sz^* \rangle$ for all $z^* \in Z^*$, and

$$\begin{aligned} \|v_x^{**}\|_{V^{**}} &= \|(S^*)^{-1}x\|_{V^{**}} \le \|S^{-1}\|\,\|x\|_{\mathcal{Y}} \le \beta\|y-u\|_{\mathcal{Y}} \\ &= \beta(1+\epsilon)\,\text{dist}(y, \mathcal{V}). \end{aligned}$$

We would be done if the spaces involved were reflexive, because then we would have v_x^{**} in $\mathcal{V}$, and we could simply set $v = u + v_x^{**}$. Unfortunately, the spaces of interest are *not* reflexive. However, the fact that Z^* is finite dimensional allows us to apply the *principle of local reflexivity* [12, p. 53], which states that for $\delta > 0$ we can find $v_x \in \mathcal{V}$ such that both $\langle v_x^{**}, z^* \rangle = \langle z^*, v_x \rangle$ for all $z^* \in Z^*$ and $\|v_x\| \le (1+\delta)\|v_x^{**}\|$. Setting $v := u + v_x$ gives an element in $\mathcal{V}$ that interpolates y on Z^* and satisfies

$$\begin{aligned} \|y - v\|_{\mathcal{Y}} &\le \|y-u\|_{\mathcal{Y}} + \|v_x\|_{\mathcal{Y}} \le \|y-u\|_{\mathcal{Y}} + (1+\delta)\|v_x^{**}\|_{\mathcal{V}^{**}} \\ &\le (1 + (1+\delta)\beta)\|y-u\|_{\mathcal{Y}} \\ &\le (1 + (1+\delta)\beta)(1+\varepsilon)\text{dist}(y, \mathcal{V}). \end{aligned}$$

To complete the proof, take $\delta < 1/5$ and $\varepsilon < 1/4$ and note that, because $\beta > 1$, we have $(1 + (1+\delta)\beta)(1+\varepsilon) < \frac{5}{4} + \frac{6}{4}\beta < 1 + 2\beta$. ∎

We are interested in the case in which $\mathcal{Y} = C_0 \cap L^2$, $Z^* = \text{span}\{\delta_{x_j}\colon x_j \in X\}$, and $\mathcal{V} = \mathcal{B}_\sigma$. To employ the proposition, the first thing that we need to do is find $\|z^*\|_{C_0 \cap L^{2*}}$ when $z^* := \sum_{x_j \in X} c_j \delta_{x_j} \in Z^*$.

We will use "bump" functions to do this. Let $g_R(x) = (1 - |x|/R)_+$. The support of g_R is the closed ball of radius R and center 0. It is easy to show that $\|g_R\|_2 = C_n R^{n/2}$ and $\|g_R\|_\infty = 1$. Next, we choose $R < \min\{q, (NC_n^2)^{-1/n}\}$, where q is the separation radius of X. Also, take $d_j = \bar{c}_j/|c_j|$ if $c_j \neq 0$ and set $d_j = 1$ if $c_j = 0$. The supports of the $g_R(x - x_j)$'s, $x_j \in X$, are then all disjoint and for $f(x) = \sum_j d_j\, g_R(x - x_j)$ we have

$$\langle z^*, f \rangle = \sum_{j,k} c_k d_j g_R(x_k - x_j) = \sum_j c_j d_j = \sum_j |c_j|.$$

In addition, we have $\|f\|_\infty = 1$ and $\|f\|_2 = N^{1/2}\|g_R\|_2 = C_n R^{n/2} N^{1/2} < 1$. Consequently, $\|f\|_{C_0 \cap L^2} = 1$, and $\langle z^*, f \rangle = \sum_j |c_j| = (\sum_j |c_j|)\|f\|_{C_0 \cap L^2}$. It follows that $\|z^*\|_{C_0 \cap L^{2*}} \ge \sum_j |c_j|$. On other hand, we have that $|\langle z^*, f \rangle| \le (\sum_j |c_j|)\|f\|_\infty \le (\sum_j |c_j|)\|f\|_{C_0 \cap L^2}$. We arrive at the following lemma.

LEMMA 3.2. *If $z^* := \sum_{x_j \in X} c_j \delta_{x_j} \in Z^*$, then*

$$\|z^*\|_{C_0 \cap L^{2*}} = \sum_j |c_j|. \tag{3}$$

The next step in showing that we can both interpolate and approximate with band-limited functions is to estimate $\|z^*|_{\mathcal{B}_\sigma^*}\|_{\mathcal{B}_\sigma^*}$. The approach we take runs parallel to the one above but is technically more difficult because we must work with functions in $\mathcal{B}_\sigma$, all of which are analytic and thus *not* compactly supported. Fortunately, many of the computations required were done in [16, section 3]. Indeed, we summarize what we need below, with appropriate adaptations to our current notation.

We begin by noting that if $\chi_{\frac{\sigma}{2}}(\xi)$ is the characteristic function for the ball $B(0,\sigma/2)$, then its inverse Fourier transform is given by [16, eq. (3.9)]

$$\check{\chi}_{\frac{\sigma}{2}}(x) = \left(\frac{\sigma}{4\pi|x|}\right)^{\frac{n}{2}} J_{\frac{n}{2}}\left(\frac{|x|\sigma}{2}\right),$$

where one also has [16, eq. (3.10)]

$$\check{\chi}_{\frac{\sigma}{2}}(0) = \frac{\left(\frac{\sigma}{(4\sqrt{\pi})}\right)^n}{\Gamma(\frac{n+2}{2})}.$$

We define φ_σ via

$$\varphi_\sigma := \check{\chi}^2_{\frac{\sigma}{2}} = (2\pi)^{-n}(\chi_{\frac{\sigma}{2}} * \chi_{\frac{\sigma}{2}})^{\check{}}, \tag{4}$$

where the second equality follows from the convolution theorem. We also have from the expression for $\check{\chi}(0)$ that

$$\varphi_\sigma(0) = \frac{\left(\frac{\sigma}{(4\sqrt{\pi})}\right)^{2n}}{\Gamma(\frac{n+2}{2})^2}. \tag{5}$$

Because the support of $\chi_{\frac{\sigma}{2}}$ is $B(0,\sigma/2)$, we have $\operatorname{supp}(\chi_{\frac{\sigma}{2}} * \chi_{\frac{\sigma}{2}}) \subset B(0,\sigma)$. This, of course, implies that $\varphi_\sigma \in \mathcal{B}_\sigma$ and that

$$\Upsilon(x) := \sum_{j=1}^{N} \varphi_\sigma(x - x_j) \in \mathcal{B}_\sigma. \tag{6}$$

LEMMA 3.3. *If* $z^* := \sum_{x_j \in X} c_j \delta_{x_j} \in Z^*$, *then*

$$\|z^*|_{\mathcal{B}_\sigma}\|_{\mathcal{B}_\sigma^*} \geq \left(\frac{\varphi_\sigma(0) - \max_k \sum_{j\neq k} \varphi_\sigma(x_j - x_k)}{\|\Upsilon\|_{C_0\cap L^2}}\right)\|z^*\|_{C_0\cap L^{2*}}. \tag{7}$$

Proof. Set $d_j = \bar{c}_j/|c_j|$ if $c_j \neq 0$ and $d_j = 1$ if $c_j = 0$. If we let

$$b_\sigma = \sum_{x_j \in X} d_j \varphi(x - x_j),$$

then we can compute $\langle z^*, b_\sigma\rangle$ using $z^* = \sum_{x_j\in X} c_j\delta_{x_j}$ and the definition of b_σ. What we obtain is $\langle z^*, b_\sigma\rangle = \sum_{j,k} c_j\, d_k \varphi_\sigma(x_j - x_k)$. We can then use this together with

$\sum_j c_j d_j = \sum_j |c_j|$ to obtain the following:

$$\begin{aligned}
|\langle z^*, b_\sigma\rangle| &= \left| \sum_j c_j\, d_j \varphi_\sigma(0) + \sum_{j\neq k} c_j\, d_k \varphi_\sigma(x_j - x_k)\right| \\
&\geq \left(\sum_j |c_j|\right)\varphi_\sigma(0) - \left| \sum_{j\neq k} c_j\, d_k \varphi_\sigma(x_j - x_k)\right| \\
&\geq \left(\sum_j |c_j|\right)\left(\varphi_\sigma(0) - \max_k \sum_{j\neq k} \varphi_\sigma(x_j - x_k)\right) \\
&\geq \|z^*\|_{C_0\cap L^{2*}}\left(\varphi_\sigma(0) - \max_k \sum_{j\neq k} \varphi_\sigma(x_j - x_k)\right), \qquad (8)
\end{aligned}$$

where the last inequality uses Lemma 3.2. Next, since $|d_j| = 1$, we have that b_σ satisfies

$$|b_\sigma(x)| \leq \sum_{x_j\in X} \varphi(x - x_j) = \Upsilon(x).$$

Thus, $\|b_\sigma\|_\infty \leq \|\Upsilon\|_\infty$ and $\|b_\sigma\|_2 \leq \|\Upsilon\|_2$, and, consequently,

$$\|b_\sigma\|_{C_0\cap L^2} \leq \|\Upsilon\|_{C_0\cap L^2}. \qquad (9)$$

Dividing both sides of (8) by $\|b_\sigma\|_{C_0\cap L^2}$ and then using (9) results in

$$\frac{|\langle z^*, b_\sigma\rangle|}{\|b_\sigma\|_{C_0\cap L^2}} \geq \left(\frac{\varphi_\sigma(0) - \max_k \sum_{j\neq k} \varphi_\sigma(x_j - x_k)}{\|\Upsilon\|_{C_0\cap L^2}}\right)\|z^*\|_{C_0\cap L^{2*}}.$$

Since $b_\sigma \in \mathcal{B}_\sigma$, the left-hand side above is bounded by $\|z^*|_{\mathcal{B}_\sigma}\|_{\mathcal{B}_\sigma^*}$. Using this bound in the last inequality results in (7). □

We will need estimates on sums of translates of φ_σ. This is the object of our next result.

PROPOSITION 3.4. *If $X = \{x_1, x_2, \ldots, x_N\} \subset \mathbb{R}^n$ is a set of N distinct points having separation radius q and satisfying* $\operatorname{diam}(X) \leq 1$, *then*

$$\max_k \sum_{j\neq k} \varphi_\sigma(x_j - x_k) \leq \varphi_\sigma(0)\frac{\pi\Gamma^2(\frac{n+2}{2})}{18}\left(\frac{\sigma q}{24}\right)^{-n-1}. \qquad (10)$$

In addition, if $\Upsilon(x) := \sum_{j=1}^N \varphi_\sigma(x - x_j)$, then we have

$$\|\Upsilon\|_{C_0\cap L^2} \leq \varphi_\sigma(0)\max\left(\frac{\Gamma(\frac{n+2}{2})8^{\frac{n}{2}}}{(\sigma q)^{\frac{n}{2}}}, 1\right)\left(1 + \frac{\pi\Gamma^2(\frac{n+2}{2})}{18}\left(\frac{\sigma q}{48}\right)^{-n-1}\right). \qquad (11)$$

Proof. The bound in (10) was actually established in [16, section III] and in [15, section IV] but with different notation. In [16], β corresponds to $\sigma/2$ and $\chi(x)/K$ to φ_σ. From [15, eqs. (4.4) and (4.11)],

$$\max_k \sum_{j\neq k} \varphi_\sigma(x_j - x_k) \leq 3^n \sum_{k=1}^\infty k^{n-1}\kappa_k := 3^n\Sigma.$$

The quantity Σ is estimated on [16, p. 96]. Adjusting the formula there by dividing by K and replacing β by $\sigma/2$, we obtain the bound in (10). To bound $\Upsilon(x)$, we first choose the point in $\{x_1, x_2, \ldots, x_N\}$ nearest to x. After renumbering, we may take this to be x_1, and we may rewrite the sum in Υ as

$$\Upsilon(x) = \varphi_\sigma(x - x_1) + \sum_{j=2}^{N} \varphi_\sigma(x - x_j).$$

The function φ_σ is a positive definite function, because $\chi_{\frac{\sigma}{2}} * \chi_{\frac{\sigma}{2}}$ is nonnegative. Consequently, $\varphi_\sigma(x - x_1) \le \varphi_\sigma(0)$. To bound the remaining terms in Υ, we first note that the separation radius for the set $\{x, x_2, \ldots, x_N\}$ is at least $\frac{1}{2}q$. Applying the bound in (10) to the remaining terms in Υ, while replacing q by $\frac{1}{2}q$, we obtain

$$\|\Upsilon\|_\infty \le \varphi_\sigma(0) \left(1 + \frac{\pi\Gamma^2(\frac{n+2}{2})}{18} \left(\frac{\sigma q}{48}\right)^{-n-1}\right). \tag{12}$$

Given this bound, we can get the bound on $\|\Upsilon\|_2$ this way. Note that $\|\Upsilon\|_1 = \sum_{j=1}^{N} \|\varphi_\sigma(x - x_j)\|_1 = N\|\varphi_\sigma\|_1$. Since $\varphi_\sigma = \check{\chi}^2_{\frac{\sigma}{2}} = (2\pi)^{-n}(\chi_{\frac{\sigma}{2}} * \chi_{\frac{\sigma}{2}})\hat{} \ge 0$, we see that

$$\|\varphi_\sigma\|_1 = \hat{\varphi}_\sigma(0) = (2\pi)^{-n}\chi_{\frac{\sigma}{2}} * \chi_{\frac{\sigma}{2}}(0) = (2\pi)^{-n}\|\chi_{\frac{\sigma}{2}}\|_2^2.$$

Moreover, $\chi_{\frac{\sigma}{2}}$ is a characteristic function; hence, $\chi^2_{\frac{\sigma}{2}} = \chi_{\frac{\sigma}{2}}$ and

$$(2\pi)^{-n}\|\chi_{\frac{\sigma}{2}}\|_2^2 = (2\pi)^{-n}\|\chi_{\frac{\sigma}{2}}\|_1 = \check{\chi}_{\frac{\sigma}{2}}(0) = \varphi_\sigma(0)^{\frac{1}{2}}.$$

Finally, we obtain

$$\|\varphi_\sigma\|_1 = \varphi_\sigma(0)^{\frac{1}{2}} \quad \text{and} \quad \|\Upsilon\|_1 = N\varphi_\sigma(0)^{\frac{1}{2}}. \tag{13}$$

Applying the standard inequality $\|\Upsilon\|_2 \le \left(\|\Upsilon\|_1\|\Upsilon\|_\infty\right)^{\frac{1}{2}}$ in conjunction with (12) and (13), we have that

$$\begin{aligned}\|\Upsilon\|_2 &\le \sqrt{\frac{N}{\varphi_\sigma(0)^{\frac{1}{2}}}}\,\varphi_\sigma(0)\left(1 + \frac{\pi\Gamma^2(\frac{n+2}{2})}{18}\left(\frac{\sigma q}{48}\right)^{-n-1}\right)^{\frac{1}{2}} \\ &\le \frac{\Gamma(\frac{n+2}{2})8^{\frac{n}{2}}}{(\sigma q)^{\frac{n}{2}}}\,\varphi_\sigma(0)\left(1 + \frac{\pi\Gamma^2(\frac{n+2}{2})}{18}\left(\frac{\sigma q}{48}\right)^{-n-1}\right)^{\frac{1}{2}}, \qquad (14)\end{aligned}$$

where in the last step we have used (1) and (5) to bound $\sqrt{N/\varphi_\sigma(0)^{\frac{1}{2}}}$. The final estimate (11) follows from taking the maximum of the right-hand sides of (12) and (14) and then using the fact that $x^{\frac{1}{2}} \le x$ when $x \ge 1$. □

We now come to the main result of this section.

THEOREM 3.5. *Let $X = \{x_1, x_2, \ldots, x_N\} \subset \mathbb{R}^n$ be a set of N distinct points having separation radius q and satisfying* $\operatorname{diam}(X) \le 1$, *and choose σ so that*

$$\sigma \ge \sigma_0 := \frac{24}{q}\left\{\frac{\sqrt{\pi}}{3}\Gamma\left(\frac{n+2}{2}\right)\right\}^{\frac{2}{n+1}}. \tag{15}$$

If $f \in C_0 \cap L^2$, there exists $f_\sigma \in \mathcal{B}_\sigma$ such that

$$f|_X = f_\sigma|_X \quad \textit{and} \quad \|f - f_\sigma\|_{C_0 \cap L^2} \le (5 + 2^{n+3})\operatorname{dist}_{C_0 \cap L^2}(f, \mathcal{B}_\sigma). \tag{16}$$

Proof. We made this choice of σ_0 so that

$$\frac{\pi\Gamma^2(\frac{n+2}{2})}{18}\left(\frac{24}{\sigma q}\right)^{n+1} \leq \frac{\pi\Gamma^2(\frac{n+2}{2})}{18}\left(\frac{24}{\sigma_0 q}\right)^{n+1} = \frac{1}{2}\,. \tag{17}$$

Thus, the inequality in (10) becomes

$$\max_k \sum_{j\neq k} \varphi_\sigma(x_j - x_k) \leq \frac{1}{2}\varphi_\sigma(0).$$

Inserting this on the right-hand side in (7) yields

$$\|z^*|_{\mathcal{B}_\sigma}\|_{\mathcal{B}_\sigma^*} \geq \left(\frac{\varphi_\sigma(0)}{2\|\Upsilon\|_{C_0\cap L^2}}\right)\|z^*\|_{C_0\cap L^{2^*}}\,. \tag{18}$$

We now need to estimate $\|\Upsilon\|_{C_0\cap L^2}$. First note that, since σ_0 satisfies the equation on the right-hand side in (17), we have

$$\frac{\Gamma(\frac{n+2}{2})8^{\frac{n}{2}}}{(\sigma q)^{\frac{n}{2}}} \leq \frac{\Gamma(\frac{n+2}{2})8^{\frac{n}{2}}}{(\sigma_0 q)^{\frac{n}{2}}} = \sqrt{\frac{\sigma_0 q}{8\pi 3^{n-1}}}\,.$$

With this and (17), the inequality in (11) becomes

$$\|\Upsilon\|_{C_0\cap L^2} \leq \varphi_\sigma(0)(1+2^n)\max\left\{\sqrt{\frac{\sigma_0 q}{8\pi 3^{n-1}}},\, 1\right\}\,. \tag{19}$$

Using $n = 1$ in (15), we have $\sigma_0 q = 4\pi$ and $\sigma_0 q/(8\pi 3^{1-1}) = \frac{1}{2} < 1$. To treat $n \geq 2$, consider the expression below for $\Gamma(x)$ [23, p. 253], which holds for $x > 0$:

$$\Gamma(x) = x^{x-\frac{1}{2}}e^{-x}(2\pi)^{\frac{1}{2}}e^{\frac{\theta}{12x}},\ \text{where } 0 < \theta = \theta(x) < 1.$$

Let $x = \frac{n+2}{2}$. Since $n \geq 2$, we have $x \geq 2$ and $\frac{\theta}{12x} < \frac{1}{24} < \frac{1}{2}$. Consequently,

$$\sqrt{\frac{2\pi}{e}}\left(\frac{n+2}{2e}\right)^{\frac{n+1}{2}} < \Gamma\left(\frac{n+2}{2}\right) < \sqrt{2\pi}\left(\frac{n+2}{2e}\right)^{\frac{n+1}{2}}\,.$$

Coupling this with (15) gives us

$$12e^{-1}\left(\frac{\sqrt{2}\pi}{3\sqrt{e}}\right)^{\frac{2}{n+1}}(n+2) < \sigma_0 q < 12e^{-1}\left(\frac{\sqrt{2}\pi}{3}\right)^{\frac{2}{n+1}}(n+2). \tag{20}$$

Using a little calculus along with the upper bound in (20), it is easy to show that $\sigma_0 q/(8\pi 3^{n-1}) < 1$. Thus, we arrive at our final bound on $\|\Upsilon\|_{C_0\cap L^2}$,

$$\|\Upsilon\|_{C_0\cap L^2} \leq \varphi_\sigma(0)(1+2^n). \tag{21}$$

Employing the bound above in (18) yields

$$\|z^*|_{\mathcal{B}_\sigma}\|_{\mathcal{B}_\sigma^*} \geq \left(\frac{1}{2+2^{n+1}}\right)\|z^*\|_{C_0\cap L^{2^*}}\,. \tag{22}$$

Equivalently, $\|z^*\|_{C_0\cap L^{2^*}} \leq (2+2^{n+1})\|z^*|_{\mathcal{B}_\sigma}\|_{\mathcal{B}_\sigma^*}$. Applying Proposition 3.1 then completes the proof. □

Remark. It is interesting to note that the connection between σ_0 and q given in (15) is asymptotically linear in the dimension n. Indeed, from (20), we easily derive this asymptotic formula. As $n \to \infty$, $\sigma_0 \sim 12e^{-1}(n+2)q^{-1}$.

3.2. Jackson-type estimates for $\mathcal{B}$. The purpose of this section is to provide estimates on $\operatorname{dist}_{C_0\cap L^2}(f,\mathcal{B}_\sigma)$ when f has certain smoothness properties. Nikolskii [18, section 5.2] obtains many of the results that we need, albeit implicitly and, unfortunately, less than transparently. For the convenience of the reader, we will obtain the results that we need here, employing methods[1] considerably different from those in [18].

The approach we take is based on the Calderón decomposition formula [5, 10], which essentially states that if $f \in L^2(\mathbb{R}^n)$, then

$$f = \int_0^\infty \bar{\psi}_t * \psi_t * f \,\frac{dt}{t},\ \ \psi_t(x) := t^{-n}\psi\left(\frac{x}{t}\right), \tag{23}$$

where $\psi \in L^1$ is an arbitrary radial function with a (radial) Fourier transform that satisfies $\int_0^\infty |\hat{\psi}(t|\xi|)|^2 \frac{dt}{t} = 1$ if $\xi \in \mathbb{R}^n \setminus \{0\}$. The integral in (23), which is improper, is understood as an L^2 limit of $\int_\epsilon^T \bar{\psi}_t * \psi_t * f \frac{dt}{t}$ as $T \to \infty$ and $\epsilon \to 0^+$ independently [10, Theorem 1.2].

We are interested in approximating f with

$$g_\sigma := \int_{\frac{1}{\sigma}}^\infty \bar{\psi}_t * \psi_t * f \,\frac{dt}{t}, \tag{24}$$

where on ψ we make the additional assumptions that ψ is in $\mathcal{S}$ and that $\operatorname{supp}(\hat{\psi}) \subset B(0,1)$ and that $\int_0^\infty |\hat{\psi}(t|\xi|)|^2 \frac{dt}{t^{1+r}} < \infty$ for all $r \geq 0$ and all $\xi \in \mathbb{R}^n \setminus \{0\}$. The first assumption implies that $\psi \in \mathcal{B}_1$ and the second that all moments of ψ vanish. From

$$\hat{g}_\sigma(\xi) = \hat{f}(\xi)\int_{\frac{1}{\sigma}}^\infty |\hat{\psi}(t|\xi|)|^2 \frac{dt}{t} = \hat{f}(\xi)\begin{cases} 0 & \text{if } |\xi| \geq \sigma, \\ \int_{\frac{|\xi|}{\sigma}}^1 |\hat{\psi}(t)|^2 \frac{dt}{t} & \text{if } |\xi| < \sigma, \end{cases} \tag{25}$$

it follows that $\operatorname{supp}(\hat{g}_\sigma) \subset B(0,\sigma)$, and so $g_\sigma \in \mathcal{B}_\sigma$. Since $\int_0^\infty |\hat{\psi}(t|\xi|)|^2 \frac{dt}{t} = 1$, we also have

$$\hat{f} - \hat{g}_\sigma = \hat{f}(\xi)\begin{cases} 1 & \text{if } |\xi| \geq \sigma, \\ \int_0^{\frac{|\xi|}{\sigma}} |\hat{\psi}(t)|^2 \frac{dt}{t} & \text{if } |\xi| < \sigma. \end{cases} \tag{26}$$

This leads to useful Sobolev norm estimates.

PROPOSITION 3.6. *Let $r \geq 0$ and $s \geq 0$. If $f \in W_2^{s+r}(\mathbb{R}^n)$, then*

$$\|f - g_\sigma\|_{W_2^s} \leq c_r \sigma^{-r} \|f\|_{W_2^{r+s}}, \tag{27}$$

where $c_r := \int_0^1 |\hat{\psi}(t)|^2 \frac{dt}{t^{1+r}}$.

Proof. Note that

$$\int_0^{\frac{|\xi|}{\sigma}} |\hat{\psi}(t)|^2 \frac{dt}{t} = \int_0^{\frac{|\xi|}{\sigma}} t^r |\hat{\psi}(t)|^2 \frac{dt}{t^{1+r}} \leq \left(\frac{|\xi|}{\sigma}\right)^r \int_0^1 |\hat{\psi}(t)|^2 \frac{dt}{t^{1+r}}.$$

Since $c_r \geq c_0 = 1$, from (26) and the previous inequality we have

$$\begin{aligned}(1+|\xi|^2)^{\frac{s}{2}}|\hat{f} - \hat{g}_\sigma| &\leq c_r \sigma^{-r}(1+|\xi|^2)^{\frac{s}{2}}|\xi|^r |\hat{f}(\xi)| \\ &\leq c_r \sigma^{-r}(1+|\xi|^2)^{\frac{s+r}{2}}|\hat{f}(\xi)|.\end{aligned}$$

[1] The authors wish to thank Professor W. R. Madych for pointing these methods out to us.

Taking the L^2 norms of both sides above yields (27). □

We now require L^∞ bounds on derivatives $D^\alpha f - D^\alpha g_\sigma$, where α is a multi-index of nonnegative integers and $D^\alpha = (\frac{\partial}{\partial x_1})^{\alpha_1} \cdots (\frac{\partial}{\partial x_n})^{\alpha_n}$. These we will use in conjunction with our bounds above to estimate $\|f - g_\sigma\|_{C_0 \cap L^2}$ and, eventually, $\|f - f_\sigma\|_{C_0 \cap L^2}$, where f_σ both interpolates and approximates f.

Suppose that f is in $C_0^k(\mathbb{R}^n)$, where $k > 0$ is an integer. For x fixed, we can use Taylor's theorem with remainder to obtain

$$D^\alpha f(x - ty) = \sum_{|\beta| \le k-1-|\alpha|} \frac{(-t)^{|\beta|} y^\beta}{\beta!} D^{\alpha+\beta} f(x) + \sum_{|\beta| = k-|\alpha|} \frac{(-t)^{k-|\alpha|} y^\beta}{\beta!} D^{\alpha+\beta} f(\tilde{x}),$$

where $\tilde{x}$ is on the line between x and $x - ty$. Next, note that

$$\begin{aligned} \psi_t * D^\alpha f(x) &= \int_{\mathbb{R}^n} t^{-n} \psi(y/t) D^\alpha f(x - y) d^n y \\ &= \int_{\mathbb{R}^n} \psi(y) D^\alpha f(x - ty) d^n y. \end{aligned}$$

Inserting the expression for $D^\alpha f(x - ty)$ into the bottom integral and noting that all of the moments of ψ are 0, we see that

$$\psi_t * D^\alpha f(x) = \sum_{|\beta| = k-|\alpha|} \frac{(-t)^{k-|\alpha|}}{\beta!} \int_{\mathbb{R}^n} \psi(y) y^\beta D^{\alpha+\beta} f(\tilde{x}) d^n y.$$

Taking absolute values and using the boundedness of the derivatives, we have

$$\begin{aligned} \|\psi_t * D^\alpha f\|_{L^\infty} &\le t^{k-|\alpha|} \big\| |y|^{k-|\alpha|} \psi \big\|_{L^1} \Big(\sum_{|\beta| = k-|\alpha|} \frac{\|D^{\alpha+\beta} f\|_{L^\infty}}{\beta!} \Big) \\ &\le t^{k-|\alpha|} \big\| |y|^{k-|\alpha|} \psi \big\|_{L^1} \|D^\alpha f\|_{C_0^{k-|\alpha|}} \\ &\le t^{k-|\alpha|} \big\| |y|^{k-|\alpha|} \psi \big\|_{L^1} \|f\|_{C_0^k}. \end{aligned}$$

From this, the fact that $\|\bar{\psi}_t\|_{L^1} = \|\psi\|_{L^1}$, and Young's inequality, we see that

$$\|\bar{\psi}_t * \psi_t * D^\alpha f\|_{L^\infty} \le t^{k-|\alpha|} \|\psi\|_{L^1} \big\| |y|^{k-|\alpha|} \psi \big\|_{L^1} \|f\|_{C_0^k}. \tag{28}$$

Next, recall that from (23) and (24) the difference $D^\alpha f - D^\alpha g_\sigma$ is

$$D^\alpha f - D^\alpha g_\sigma = \int_0^{\frac{1}{\sigma}} \bar{\psi}_t * \psi_t * D^\alpha f \, \frac{dt}{t}. \tag{29}$$

Taking the L^∞ norm of $D^\alpha f - D^\alpha g_\sigma$, using the bound in (28) above, and doing the simple integral involved, we have proven the following result.

PROPOSITION 3.7. *Let $k > 0$ be an integer, and let α be a multi-index satisfying $k > |\alpha|$. If $f \in C_0^k \cap W_2^k$, then*

$$\|D^\alpha f - D^\alpha g_\sigma\|_{L^\infty} \le \sigma^{|\alpha|-k} c'_{k-|\alpha|} \|f\|_{C_0^k},$$

where $c'_{k-|\alpha|} = (k - |\alpha|)^{-1} \big\| |y|^{k-|\alpha|} \psi \big\|_{L^1} \|\psi\|_{L^1}$.

Combining Propositions 3.6 and 3.7 immediately yields these distance estimates.

THEOREM 3.8. *Let $k > 0$ be an integer, and let $\sigma > 0$. If $f \in C_0^k \cap W_2^k$, then there is a constant $C = C(k, n)$ such that*

$$\operatorname{dist}_{C_0 \cap L^2}(f, \mathcal{B}_\sigma) \le C\sigma^{-k} \|f\|_{C_0^k \cap W_2^k}.$$

Proof. From Proposition 3.6, with $s = 0$ and $r = k$, we have $\|f - g_\sigma\|_{L^2} \le c_k \sigma^{-k} \|f\|_{W_2^k}$. In addition, from Proposition 3.7, with $\alpha = 0$, we have that $\|f - g_\sigma\|_{C_0} \le c_k' \sigma^{-k} \|f\|_{C_0^k}$. Consequently, we obtain

$$\begin{aligned} \|f - g_\sigma\|_{C_0 \cap L^2} &\le \max(c_k, c_k') \sigma^{-k} \max\left(\|f\|_{C_0^k}, \|f\|_{W_2^k} \right) \\ &\le C\sigma^{-k} \|f\|_{C_0^k \cap W_2^k}, \end{aligned}$$

where $C = \max(c_k, c_k')$. Since $\operatorname{dist}_{C_0 \cap L^2}(f, \mathcal{B}_\sigma) \le \|f - g_\sigma\|_{C_0 \cap L^2}$, the result follows immediately. ∎

3.3. Error estimates for $\mathcal{B}$ interpolants. We will conclude our discussion of band-limited functions by proving error estimates for the interpolant f_σ for $f \in C_0 \cap L^2$, whose existence was shown in Theorem 3.5. The first is an immediate corollary to Theorems 3.5 and 3.8.

COROLLARY 3.9. *Let $f \in C_0^k \cap W_2^k$, and let f_σ be as in Theorem* 3.5. *Then*

$$\|f - f_\sigma\|_{C_0 \cap L^2} \le C\sigma^{-k} \|f\|_{C_0^k \cap W_2^k},$$

where $C = C(k, n)$.

We turn to obtaining estimates on $\|D^\alpha f - D^\alpha f_\sigma\|_{L^\infty}$. To do this, we will make use of Bernstein's theorem for functions of exponential type [18, section 3.2.2, eq. (8)]: If $h_\sigma \in \mathcal{B}_\sigma$, then

$$\|D^\alpha h_\sigma\|_{L^p} \le \sigma^{|\alpha|} \|h_\sigma\|_{L^p}, \quad 1 \le p \le \infty. \tag{30}$$

Our result is the following theorem.

THEOREM 3.10. *Let $k > 0$ be an integer and α be a multi-index with $|\alpha| < k$, $\sigma > 0$, and $f \in C_0^k \cap W_2^k$. If $f_\sigma \in \mathcal{B}_\sigma$ is the interpolant to f from Theorem* 3.5, *then there is a constant $C = C(|\alpha|, k, n)$ for which*

$$\|D^\alpha f - D^\alpha f_\sigma\|_{L^\infty} \le C\sigma^{|\alpha|-k} \|f\|_{C_0^k \cap W_2^k}. \tag{31}$$

Proof. Let g_σ be defined by (24). Then we have

$$\|D^\alpha f - D^\alpha f_\sigma\|_{L^\infty} \le \|D^\alpha f - D^\alpha g_\sigma\|_{L^\infty} + \|D^\alpha g_\sigma - D^\alpha f_\sigma\|_{L^\infty}.$$

By Proposition 3.7, we can bound the first term on the right-hand side by $\sigma^{|\alpha|-k} C_1 \|f\|_{C_0^k}$. Using Bernstein's inequality (30) for $p = \infty$, with $h_\sigma = g_\sigma - f_\sigma$, we may bound the second term by $\sigma^{|\alpha|} \|g_\sigma - f_\sigma\|_{L^\infty}$. Putting these two together yields

$$\|D^\alpha f - D^\alpha f_\sigma\|_{L^\infty} \le \sigma^{|\alpha|-k} C_1 \|f\|_{C_0^k} + \sigma^{|\alpha|} \|g_\sigma - f_\sigma\|_{L^\infty}.$$

Next, observe that $\|g_\sigma - f_\sigma\|_{L^\infty} \le \|f - f_\sigma\|_{L^\infty} + \|g_\sigma - f\|_{L^\infty}$. By Proposition 3.7 and Corollary 3.9, we have

$$\begin{aligned} \|g_\sigma - f_\sigma\|_{L^\infty} &\le C_2 \sigma^{-k} \|f\|_{C_0^k \cap W_2^k} + C_3 \sigma^{-k} \|f\|_{C_0^k} \\ &\le (C_2 + C_3)\, \sigma^{-k} \|f\|_{C_0^k \cap W_2^k}. \end{aligned}$$

Finally, inserting this bound on the right-hand side in the previous one gives us (31), with $C = C_1 + C_2 + C_3$. ∎

It is easy to obtain bounds in W_2^r similar to the ones above. We do not need them, however, except in the following case.

PROPOSITION 3.11. *Let $f \in C_0^k \cap W_2^k$, and let f_σ be as in Theorem* 3.5. *Assume that $\sigma \geq 1$. Then there is a constant $C' = C'(k,n)$ such that*

$$\|f_\sigma\|_{W_2^k} \leq C' \|f\|_{C_0^k \cap W_2^k} . \tag{32}$$

Proof. To bound $\|f_\sigma\|_{W_2^k}$, we begin with the inequality

$$\|f_\sigma\|_{W_2^k} \leq \|f_\sigma - g_\sigma\|_{W_2^k} + \|f - g_\sigma\|_{W_2^k} + \|f\|_{W_2^k}.$$

Proposition 3.6, with $s = k$ and $r = 0$, gives us $\|f_\sigma - g_\sigma\|_{W_2^k} \leq c_0 \|f\|_{W_2^k}$. In addition, Bernstein's inequality (30) in conjunction with the definition of the Sobolev norm for W_2^k yields

$$\|f_\sigma - g_\sigma\|_{W_2^k} \leq (1 + \sigma^k) \|f_\sigma - g_\sigma\|_{L^2}.$$

Thus, we have that

$$\|f_\sigma\|_{W_2^k} \leq (1 + \sigma^k) \|f_\sigma - g_\sigma\|_{L^2} + (c_0 + 1) \|f\|_{W_2^k} .$$

Next, by Corollary 3.9 and Proposition 3.6, we also have

$$\begin{aligned} \|f_\sigma - g_\sigma\|_{L^2} &\leq \|f_\sigma - f\|_{C_0 \cap L^2} + \|f - g_\sigma\|_{L^2} \\ &\leq C\sigma^{-k} \|f\|_{C_0^k \cap W_2^k} + c_k \sigma^{-k} \|f\|_{W_2^k} \\ &\leq (C + c_k) \sigma^{-k} \|f\|_{C_0^k \cap W_2^k} . \end{aligned}$$

Combining this with the inequality previous to it yields

$$\|f_\sigma\|_{W_2^k} \leq \left((1 + \sigma^{-k})(C + c_k) + (c_0 + 1) \right) \|f\|_{C_0^k \cap W_2^k} \leq C' \|f\|_{C_0^k \cap W_2^k} ,$$

which completes the proof. ∎

4. RBFs. Let m be a nonnegative integer, and let $\Phi(x) = \Phi(|x|)$ be continuous. We say Φ is an order $m \geq 0$ RBF if for every subset $X = \{x_1, \ldots, x_n\}$ comprising distinct points in $\mathbb{R}^n$ and for every $c \in \mathbb{C}^N \setminus \{0\}$ satisfying $\sum_{j=1}^N c_j p(x_j) = 0$ for all $p \in \pi_{m-1}(\mathbb{R}^n)$ we have that

$$c^H A c = \sum_{j,k=1}^N \bar{c}_j c_k \Phi(x_k - x_j) > 0.$$

That is, the function Φ is *strictly* conditionally positive definite of order m. Interpolants are formed from an order m RBF Φ in the following way. When X is a unisolvent set for $\pi_{m-1}(\mathbb{R}^n)$ and data is generated by a continuous function f, then there is a unique $c \in \mathbb{C}^N$ satisfying $\sum_{j=1}^N c_j p(x_j) = 0$ for every $p \in \pi_{m-1}(\mathbb{R}^n)$ and a unique $q \in \pi_{m-1}(\mathbb{R}^n)$ such that

$$I_X f(x) = \sum_{x_j \in X} c_j \Phi(x - x_j) + q(x)$$

satisfies $I_X f|_X = f|_X$. Moreover, if f is a polynomial in $\pi_{m-1}(\mathbb{R}^n)$, then $I_X f = q$. That is, the method reproduces polynomials of degree less than m.

4.1. RKHS and RBFs. An order m RBF Φ has $\mathcal{O}(|x|^{2m})$ growth as $|x| \to \infty$ [13, Corollary 2.3]. Consequently, we may view it as being a tempered distribution. As such, it has a (radial) Fourier transform $\hat{\Phi} \in \mathcal{S}'$. In all cases of practical interest, $\hat{\Phi}(\xi)$ in $\mathbb{R}^n$ is a positive, continuous function for $\xi \in \mathbb{R}^n \setminus \{0\}$ that may have a singularity of the form $|\xi|^{-\tau}$ as $|\xi| \to 0$. With these RBFs, we can associate a RKHS, the native space of Φ,

$$\mathcal{N}_\Phi := \left\{ f \in L^2(\mathbb{R}^n) \colon \|f\|_\Phi^2 := \int_{\mathbb{R}^n} |\hat{f}(\xi)|^2 \widehat{\Phi}(\xi)^{-1} d^n\xi < \infty \right\}. \tag{33}$$

We will mention two important classes of RBFs. Duchon's thin-plate splines

$$\Phi_\nu^{\mathrm{TPS}}(x) = \begin{cases} (-1)^{\lfloor \nu \rfloor + 1} |x|^{2\nu}, & \nu > 0,\ \nu \notin \mathbb{N}, \\ (-1)^{\nu+1} |x|^{2\nu} \log(|x|), & \nu \in \mathbb{N}, \end{cases}$$

are of order $m_\nu = \lfloor \nu \rfloor + 1$ and have distributional Fourier transforms (in $\mathbb{R}^n$) given by [20, Table 1]

$$\widehat{\Phi}_\nu^{\mathrm{TPS}}(\xi) = C_{\nu,n} |\xi|^{-n-2\nu}.$$

Wendland's compactly supported RBFs [21, 22] also display similar behavior in their Fourier transforms. The functions themselves are all (order 0) RBFs but only on $\mathbb{R}^d$, $d \le n$. Each has the form

$$\Phi_{n,k}^{\mathrm{WEN}}(x) = \begin{cases} p_{n,k}(|x|), & 0 \le |x| \le 1, \\ 0, & |x| > 1, \end{cases}$$

where $p_{n,k}$ is a univariate polynomial of degree $\lfloor \frac{n}{2} \rfloor + 3k + 1$; also, $\Phi_{n,k}$ is in $C^{2k}(\mathbb{R}^n)$. Their Fourier transforms satisfy the bounds [22, Theorem 2.1]

$$c_{n,k}(1 + |\xi|^2)^{-\frac{n}{2}-k-\frac{1}{2}} \le \widehat{\Phi}_{n,k}^{\mathrm{WEN}}(\xi) \le C_{n,k}(1 + |\xi|^2)^{-\frac{n}{2}-k-\frac{1}{2}}.$$

Error estimates on $I_X f$ and $D^\alpha I_X f$, where α is a standard multi-index, are known in the case where f belongs to a native space $\mathcal{N}_\Phi$ [13, 24] stemming from an order m RBF.

THEOREM 4.1. *Let α be a multi-index, $r, s \in \mathbb{R}$, with $\frac{n}{2} + |\alpha| < r$ and $s + \frac{n}{2} < m$, and suppose that $\widehat{\Phi}(\xi)$ is positive and continuous on $\xi \in \mathbb{R}^n \setminus \{0\}$ and satisfies*

$$\widehat{\Phi}(\xi)^{-1} = \mathcal{O}(|\xi|^{2s}) \text{ as } |\xi| \to 0 \quad \text{and} \quad \widehat{\Phi}(\xi)^{-1} = \mathcal{O}(|\xi|^{2r}) \text{ as } |\xi| \to \infty. \tag{34}$$

If $\Omega \supset X$ is a compact region satisfying a uniform interior cone condition and if $f \in \mathcal{N}_\Phi$, then

$$\|D^\alpha f - D^\alpha I_X f\|_{L^\infty(\Omega)} \le C h^{r - \frac{n}{2} - |\alpha|} \|f\|_\Phi, \quad C = C(|\alpha|, \Omega, \Phi), \tag{35}$$

where $h = h_{X,\Omega}$ is the mesh norm.

Various versions of these estimates are found in Madych and Nelson [13, Theorem 4.4] and in Wu and Schaback [24, Theorems 4.5 and 5.14]. A cone condition on Ω is alluded to in [13] and explicitly incorporated in the estimates in [19, p. 333].

4.2. Extended error estimates for RBF interpolants. We are now ready to obtain the RBF interpolation error estimates discussed in section 1. Let $f \in C_0^k \cap W_2^k$, σ, and f_σ be as in Theorem 3.5, and let α be a multi-index with $\frac{n}{2} + |\alpha| < k$. We will assume that Ω satisfies a uniform interior cone condition and that $\mathrm{diam}(\Omega) \leq 1$. Of course, $\Omega \supset X$ implies that $\mathrm{diam}(X) \leq 1$. Next, we will take $\sigma = \sigma_0$ in (15). This choice of σ implies two things: first, since $q \leq \mathrm{diam}(X) \leq 1$, $\sigma > 1$; and, second, σh has the form

$$\sigma h = \gamma_n \rho, \text{ where } \rho := \frac{h}{q} \text{ and } \gamma_n := 24 \left\{ \frac{\sqrt{\pi}}{3} \Gamma\left(\frac{n+2}{2}\right) \right\}^{\frac{2}{n+1}}. \tag{36}$$

Since the ratio $\rho = h/q \geq 1$, we also see that the product $\sigma h > 1$ and that $\sigma > 1$.

The idea for obtaining new bounds is to write the difference $D^\alpha f - D^\alpha I_X f$ as the sum of three pieces:

$$D^\alpha f - D^\alpha I_X f = (D^\alpha f - D^\alpha f_\sigma) + (D^\alpha f_\sigma - D^\alpha I_X f_\sigma) + D^\alpha (I_X f_\sigma - I_X f).$$

Now, $(f_\sigma - f)|_X = 0$, and so the uniqueness of the RBF interpolant implies that $I_X f_\sigma - I_X f \equiv 0$. The third term above is thus 0. Our error then becomes

$$\|D^\alpha f - D^\alpha I_X f\|_{L^\infty(\Omega)} \leq \|D^\alpha f - D^\alpha f_\sigma\|_{L^\infty(\Omega)} + \|D^\alpha f_\sigma - D^\alpha I_X f_\sigma\|_{L^\infty(\Omega)}.$$

We can replace the first term on the right-hand side above by its bound from Theorem 3.10 and the second term by its RKHS bound from (35), since Ω satisfies the requisite cone condition. Doing so yields

$$\|D^\alpha f - D^\alpha I_X f\|_{L^\infty(\Omega)} \leq C_1 \sigma^{|\alpha|-k} \|f\|_{C_0^k \cap W_2^k} + C_2 h^{r-\frac{n}{2}-|\alpha|} \|f_\sigma\|_\Phi, \tag{37}$$

provided $\widehat{\Phi}$ satisfies the conditions in Theorem 4.1. The norm $\|\cdot\|_\Phi$ is given in (33). Under these assumptions on Φ, plus the additional assumption that $k \leq r$, the norm $\|f_\sigma\|_\Phi$ can be estimated as follows:

$$\begin{aligned}
\|f_\sigma\|_\Phi^2 &= \int_{|\xi| \leq \sigma} \frac{|\hat{f}_\sigma|^2}{\widehat{\Phi}} d^n\xi \\
&\leq C_3^2 \int_{|\xi| \leq \sigma} (1 + |\xi|^{2r}) |\hat{f}_\sigma|^2 d^n\xi \\
&\leq C_3^2 \int_{|\xi| \leq \sigma} \frac{1 + |\xi|^{2r}}{(1 + |\xi|^2)^k} (1 + |\xi|^2)^k |\hat{f}_\sigma|^2 d^n\xi \\
&\leq C_3^2 \sigma^{2r-2k} \|f_\sigma\|_{W_2^k}^2.
\end{aligned}$$

Taking square roots and employing Proposition 3.11, which holds since $\sigma > 1$, then gives us

$$\|f_\sigma\|_\Phi \leq C_4 \sigma^{r-k} \|f\|_{C_0^k \cap W_2^k}.$$

Inserting this in (37) yields

$$\|D^\alpha f - D^\alpha I_X f\|_{L^\infty(\Omega)} \leq \left(C_1 \sigma^{|\alpha|-k} + C_5 h^{r-\frac{n}{2}-|\alpha|} \sigma^{r-k} \right) \|f\|_{C_0^k \cap W_2^k}.$$

Factoring $h^{k-|\alpha|-\frac{n}{2}}$ from the right-hand side and manipulating the resulting expression, we have

$$\|D^\alpha f - D^\alpha I_X f\|_{L^\infty(\Omega)} \leq h^{k-|\alpha|-\frac{n}{2}} \left(C_1 (\sigma h)^{|\alpha|-k} h^{\frac{n}{2}} + C_5 (\sigma h)^{r-k} \right) \|f\|_{C_0^k \cap W_2^k}.$$

From $|\alpha| + \frac{n}{2} < k \le r$, $h \le 1$, and $\sigma h > 1$, it follows that $(\sigma h)^{|\alpha|-k} h^{\frac{n}{2}} < (\sigma h)^{r-k}$, and so the coefficient on the right-hand side above is less than a multiple of $h^{k-|\alpha|-\frac{n}{2}}(\sigma h)^{r-k}$. From (36), $\sigma h = \gamma_n \rho$, and the final bound is $Ch^{k-|\alpha|-\frac{n}{2}}\rho^{r-k}$. In summary, we have obtained this result.

THEOREM 4.2. *Let the notation and assumptions of Theorem* 4.1 *hold. In addition, suppose that* $\operatorname{diam}(\Omega) \le 1$. *If* k *is an integer satisfying* $|\alpha| + \frac{n}{2} < k \le r$ *and if* $f \in C_0^k \cap W_2^k$, *then there is a constant* $C = C(k, |\alpha|, n, \Omega, \Phi)$ *such that*

$$\|D^\alpha f - D^\alpha I_X f\|_{L^\infty(\Omega)} \le Ch^{k-|\alpha|-\frac{n}{2}}\rho^{r-k}\|f\|_{C_0^k \cap W_2^k}$$

holds. Here, $\rho = \rho_{X,\Omega}$ *is the mesh ratio, and* $h = h_{X,\Omega}$ *is the mesh norm.*

Note that if we require a uniform bound on the mesh ratio, we immediately get uniform bounds on the error.

COROLLARY 4.3. *For any set of centers* $X \subset \Omega$ *for which* $\rho_{X,\Omega} \le R$, *with* R *fixed, we have*

$$\frac{\|D^\alpha f - D^\alpha I_X f\|_{L^\infty(\Omega)}}{\|f\|_{C_0^k \cap W_2^k}} = \mathcal{O}(h^{k-|\alpha|-\frac{n}{2}}).$$

We have assumed throughout the paper that f is in $C_0 \cap L^2$, and so it is defined and, at the very least, continuous on all of $\mathbb{R}^n$. In many applications, for example solving a partial differential equation, this is not the case. The function f is then only defined on Ω. To deal with this situation, one can appeal to one of the many extension theorems available (cf. [1, Chapter IV] or [11, section 6.9]) and again have a function defined on $\mathbb{R}^n$. This will usually require some additional assumptions concerning the smoothness of the boundary.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] A. ALDROUBI AND K. GRÖCHENIG, *Nonuniform sampling and reconstruction in shift-invariant spaces*, SIAM Rev., 43 (2001), pp. 585–620.

[3] R. K. BEATSON, J. B. CHERRIE, AND D. L. RAGOZIN, *Fast evaluation of radial basis functions: Methods for four-dimensional polyharmonic splines*, SIAM J. Math. Anal. 32 (2001), pp. 1272–1310.

[4] C. DE BOOR, R. DEVORE, AND A. RON, *Approximation from shift-invariant subspaces of* $L_2(R^d)$, Trans. Amer. Math. Soc., 341 (1994), pp. 787–806.

[5] A. P. CALDERÓN, *Intermediate spaces and interpolation, the complex method*, Studia Math., 24 (1964), pp. 113–190.

[6] J. C. CARR, R. K. BEATSON, J. B. CHERRÌE, T. J. MITCHELL, W. R. FRIGHT, B. C. MCCALLUM, AND T. R. EVANS, *Reconstruction and representation of* 3*D objects with radial basis functions*, in Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, Los Angeles, CA, 2001, ACM, New York, 2001, pp. 67–76.

[7] J. DUCHON, *Splines minimizing rotation invariant semi-norms in Sobolev spaces*, in Constructive Theory of Functions of Several Variables, Lecture Notes in Math. 571, W. Schempp and K. Zeller, eds., Springer, Berlin, 1977, pp. 85–100.

[8] J. DUCHON, *Sur l'erreur d'interpolation des fonctions de plusieurs variables par les* D^m*-splines*. RAIRO Anal. Numér., 12 (1978), pp. 325–334.

[9] P. ERDŐS, *On some convergence properties of the interpolation polynomials*, Ann. of Math. (2), 44 (1943), pp. 330–337.

[10] M. FRAZIER, B. JAWERTH, AND G. WEISS, *Littlewood-Paley Theory and the Study of Function Spaces*, CBMS Reg. Conf. Ser. Math. 79, AMS, Providence, RI, 1991.

[11] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, New York, 1977.

[12] W. B. JOHNSON AND J. LINDENSTRAUSS, *Handbook of the Geometry of Banach Spaces,* Vol. I, Elsevier, Amsterdam, 2001.

[13] W. R. Madych and S. A. Nelson, *Multivariate interpolation and conditionally positive definite functions.* II, Math. Comp., 54 (1990), pp. 211–230.
[14] H. N. Mhaskar, F. J. Narcowich, N. Sivakumar, and J. D. Ward, *Approximation with interpolatory constraints*, Proc. Amer. Math. Soc., 130 (2002), pp. 1355–1364.
[15] F. J. Narcowich and J. D. Ward, *Norms of inverses and condition numbers for matrices associated with scattered data*, J. Approx. Theory, 64 (1991), pp. 69–94.
[16] F. J. Narcowich and J. D. Ward, *Norm Estimates for the Inverses of a General Class of Scattered-Data Radial-Function Interpolation Matrices*, J. Approx. Theory, 69 (1992), pp. 84–109.
[17] F. J. Narcowich and J. D. Ward, *Scattered data interpolation on spheres: Error estimates and locally supported basis functions*, SIAM J. Math. Anal., 33 (2002), pp. 1393–1410.
[18] S. M. Nikolskii, *Approximation of Functions of Several Variables and Imbedding Theorems*, Springer-Verlag, New York, 1975.
[19] R. Schaback, *Approximation by radial basis functions with finitely many centers*, Constr. Approx., 12 (1996), pp. 331–340.
[20] R. Schaback, *Improved error bounds for scattered data interpolation by radial basis functions*, Math. Comp., 68 (1999), pp. 201–216.
[21] H. Wendland, *Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree*, Adv. Comput. Math., 4 (1995), pp. 389–396.
[22] H. Wendland, *Error estimates for interpolation by compactly supported radial basis functions of minimal degree*, J. Approx. Theory, 93 (1998), pp. 258–272.
[23] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, 4th ed., Cambridge University Press, London, 1965.
[24] Z. Wu and R. Schaback, *Local error estimates for radial basis function interpolation of scattered data*, IMA J. Numer. Anal., 13 (1993), pp. 13–27.
[25] J. Yoon, *Spectral approximation orders of radial basis function interpolation on the Sobolev space*, SIAM J. Math. Anal., 33 (2001), pp. 946–958.

SIAM J. MATH. ANAL.
Vol. 36, No. 1, pp. 301–322

ANALYSIS OF A MULTIDIMENSIONAL PARABOLIC POPULATION MODEL WITH STRONG CROSS-DIFFUSION*

LI CHEN† AND ANSGAR JÜNGEL‡

Abstract. The global existence of a nonnegative weak solution to a multidimensional parabolic strongly coupled model for two competing species is proved. The main feature of the model is that the diffusion matrix is nonsymmetric and generally not positive definite and that the nondiagonal matrix elements (the cross-diffusion terms) are allowed to be "large." The ideas of the existence proof are a careful approximation of the cross-diffusion terms using finite differences and the use of an entropy inequality yielding a priori estimates.

Key words. cross-diffusion system, entropy functional, existence of weak solutions, Orlicz space

AMS subject classifications. 35K55, 35D05, 92D25

DOI. 10.1137/S0036141003427798

1. Introduction.

For the time evolution of two competing species with homogeneous population density, usually the Lotka–Volterra differential equations are used as an appropriate mathematical model. In the case of nonhomogeneous densities, diffusion effects have to be taken into account leading to reaction-diffusion equations. Shigesada, Kawasaki, and Teramoto proposed in their pioneering work [25] to introduce further so-called cross-diffusion terms modeling segregation phenomena of the competing species. Denoting by $u_i(x,t)$ the population density of the ith species and by $J_i(x,t)$ the corresponding population flows, the time-dependent equations can be written as

$$\partial_t u_i - \mathrm{div} J_i = f_i(u_1,u_2), \quad J_i = \nabla(c_i u_i + a_i u_i^2 + u_1 u_2) + d_i u_i q, \tag{1.1}$$

where $i = 1,2$. The equations are solved in the bounded domain $\Omega \subset \mathbb{R}^N$ $(N \le 3)$ with time $t > 0$. The function q is given by $q = \nabla U$, and $U = U(x,t)$ is a prescribed environmental potential, modeling areas where the environmental conditions are more or less favorable [20, 25]. The diffusion coefficients c_i and a_i are nonnegative, and $d_i \in \mathbb{R}$ $(i = 1,2)$. The source terms are in Lotka–Volterra form:

$$f_i(u_1,u_2) = (R_i - \beta_{i1}u_1 - \beta_{i2}u_2)u_i, \quad i = 1,2, \tag{1.2}$$

where $R_i \ge 0$ is the intrinsic growth rate of the ith species, $\beta_{ii} > 0$ are the coefficients of intraspecific competition, and $\beta_{12} \ge 0$ and $\beta_{21} \ge 0$ are those of interspecific competition. The above system of equations is supplemented with (biologically motivated) homogeneous Neumann boundary conditions and initial conditions:

$$J_i \cdot \gamma = 0 \quad \text{on } \partial\Omega \times (0,\infty), \tag{1.3}$$

$$u_i(\cdot,0) = u_i^0 \quad \text{in } \Omega, \quad i = 1,2, \tag{1.4}$$

*Received by the editors May 12, 2003; accepted for publication (in revised form) December 12, 2003; published electronically July 14, 2004. The authors were partially supported from the IHP Project "Hyperbolic and Kinetic Equations" of the European Union, grant HPRN-CT-2002-00282, and from the Gerhard–Hess Award of the Deutsche Forschungsgemeinschaft, grant JU 359/3.
http://www.siam.org/journals/sima/36-1/42779.html

†Department of Mathematical Sciences, Tsinghua University, Beijing, 100084, People's Republic of China (lchen@math.tsinghua.edu.cn).

‡Fachbereich Mathematik und Informatik, Universität Mainz, Staudingerweg 9, 55099 Mainz, Germany (juengel@mathematik.uni-mainz.de).

and γ denotes the exterior unit normal to $\partial\Omega$, which is assumed to exist almost everywhere.

Notice that the above system is scaled in such a way that the coefficient of the *cross-diffusion term* $\nabla(u_1u_2)$ is equal to one (see [8] for details).

The problem (1.1)–(1.4) is strongly coupled with full diffusion matrix

$$A(u_1,u_2)=\begin{pmatrix} c_1+2a_1u_1+u_2 & u_1 \\ u_2 & c_2+2a_2u_2+u_1 \end{pmatrix}.$$

Nonlinear problems of this kind are quite difficult to deal with since the usual idea of applying maximum principle arguments to get a priori estimates cannot be used here. Furthermore, the diffusion matrix is not symmetric and of degenerate type if $c_1=c_2=0$.

Until now, only partial results were available in the literature concerning the well posedness of the above problem. We summarize some of the available results for the time-dependent equations (see [28] for a review) and refer the reader to [16, 17, 23, 24] for the stationary problem. Global existence of solutions and their qualitative behavior for $a_1=a_2=0$ and no cross-diffusion for the second species have been proved in, e.g., [3, 18, 21, 22, 27]. In this case, (1.1) for $i=2$ is only weakly coupled. The existence of an attractor has been studied in [15, 22]. Notice that in chemotaxis, related models appear [7, 9, 19].

For sufficiently small cross-diffusion terms (or "small" initial data) and vanishing self-diffusion coefficients $a_1=a_2=0$, Deuring proved the global existence of solutions in [6]. For the case $c_1=c_2$, a global existence result in one space dimension has been obtained by Kim [12]. Furthermore, under the condition

$$2a_1>1, \quad 2a_2>1, \tag{1.5}$$

Yagi [29] has shown the global existence of solutions in two space dimensions. A global existence result for weak solutions in any space dimension under assumption (1.5) can be found in [8]. Condition (1.5) can be easily understood by observing that in this case, the diffusion matrix is positive definite:

$$\xi^TA(u_1,u_2)\xi\ge\min\{c_1,c_2\}|\xi|^2 \quad \text{for all } \xi\in\mathbb{R}^2,$$

hence yielding an elliptic operator. If the condition (1.5) does not hold, there are choices of c_i, a_i, $u_i\ge 0$ for which the matrix $A(u_1,u_2)$ is *not* positive definite. Finally, Galiano, Garzòn, and Jüngel [8] proved the existence of global weak solutions for *any* a_1, $a_2>0$. However, the proof uses the embedding $H^1(\Omega)\subset L^\infty(\Omega)$ in a crucial way such that the result is restricted to one space dimension only.

In this paper we solve the problem (1.1)–(1.4) for (up to) *three* space dimensions without any restriction on the diffusion coefficients. More precisely, we prove the following result.

THEOREM 1.1. *Let $T>0$, and assume that*

- *$\Omega\subset\mathbb{R}^N$ $(N\le 3)$ is a bounded domain with boundary $\partial\Omega\in C^{0,1}$;*
- *the parameters satisfy $c_i\ge 0$, $a_i>0$; $R_i\ge 0$, $\beta_{ii}>0$ $(i=1,2)$, $\beta_{12}=\beta_{21}\ge 0$; $q\in(L^2(Q_T))^N$, where $Q_T=\Omega\times(0,T)$;*
- *the initial data satisfy $u_i^0\in L_\Psi(\Omega)$ and $u_i^0\ge 0$ in Q_T $(i=1,2)$.*

Then problem (1.1)–(1.4) has a weak solution (u_1,u_2) satisfying $u_i\ge 0$ in Q_T and

$$u_i\in L^2(0,T;H^1(\Omega))\cap L^\infty(0,T;L_\Psi(\Omega))\cap W^{1,r}(0,T;(W^{1,r'}(\Omega))'), \quad i=1,2,$$

where $r = (2N+2)/(2N+1)$ *and* $r' = r/(r-1) = 2N+2$, *in the sense that for all* $\varphi \in L^{r'}(0,T;W^{1,r'}(\Omega))$, $i = 1,2$,

$$\int_0^T \langle \partial_t u_i, \varphi\rangle\, dt + \int_{Q_T} (c_i\nabla u_i + 2a_i u_i \nabla u_i + \nabla(u_1u_2) + d_i u_i q)\cdot\nabla\varphi dxdt = \int_{Q_T} f_i(u_1,u_2)\varphi dxdt,$$

and $\langle\cdot,\cdot\rangle$ *denotes the dual product between* $W^{1,r'}(\Omega)$ *and its dual* $(W^{1,r'}(\Omega))'$.

Here, $L_\Psi(\Omega)$ denotes the Orlicz space for $\Psi(s) = (1+s)\ln(1+s) - s$, $s \geq 0$. Orlicz space techniques for a related parabolic system have already been employed in [13]. We refer the reader to the appendix for its definition and some properties.

In order to explain the method of our proof it is convenient to recall the ideas of [8]. By using the exponential transformation of variables $u_1 = \exp(w_1)$, $u_2 = \exp(w_2)$, (1.1) transform into

$$\partial_t \begin{pmatrix} e^{w_1} \\ e^{w_2}\end{pmatrix} - \operatorname{div}\left(B(w_1,w_2)\nabla\begin{pmatrix} w_1\\ w_2\end{pmatrix} + \begin{pmatrix} d_1e^{w_1}\\ d_2e^{w_2}\end{pmatrix} q\right) = \begin{pmatrix} f_1\\ f_2\end{pmatrix},$$

and the new diffusion matrix

$$B(w_1,w_2) = \begin{pmatrix} c_1e^{w_1} + 2a_1e^{2w_1} + e^{w_1+w_2} & e^{w_1+w_2} \\ e^{w_1+w_2} & c_2e^{w_2} + 2a_2e^{2w_2} + e^{w_1+w_2}\end{pmatrix}$$

is symmetric and positive definite:

$$\det B(w_1,w_2) \geq (c_1e^{w_1} + 2a_1e^{2w_1})(c_2e^{w_2} + 2a_2e^{2w_2}) > 0.$$

In this formulation the matrix B provides an elliptic operator for all $c_i > 0$, $a_i \geq 0$ or $c_i \geq 0$, $a_i > 0$ $(i = 1,2)$. In this sense, the system (1.1)–(1.2) is called *parabolic*. We remark that exponential transformations of variables have also been used in other applications, such as chemotaxis [19] and semiconductor modeling [10].

The above change of unknowns symmetrizing the problem implies the existence of an *entropy functional*

$$E(t) = \sum_{i=1}^2 \int_\Omega (u_i(\ln u_i - 1) + 1)dx \geq 0,$$

with the corresponding entropy inequality

$$E(t) + 2\int_0^t \int_\Omega \Big(\sum_{i=1}^2 (2c_i|\nabla\sqrt{u_i}|^2 + a_i|\nabla u_i|^2) + 2|\nabla\sqrt{u_1u_2}|^2\Big)dxdt \leq E(0) + C \tag{1.6}$$

for $0 < t < T$ and any $T > 0$, where the constant $C > 0$ depends on T, q, and the source terms. It can be formally derived by using $\ln u_i$ as a test function in the weak formulation of (1.1)–(1.4). This inequality provides an $L^2(0,T;H^1(\Omega))$ estimate for u_1 and u_2 if a_1, $a_2 > 0$. The existence of a symmetric formulation of the problem is even equivalent to the existence of an entropy functional [5, 11]. We notice that the above entropy functional has also been employed in angiogenesis-chemotactic applications as an analytical tool [4].

However, the entropy inequality can be made rigorous only if $u_i \geq 0$, which cannot be easily obtained from the minimum principle. The nonnegativity of the solutions is obtained in [8] by proving that the transformed variable satisfies $w_i \in L^2(0,T;H^1(\Omega))$. As $H^1(\Omega)$ embeds continuously into $L^\infty(\Omega)$ in one space dimension, this implies $w_i(\cdot,t) \in L^\infty(\Omega)$ for almost every $t > 0$, and hence $u_i(\cdot,t) = \exp(w_i(\cdot,t)) > 0$ in Ω. Clearly, this method cannot be used in several space dimensions.

The main idea of our proof is to *discretize* the cross-diffusion term $\nabla(u_1 u_2)$ by *finite differences* and first to prove the existence of solutions to the approximate problem, which is now only weakly coupled. The precise approximation has to be chosen in such a way that the above entropy inequality also holds for the approximate problem. This provides the a priori estimates necessary to perform the limit of vanishing approximation parameters. The idea is inspired from [13], where a different problem is studied.

One possibility is to approximate the cross-diffusion term $\Delta(u_1 u_2) = \operatorname{div}(u_1 u_2 \nabla \ln(u_1 u_2))$ by the finite differences

$$D^{-h}[\chi_h u_1 u_2 D^h(\ln(u_1 u_2))],$$

where D^h is an approximation of the gradient,

$$D^h f = (D_1^h f, \ldots, D_N^h f) \quad \text{and} \quad D_j^h f(x,t) = \frac{f(x+he_j,t) - f(x,t)}{h}, \tag{1.7}$$

D^{-h} is an approximation of the divergence,

$$D^{-h}F(x,t) = \sum_{j=1}^N \frac{F_j(x-he_j,t) - F_j(x,t)}{-h}, \tag{1.8}$$

with the jth unit vector e_j of $\mathbb{R}^N$, $j = 1, \ldots, N$, and χ_h is the characteristic function of $\{x \in \Omega : \operatorname{dist}(x, \partial\Omega) > h\}$. It can be shown formally that the problem with this discrete cross-diffusion term possesses the entropy inequality

$$E(t) + \int_0^t \int_\Omega \Big(\sum_{i=1}^2 (4c_i |\nabla \sqrt{u_i}|^2 + 2a_i |\nabla u_i|^2) + \chi_h u_1 u_2 |D^h \ln(u_1 u_2)|^2 \Big) dt dx \leq E(0) + C$$

for some constant $C > 0$.

However, this estimate is valid only for *positive* population densities u_i. In order to deal with this difficulty, we employ Stampacchia's truncation method; i.e., we replace u_i by $(u_i)_+ + \eta$, where $(u_i)_+ = \max\{0, u_i\}$ and $\eta > 0$. This allows us to define the expression $\ln\big(((u_1)_+ + \eta)((u_2)_+ + \eta)\big)$, for instance.

The above estimate is formally derived by employing $\ln((u_i)_+ + \eta)$ as a test function in the weak formulation. Therefore, we obtain only estimates for $(u_i)_+$. In order to derive estimates also for $(u_i)_- = \min\{0, u_i\}$, we employ $(u_i)_-$ as a test function. This yields, for instance, an estimate of the type $\|(u_i)_-\|_{L^\infty(0,T;L^2(\Omega))} \leq C/|\ln \eta|$ for some constant $C > 0$ which is independent of η. In the limit $\eta \to 0$ this gives $(u_i)_- = 0$ in Q_T, and hence the nonnegativity of the population densities.

We notice that our strategy can also be applied to general systems of the type

$$\partial_t u - \operatorname{div}(A(u)\nabla u) = f(u),$$

where $u = u(x,t) \in \mathbb{R}^n$, $f(u) \in \mathbb{R}^n$ satisfies some growth condition, and $A(u) \in \mathbb{R}^{n\times n}$ is a diffusion matrix, maybe nonsymmetric and not positive definite, provided that

the system is symmetrizable in the sense given above and that the a priori estimates derived from the entropy inequality (which exists due to the symmetrizability) are sufficient to define a weak solution.

Let us summarize the main features of the presented method of proof:

- No restrictions on the diffusion coefficients c_i and a_i are needed.
- The global existence result holds in up to three space dimensions.
- The method provides the nonnegativity of the solutions.
- The degenerate case $c_i = 0$ can also be treated.

The idea of discretizing the cross-diffusion term by finite differences can be used for numerical purposes. We will exploit this idea in [2].

The paper is organized as follows. In section 2 we define and solve an approximate problem yielding an discrete entropy inequality. The key estimates are contained in Lemma 2.5. The limit of vanishing approximation parameters is then performed in section 3. Finally, in the appendix we recall the definition of Orlicz spaces and some of its properties.

2. Existence of solutions to an approximate problem. We use semidiscretization in time to construct the approximate problem. Moreover, as explained in the introduction, we also discretize the cross-diffusion terms by finite differences. For this, we decompose $(0,T] = \cup_{k=1}^K ((k-1)\tau, k\tau]$ for some $\tau > 0$ such that $\tau = T/K$. Furthermore, let $h > 0$, and let χ_h be the characteristic function of $\{x \in \Omega : \mathrm{dist}(x, \partial\Omega) > h\}$. Finally, let $0 < \eta < 1$, and set $\overline{s} = s/(1 + \eta(s)_+)$.

As the proof of Theorem 1.1 is highly technical, it is convenient, for the sake of a smoother presentation, to assume in this section the regularity $u_i^0 \in L^2(\Omega)$ and $q \in (L^\infty(Q_T))^N$ instead of the weaker conditions $u_i^0 \in L_\Psi(\Omega)$ and $q \in (L^2(Q_T))^N$. The general result can be proved by using appropriate smooth approximations and passing to the limit. Details are left to the reader. In fact, we simplify further and assume that $q \in (L^\infty(\Omega))^N$. The time-dependence can be treated as in [5], for instance, by averaging $q(x,t)$ over $((k-1)\tau, k\tau]$. Moreover, we assume that c_1, c_2 are *positive* numbers. We refer the reader to Remark 3.5 for the case $c_1 = 0$ or $c_2 = 0$.

For given u_1^{k-1}, $u_2^{k-1} \in L^2(\Omega)$, we solve recursively the problem

$$\begin{aligned}
&\frac{u_i^k - u_i^{k-1}}{\tau} - \mathrm{div}\big(c_i\nabla u_i^k + 2a_i((u_i^k)_+ + \eta)\nabla u_i^k + d_i(u_i^k)_+ q\big) \\
&= D^{-h}\left[\chi_h \overline{u_1^k}\,\overline{u_2^k} D^h \ln\big(((u_1^k)_+ + \eta)((u_2^k)_+ + \eta)\big)\right] + f_i\big((u_1^k)_+ + \eta, (u_2^k)_+ + \eta\big) \text{ in } \Omega, \\
&\big(c_i\nabla u_i^k + 2a_i((u_i^k)_+ + \eta)\nabla u_i^k + d_i(u_i^k)_+ q\big)\cdot\gamma = 0 \quad \text{on } \partial\Omega,
\end{aligned} \tag{2.1}$$

where $i = 1, 2$. The finite difference operators are defined in (1.7) and (1.8).

The existence of solutions to the approximate system (2.1) is proved in two steps. In order to apply Lax–Milgram's lemma we need bounded diffusion coefficients. Therefore, we approximate the diffusion coefficients $2a_i((u_i^k)_+ + \eta)$ by

$$2a_i \frac{(u_i^k)_+ + \eta}{1 + \nu((u_i^k)_+ + \eta)}$$

for some $\nu > 0$ and prove the existence of solutions to the resulting system. Then we derive uniform bounds with respect to ν which allows us to pass to the limit $\nu \to 0$.

The second approximate system reads as follows:

$$\frac{u_i^k-u_i^{k-1}}{\tau}-\operatorname{div}\Big(c_i\nabla u_i^k+2a_i\frac{(u_i^k)_+ +\eta}{1+\nu((u_i^k)_+ +\eta)}\nabla u_i^k+d_i(u_i^k)_+ q\Big)$$
$$=D^{-h}\Big[\chi_h\overline{u_1^k}\,\overline{u_2^k}D^h\ln\big(((u_1^k)_+ +\eta)((u_2^k)_+ +\eta)\big)\Big]+f_i\big((u_1^k)_+ +\eta,(u_2^k)_+ +\eta\big)\text{ in }\Omega,$$

$$\Big(c_i\nabla u_i^k+2a_i\frac{(u_i^k)_+ +\eta}{1+\nu((u_i^k)_+ +\eta)}\nabla u_i^k+d_i(u_i^k)_+ q\Big)\cdot\gamma=0\quad\text{on }\partial\Omega,\ i=1,2. \tag{2.2}$$

In subsection 2.1 we prove some bounds uniform in ν and the existence of weak solutions to (2.2). Then by letting $\nu\to 0$ in subsection 2.2 we conclude the solvability of (2.1).

In the following, C and $C(\cdots)$ denote positive constants with values varying from occurrence to occurrence and depending on the quantities indicated in the brackets.

2.1. Existence of solutions to the second approximate problem (2.2).

LEMMA 2.1. *Assume that the time discretization parameter $\tau>0$ is so small that*

$$\frac{3}{16\tau}\ge\max_{i=1,2}\Big\{\frac{d_i^2}{2c_i}\|q\|_{L^\infty(\Omega)}^2+2(R_i+\beta_{i1}+\beta_{i2})\Big\}\quad\textit{and}\quad 32\tau\le h^2\eta^2. \tag{2.3}$$

Then there exists a solution $(u_1,u_2)\in(H^1(\Omega))^2$ of problem (2.2) satisfying the following estimate:

$$\int_\Omega\sum_{i=1}^2\left(\frac{c_i}{2}|\nabla u_i|^2+\frac{u_i^2}{4\tau}+2a_i\frac{(u_i)_+ +\eta}{1+\nu((u_i)_+ +\eta)}|\nabla u_i|^2\right)dx\le C(\tau), \tag{2.4}$$

where the constant $C(\tau)>0$ depends on τ but not on ν.

The above estimate is used only to pass to the limit $\nu\to 0$ for *fixed* parameters τ, h, and η. For the limits $\tau,h\to 0$ and $\eta\to 0$ we need other estimates.

Remark 2.2. The second restriction on the time discretization parameter τ in (2.3) is similar to the well-known condition $\tau/h^2\le$ const needed for explicit finite difference approximations of parabolic equations since we treat the discrete cross-diffusion term in an "explicit" way. Clearly, this condition has no importance for the existence result.

Proof. Construct a mapping

$$\mathcal{T}:(\sigma,v_1,v_2)\in[0,1]\times(L^4(\Omega))^2\to(L^4(\Omega))^2$$

by solving the following linear problem:

$$\begin{aligned}-\operatorname{div}(c_i\nabla u_i)+\frac{u_i}{\tau}-\sigma\operatorname{div}\Big(2a_i\frac{(v_i)_+ +\eta}{1+\nu((v_i)_+ +\eta)}\nabla u_i\Big)-\sigma\operatorname{div}(d_i(v_i)_+ q)&\\ =\sigma\frac{u_i^{k-1}}{\tau}+\sigma F_i(v_1,v_2)\quad\text{in }\Omega,&\\ \Big(c_i\nabla u_i+2\sigma a_i\frac{(v_i)_+ +\eta}{1+\nu((v_i)_+ +\eta)}\nabla u_i+\sigma d_i(v_i)_+ q\Big)\cdot\gamma=0\quad\text{on }\partial\Omega,&\end{aligned} \tag{2.5}$$

where

$$F_i(v_1,v_2)=D^{-h}\left[\chi_h\overline{v_1}\,\overline{v_2}D^h\ln\big(((v_1)_+ +\eta)((v_2)_+ +\eta)\big)\right]+f_i((v_1)_+ +\eta,(v_2)_+ +\eta)$$

and v_1, $v_2 \in L^4(\Omega)$, $i = 1, 2$. The functionals F_i satisfy the estimate $\|F_i(v_1, v_2)\|_{L^2(\Omega)} \le C(1 + \|v_1\|^2_{L^4(\Omega)} + \|v_2\|^2_{L^4(\Omega)})$ for $i = 1, 2$. The above problem has a unique solution (by Lax–Milgram's lemma) since the diffusion coefficients are bounded. Thus, the mapping $\mathcal{T}$ is well defined. It is not difficult to prove the continuity of $\mathcal{T}$. Moreover, since the embedding $H^1(\Omega) \hookrightarrow L^4(\Omega)$ is compact, for every $\sigma \in [0, 1]$, the mapping $\mathcal{T}$ is compact. Here, we use the restriction $N \le 3$ of the space dimension (see Remark 2.3). When $\sigma = 0$, the equation $\mathcal{T}(0, u_1, u_2) = (u_1, u_2)$ immediately yields $u_1 = u_2 = 0$ in Ω.

It remains to establish uniform estimates for every fixed point of $\mathcal{T}$. Any fixed point (u_1, u_2) satisfies the equation

$$
\begin{aligned}
&- \operatorname{div}(c_i \nabla u_i) + \frac{u_i}{\tau} - \sigma \operatorname{div}\Big(2a_i \frac{(u_i)_+ + \eta}{1 + \nu((u_i)_+ + \eta)} \nabla u_i\Big) - \sigma \operatorname{div}(d_i (u_i)_+ q) \\
&\qquad = \sigma \frac{u_i^{k-1}}{\tau} + \sigma F_i(u_1, u_2) \quad \text{in } \Omega,\ i = 1, 2,
\end{aligned} \tag{2.6}
$$

together with homogeneous Neumann boundary conditions. We use $u_i \in H^1(\Omega)$ as a test function in the weak formulation of (2.6) for $i = 1, 2$ and add the resulting equations:

$$
\begin{aligned}
&\sum_{i=1}^2 \int_\Omega \Big(c_i |\nabla u_i|^2 + \frac{u_i^2}{\tau} + 2\sigma a_i \frac{(u_i)_+ + \eta}{1 + \nu((u_i)_+ + \eta)} |\nabla u_i|^2 \Big) dx \\
&= - \sum_{i=1}^2 \int_\Omega \sigma d_i (u_i)_+ q \cdot \nabla u_i dx + \frac{\sigma}{\tau} \sum_{i=1}^2 \int_\Omega u_i u_i^{k-1} dx \\
&\quad + \sum_{i=1}^2 \int_\Omega \sigma D^{-h} \left[\chi_h \overline{u_1}\, \overline{u_2} D^h \ln\big(((u_1)_+ + \eta)((u_2)_+ + \eta)\big) \right] u_i dx \\
&\quad + \sum_{i=1}^2 \int_\Omega \sigma \big[R_i - \beta_{i1}((u_1)_+ + \eta) - \beta_{i2}((u_2)_+ + \eta) \big] ((u_i)_+ + \eta) u_i dx.
\end{aligned} \tag{2.7}
$$

The terms on the right-hand side are estimated by Young's inequality. For the third term we also use the elementary inequalities $|\overline{u_i}| \le 1/\eta$ and $|\ln(x + \eta)| \le x + |\ln \eta|$ for all $x \ge 0$ and $0 < \eta < 1$. This yields after some computations

$$
\begin{aligned}
&\sum_{i=1}^2 \int_\Omega \Big(\frac{c_i}{2} |\nabla u_i|^2 + \frac{u_i^2}{4\tau} + 2\sigma a_i \frac{(u_i)_+ + \eta}{1 + \nu((u_i)_+ + \eta)} |\nabla u_i|^2 \Big) dx \\
&\le \frac{1}{\tau} \sum_{i=1}^2 \int_\Omega (u_i^{k-1})^2 dx + 2|\Omega| \sum_{i=1}^2 (R_i + \beta_{i1} + \beta_{i2}) + \frac{128\tau}{h^4 \eta^4} |\ln \eta|^2 |\Omega| \\
&\quad + \sum_{i=1}^2 \int_\Omega u_i^2 \Big(- \frac{1}{4\tau} + \frac{d_i^2}{2c_i} \|q\|^2_{L^\infty} + \frac{64\tau}{h^4 \eta^4} + 2(R_i + \beta_{i1} + \beta_{i2}) \Big) dx,
\end{aligned}
$$

where $|\Omega|$ is the measure of Ω. By choosing τ so small that (2.3) is satisfied, in particular $64\tau/(h^4\eta^4) \le 1/16\tau$, we obtain

$$
\sum_{i=1}^2 \int_\Omega \Big(\frac{c_i}{2} |\nabla u_i|^2 + \frac{u_i^2}{4\tau} + 2\sigma a_i \frac{(u_i)_+ + \eta}{1 + \nu((u_i)_+ + \eta)} |\nabla u_i|^2 \Big) dx
$$

$$\le \frac{1}{\tau}\sum_{i=1}^2\int_\Omega (u_i^{k-1})^2dx + \sum_{i=1}^2\int_\Omega u_i^2\Big(-\frac{3}{16\tau}+\frac{d_i^2}{2c_i}\|q\|_{L^\infty}^2+2(R_i+\beta_{i1}+\beta_{i2})\Big)dx + C(\tau)$$
$$\le \frac{1}{\tau}\sum_{i=1}^2\int_\Omega (u_i^{k-1})^2dx + C(\tau) \le C(\tau).$$

By the Leray–Schauder theorem, $\mathcal{T}(1,\cdot)$ has a fixed point. Thus we conclude the existence of a weak solution of problem (2.2). The inequality (2.4) follows from the above estimate with $\sigma = 1$. □

Remark 2.3. From the proof of the above lemma we see that if $\beta_{ij} = 0$ for all $i,j = 1,2$, then the fixed-point mapping $\mathcal{T}$ can be defined on $[0,1]\times(L^2(\Omega))^2$. This allows us to prove the above result for *any* space dimension N (see Remark 3.6).

2.2. The limit $\nu \to 0$. We show in the following that the limit $\nu\to 0$ can be performed in (2.2).

LEMMA 2.4. *There exists a weak solution $(u_1,u_2)\in(H^1(\Omega))^2$ of problem (2.1) in the sense that for all $\varphi\in W^{1,2\cdot 2^*/(2^*-2)}(\Omega)$,*

$$\begin{aligned}&\int_\Omega \frac{u_i-u_i^{k-1}}{\tau}\varphi dx + \int_\Omega (c_i\nabla u_i + 2a_i((u_i)_+ +\eta)\nabla u_i + d_i(u_i)_+ q)\cdot\nabla\varphi dx\\ &\quad -\int_\Omega D^{-h}\left[\chi_h\overline{u_1}\,\overline{u_2}D^h\ln\big(((u_1)_+ +\eta)((u_2)_+ +\eta)\big)\right]\varphi dx\\ &\quad = \int_\Omega f_i\big((u_1)_+ +\eta,(u_2)_+ +\eta\big)\varphi dx,\end{aligned}\tag{2.8}$$

where $2^*=\infty$ if $N=1$, 2^* can be any real number if $N=2$, and $2^*=2N/(N-2)$ if $N\ge 3$.

Proof. Let $(u_1^\nu,u_2^\nu)\in(H^1(\Omega))^2$ be a weak solution of (2.2). From the uniform estimate (2.4) we conclude the existence of a subsequence of (u_1^ν,u_2^ν) (not relabeled) such that, as $\nu\to 0$,

$$\begin{aligned}\nabla u_i^\nu &\rightharpoonup \nabla u_i \quad \text{weakly in } (L^2(\Omega))^N,\\ u_i^\nu &\to u_i \quad \text{strongly in } L^r(\Omega),\ 1\le r<2^*,\ i=1,2.\end{aligned}\tag{2.9}$$

The last convergence result follows from the compactness of the embedding $H^1(\Omega)\hookrightarrow L^r(\Omega)$ for all $r<2^*$. In particular, we have $(u_i^\nu)_+\to(u_i)_+$ strongly in $L^r(\Omega)$ and

$$((u_i^\nu)_+ +\eta)\nabla u_i^\nu \rightharpoonup ((u_i)_+ +\eta)\nabla u_i \quad \text{weakly in } (L^s(\Omega))^N \text{ for all } 1\le s\le\frac{2r}{r+2}.$$

Here, we used the fact that the product of a strongly convergent and a weakly convergent sequence is weakly convergent (in an appropriate space). Since (u_i^ν) is uniformly bounded in $H^1(\Omega)$, Hölder's inequality implies

$$\|((u_i^\nu)_+ +\eta)\nabla u_i^\nu\|_{L^{2\cdot 2^*/(2+2^*)}(\Omega)} \le \|(u_i^\nu)_+ +\eta\|_{L^{2^*}(\Omega)}\|\nabla u_i^\nu\|_{L^2(\Omega)}\le C,$$

where $C>0$ is independent of ν. Thus, the above weak convergence also holds for $s = 2\cdot 2^*/(2+2^*)$.

Now we use the following result: Let $(v_\nu)\subset L^\infty(\Omega)$ and $(w_\nu)\subset L^s(\Omega)$ with $s\ge 1$ be two sequences such that (v_ν) is bounded in $L^\infty(\Omega)$, $v_\nu\to v$ pointwise

almost everywhere in Ω as $\nu \to 0$, and $w_\nu \rightharpoonup w$ weakly in $L^s(\Omega)$. Then, as $\nu \to 0$, $v_\nu w_\nu \rightharpoonup vw$ weakly in $L^s(\Omega)$. Applying this result to $v_\nu = 1/(1+\nu((u_i^\nu)_+ + \eta))$ and $w_\nu = ((u_i^\nu)_+ + \eta)\nabla u_i^\nu$ with $s = 2\cdot 2^*/(2+2^*)$ yields

$$\frac{(u_i^\nu)_+ + \eta}{1+\nu((u_i^\nu)_+ + \eta)}\nabla u_i^\nu \rightharpoonup ((u_i)_+ + \eta)\nabla u_i \quad \text{weakly in } (L^s(\Omega))^N, \quad s = \frac{2\cdot 2^*}{2+2^*}.$$

Moreover, by similar arguments as above, as $\nu \to 0$,

$$\begin{aligned}
&f_i((u_1^\nu)_+ + \eta, (u_2^\nu)_+ + \eta) = \big(R_i - \beta_{i1}((u_1^\nu)_+ + \eta) - \beta_{i2}((u_2^\nu)_+ + \eta)\big)((u_i^\nu)_+ + \eta) \\
&\rightharpoonup \big(R_i - \beta_{i1}((u_1)_+ + \eta) - \beta_{i2}((u_2)_+ + \eta)\big]((u_i)_+ + \eta) \quad \text{weakly in } L^{2^*/2}(\Omega), \\
&D^{-h}\left[\chi_h \overline{u_1^\nu}\ \overline{u_2^\nu} D^h \ln\left(((u_1^\nu)_+ + \eta)((u_2^\nu)_+ + \eta)\right)\right] \\
&\rightharpoonup D^{-h}\left[\chi_h \overline{u_1}\ \overline{u_2} D^h \ln\left(((u_1)_+ + \eta)((u_2)_+ + \eta)\right)\right] \quad \text{weakly in } L^s(\Omega)
\end{aligned}$$

for all $1 < s < \infty$. These convergence results allow us to pass to the limit $\nu \to 0$ in the weak formulation of (2.2) which yields (2.8), and hence the conclusion. ∎

2.3. Uniform estimates with respect to τ and h. The following entropy inequality is the key estimate of this paper providing uniform bounds in τ, h, and η.

LEMMA 2.5. *Let $(u_1, u_2) \in (H^1(\Omega))^2$ be a solution of (2.1). Then the following estimates hold:*

$$\begin{aligned}
&\int_\Omega \left[\sum_{i=1}^2 \Big(c_i \frac{|\nabla(u_i)_+|^2}{(u_i)_+ + \eta} + a_i |\nabla(u_i)_+|^2\Big)\right. \\
&\left. + \chi_h \overline{u_1}\ \overline{u_2} \left|D^h \ln\left(((u_1)_+ + \eta)((u_2)_+ + \eta)\right)\right|^2\right] dx \\
&+ \frac{1}{\tau}\sum_{i=1}^2 \int_\Omega \left[((u_i)_+ + \eta)\left(\ln((u_i)_+ + \eta) - 1\right) + (u_i)_- \ln \eta\right] dx \\
&\le \frac{1}{\tau}\sum_{i=1}^2 \int_\Omega \left[((u_i^{k-1})_+ + \eta)\left(\ln((u_i^{k-1})_+ + \eta) - 1\right) + (u_i^{k-1})_- \ln \eta\right] dx + C
\end{aligned} \tag{2.10}$$

and

$$\begin{aligned}
&\sum_{i=1}^2 \int_\Omega \Big(\frac{c_i}{2}|\nabla(u_i)_-|^2 + 2a_i\eta|\nabla(u_i)_-|^2\Big)dx + \frac{1}{2\tau}\sum_{i=1}^2 \int_\Omega |(u_i)_-|^2 dx \\
&\le \frac{1}{2\tau}\sum_{i=1}^2 \int_\Omega |(u_i^{k-1})_-|^2 dx + \eta C \int_\Omega \sum_{i=1}^2 (|(u_i)_+|^2 + |(u_i)_-|^2) dx \\
&+ \frac{C(c_1,c_2)}{\eta^2}\int_\Omega \chi_h \overline{u_1}\ \overline{u_2} \left|D^h \ln\left(((u_1)_+ + \eta)((u_2)_+ + \eta)\right)\right|^2 dx + C,
\end{aligned} \tag{2.11}$$

where $C > 0$ depends only on R_i, β_{ij} $(i,j = 1,2)$, and $\|q\|_{L^2(\Omega)}$.

Proof. Let (u_1, u_2) be a solution of (2.1); i.e., $u_i \in H^1(\Omega)$ satisfies (2.8), $i = 1, 2$. As $\ln((u_i)_+ + \eta) \notin W^{1,2\cdot 2^*/(2^*-2)}(\Omega)$ in general, we cannot use this function as a test function in the weak formulation (2.8). Therefore, we choose a sequence (v^ε) of smooth functions satisfying $v^\varepsilon \to (u_i)_+$ in $H^1(\Omega)$ (for some fixed i) and $v^\varepsilon \ge 0$ in Ω

and use $\varphi = \ln(v^\varepsilon + \eta)$ as a test function in (2.8):

$$
\begin{aligned}
&\int_\Omega \frac{u_i - u_i^{k-1}}{\tau} \ln(v^\varepsilon + \eta) dx \\
&+ \int_\Omega \big(c_i \nabla u_i + 2a_i((u_i)_+ + \eta)\nabla u_i + d_i (u_i)_+ q\big) \cdot \nabla \ln(v^\varepsilon + \eta) dx \\
&- \int_\Omega D^{-h}\left[\chi_h \overline{u_1}\,\overline{u_2} D^h \ln\left(((u_1)_+ + \eta)((u_2)_+ + \eta)\right)\right] \ln(v^\varepsilon + \eta) dx \\
&= \int_\Omega \big[R_i - \beta_{i1}((u_1)_+ + \eta) - \beta_{i2}((u_2)_+ + \eta)\big]((u_i)_+ + \eta)\ln(v^\varepsilon + \eta) dx.
\end{aligned} \tag{2.12}
$$

We claim that, as $\varepsilon \to 0$,

$$
\int_\Omega ((u_i)_+ + \eta)\nabla u_i \cdot \nabla \ln(v^\varepsilon + \eta) dx \to \int_\Omega |\nabla (u_i)_+|^2 dx. \tag{2.13}
$$

In order to prove this claim we observe that

$$
((u_i)_+ + \eta)\nabla \ln(v^\varepsilon + \eta) \rightharpoonup ((u_i)_+ + \eta)\nabla \ln((u_i)_+ + \eta) = \nabla (u_i)_+
$$

weakly in $L^{2\cdot 2^*/(2+2^*)}(\Omega)$ and

$$
\|((u_i)_+ + \eta)\nabla \ln(v^\varepsilon + \eta)\|_{L^2(\Omega)} \le \left\| \frac{(u_i)_+ + \eta}{v^\varepsilon + \eta} \right\|_{L^\infty(\Omega)} \|\nabla v^\varepsilon\|_{L^2(\Omega)} \le C,
$$

where $C > 0$ is a constant independent of ε. Therefore, the above weak convergence holds also in $L^2(\Omega)$. Since $\nabla u_i \in L^2(\Omega)$, the claim follows.

As $\ln(v^\varepsilon + \eta) \to \ln((u_i)_+ + \eta)$ in $H^1(\Omega)$, we can pass to the limit $\varepsilon \to 0$ in (2.12). Adding (2.12) for $i = 1$ and $i = 2$ and using (2.13) then gives in the limit $\varepsilon \to 0$

$$
\begin{aligned}
&\int_\Omega \Big[\sum_{i=1}^2 \Big(c_i \frac{|\nabla (u_i)_+|^2}{(u_i)_+ + \eta} + 2a_i |\nabla (u_i)_+|^2 \Big) \\
&+ \chi_h \overline{u_1}\,\overline{u_2} \left| D^h \ln\left(((u_1)_+ + \eta)((u_2)_+ + \eta)\right)\right|^2 \Big] dx \\
&+ \sum_{i=1}^2 \int_\Omega \frac{u_i - u_i^{k-1}}{\tau} \ln((u_i)_+ + \eta) dx - \sum_{i=1}^2 \int_\Omega |d_i q \nabla (u_i)_+| dx \\
&\le \sum_{i=1}^2 \int_\Omega \big[R_i - \beta_{i1}((u_1)_+ + \eta) - \beta_{i2}((u_2)_+ + \eta)\big]((u_i)_+ + \eta)\ln((u_i)_+ + \eta) dx.
\end{aligned} \tag{2.14}
$$

In the following we estimate the terms of the above inequality. With the elementary inequality $x(\ln x - \ln y) \ge x - y$ for all $x, y > 0$ (which is a consequence of the convexity of $x \mapsto \ln x$), we obtain

$$
\begin{aligned}
&\int_\Omega \frac{u_i - u_i^{k-1}}{\tau} \ln((u_i)_+ + \eta) dx \\
&= \frac{1}{\tau} \int_\Omega \big[((u_i)_+ + \eta)\ln((u_i)_+ + \eta) - ((u_i^{k-1})_+ + \eta)\ln((u_i^{k-1})_+ + \eta) \\
&\quad + ((u_i^{k-1})_+ + \eta)(\ln((u_i^{k-1})_+ + \eta) - \ln((u_i)_+ + \eta))\big] dx \\
&\quad + \frac{1}{\tau} \int_\Omega ((u_i)_- - (u_i^{k-1})_-)\ln((u_i)_+ + \eta) dx
\end{aligned} \tag{2.15}
$$

$$\geq \frac{1}{\tau}\int_\Omega [((u_i)_+ + \eta)(\ln((u_i)_+ + \eta) - 1) + (u_i)_- \ln\eta]dx$$
$$- \frac{1}{\tau}\int_\Omega [((u_i^{k-1})_+ + \eta)(\ln((u_i^{k-1})_+ + \eta) - 1) + (u_i^{k-1})_- \ln\eta]dx.$$

The last term on the left-hand side in (2.14) is estimated by employing Young's inequality:

$$\sum_{i=1}^2 \int_\Omega |d_i q \nabla(u_i)_+|dx \leq \sum_{i=1}^2 a_i \int_\Omega |\nabla(u_i)_+|^2 dx + C(a_1, a_2, d_1, d_2, \|q\|_{L^2(\Omega)}). \tag{2.16}$$

Finally, by the assumptions $\beta_{ii} > 0$ and $\beta_{12} = \beta_{21}$, the right-hand side of (2.14) is uniformly bounded. Putting the above estimates (2.15)–(2.16) together, the first inequality (2.10) follows from (2.14).

In order to derive the second inequality (2.11), we take a sequence (v^ε) of smooth functions satisfying $v^\varepsilon \to (u_i)_-$ in $H^1(\Omega)$ and $v^\varepsilon = 0$ in $\{u_i \geq 0\}$, and we choose $\varphi = v^\varepsilon$ as a test function in the weak formulation (2.8):

$$\sum_{i=1}^2 \int_\Omega \Big(c_i \nabla(u_i)_- \cdot \nabla v^\varepsilon + 2a_i\eta\nabla(u_i)_- \cdot \nabla v^\varepsilon\Big)dx + \sum_{i=1}^2 \int_\Omega \frac{u_i - u_i^{k-1}}{\tau} v^\varepsilon dx$$
$$\leq -\sum_{i=1}^2 \int_\Omega \chi_h \overline{u_1}\,\overline{u_2} D^h \ln\big(((u_1)_+ + \eta)((u_2)_+ + \eta)\big) \cdot D^h v^\varepsilon dx$$
$$+ \sum_{i=1}^2 \int_\Omega \big[R_i - \beta_{i1}((u_1)_+ + \eta) - \beta_{i2}((u_2)_+ + \eta)\big]\eta v^\varepsilon dx.$$

As above we can let $\varepsilon \to 0$ to obtain

$$\sum_{i=1}^2 \int_\Omega \Big(c_i|\nabla(u_i)_-|^2 + 2a_i\eta|\nabla(u_i)_-|^2\Big)dx + \sum_{i=1}^2 \int_\Omega \frac{u_i - u_i^{k-1}}{\tau}(u_i)_- dx$$
$$\leq -\sum_{i=1}^2 \int_\Omega \chi_h \overline{u_1}\,\overline{u_2} D^h \ln\big(((u_1)_+ + \eta)((u_2)_+ + \eta)\big) \cdot D^h(u_i)_- dx \tag{2.17}$$
$$+ \sum_{i=1}^2 \int_\Omega \big[R_i - \beta_{i1}((u_1)_+ + \eta) - \beta_{i2}((u_2)_+ + \eta)\big]\eta(u_i)_- dx.$$

The second term on the left-hand side can be estimated as follows:

$$\int_\Omega \frac{u_i - u_i^{k-1}}{\tau}(u_i)_- dx = \frac{1}{\tau}\int_\Omega \big(|(u_i)_-|^2 - (u_i^{k-1})_+(u_i)_- - (u_i^{k-1})_-(u_i)_-\big)\,dx$$
$$\geq \frac{1}{2\tau}\int_\Omega \big(|(u_i)_-|^2 - |(u_i^{k-1})_-|^2\big)\,dx. \tag{2.18}$$

For the first term on the right-hand side of (2.17) we employ Young's inequality:

$$-\int_\Omega \chi_h \overline{u_1}\,\overline{u_2} D^h \ln\big(((u_1)_+ + \eta)((u_2)_+ + \eta)\big) \cdot D^h(u_i)_- dx$$
$$\leq \frac{c_i}{2}\int_\Omega |\nabla(u_i)_-|^2 dx + \frac{C(c_i)}{\eta^2}\int_\Omega \chi_h \overline{u_1}\,\overline{u_2}\,\big|D^h \ln\big(((u_1)_+ + \eta)((u_2)_+ + \eta)\big)\big|^2 dx + C. \tag{2.19}$$

Finally, for the last term on the right-hand side of (2.17) follows

$$\sum_{i=1}^2\int_\Omega \big[R_i-\beta_{i1}((u_1)_+ +\eta)-\beta_{i2}((u_2)_+ +\eta)\big]\eta(u_i)_-dx$$
$$\le \eta C\sum_{i=1}^2\int_\Omega (|(u_i)_+|^2+|(u_i)_-|^2)dx. \tag{2.20}$$

Hence, (2.11) is a consequence of (2.17)–(2.20). □

3. Proof of Theorem 1.1. Let $(u_1^k,u_2^k)\in(H^1(\Omega))^2$ be a solution to (2.1). We set $u_i^{(\tau)}(x,t)=u_i^k(x)$ if $(x,t)\in\Omega\times((k-1)\tau,k\tau]$. With the discrete time derivative

$$D_t^\tau v(x,t):=\frac{v(x,t+\tau)-v(x,t)}{\tau},\quad (x,t)\in\Omega\times[0,\infty),$$

we can rewrite the approximate problem (2.1) as

$$\begin{aligned}
&D_t^\tau u_i^{(\tau)}-\operatorname{div}\Big(c_i\nabla u_i^{(\tau)}+2a_i((u_i^{(\tau)})_+ +\eta)\nabla u_i^{(\tau)}+d_i(u_i^{(\tau)})_+q\Big)\\
&\quad -D^{-h}\Big[\chi_h\overline{u_1^{(\tau)}}\,\overline{u_2^{(\tau)}}D^h\ln\big(((u_1^{(\tau)})_+ +\eta)((u_2^{(\tau)})_+ +\eta)\big)\Big]\\
&=f_i\big((u_1^{(\tau)})_+ +\eta,(u_2^{(\tau)})_+ +\eta\big)\quad\text{in }\Omega,\\
&\Big(c_i\nabla u_i^{(\tau)}+2a_i((u_i^{(\tau)})_+ +\eta)\nabla u_i^{(\tau)}+d_i(u_i^{(\tau)})_+q\Big)\cdot\gamma=0\quad\text{on }\partial\Omega,
\end{aligned}\tag{3.1}$$

together with the initial conditions corresponding to (1.4).

The proof of Theorem 1.1 is divided into two parts. In subsection 3.1, we assume that $\eta>0$ is fixed and perform the limit $\tau,h\to 0$. In subsection 3.2, we prove the limit $\eta\to 0$. At this step we show the nonnegativity of the solution.

3.1. The limit $\tau,h\to 0$. The problem (2.1) has a solution under the condition that the parameters τ and h are related by the inequality $32\tau\le h^2\eta^2$. Therefore we let τ and h tend to zero simultaneously in such a way that the inequality $32\tau\le h^2\eta^2$ is satisfied (for fixed $\eta>0$).

LEMMA 3.1. *Let $T>0$. The following estimates hold for $i=1,2$:*

$$\|\nabla(u_i^{(\tau)})_+\|_{L^2(Q_T)}+\|(u_i^{(\tau)})_+\|_{L^\infty(0,T;L_\Psi(\Omega))}\le C, \tag{3.2}$$

$$\Big\|\chi_h\sqrt{\overline{u_1^{(\tau)}}\,\overline{u_2^{(\tau)}}}D^h\ln\Big((u_1^{(\tau)})_+ +\eta)((u_2^{(\tau)})_+ +\eta)\Big)\Big\|_{L^2(Q_T)}\le C, \tag{3.3}$$

$$\|\nabla(u_i^{(\tau)})_-\|_{L^2(Q_T)}+\|(u_i^{(\tau)})_-\|_{L^\infty(0,T;L^2(\Omega))}\le C/\eta, \tag{3.4}$$

where $C>0$ is independent of c_1, c_2, h, τ, and η. Furthermore,

$$\|u_i^{(\tau)}\|_{L^2(0,T;H^1(\Omega))}+\|u_i^{(\tau)}\|_{L^p(Q_T)}\le C(\eta), \tag{3.5}$$

$$\|D_t^\tau u_i^{(\tau)}\|_{L^r(0,T;(W^{1,r'}(\Omega))')}\le C(\eta), \tag{3.6}$$

where $p=(2N+2)/N$, $r=(2N+2)/(2N+1)$, $r'=r/(r-1)=2N+2$, and $C(\eta)>0$ does not depend on τ or h.

Proof. The estimates (3.2)–(3.5) are consequences of the key inequalities (2.10) and (2.11). First, we prove (3.2) and (3.3). Let $K\in\mathbb{N}$, and set $\tau=T/K$. The

estimate (2.10) can be rewritten at $t_k = k\tau$ as

$$\begin{aligned}
&\int_0^{t_k}\int_\Omega \Big[\sum_{i=1}^2 \Big(4c_i\Big|\nabla\sqrt{(u_i^{(\tau)})_+ + \eta}\Big|^2 + a_i|\nabla(u_i^{(\tau)})_+|^2\Big) \\
&\quad + \chi_h \overline{u_1^{(\tau)}}\,\overline{u_2^{(\tau)}}\Big|D^h \ln\big(((u_1^{(\tau)})_+ + \eta)((u_2^{(\tau)})_+ + \eta)\big)\Big|^2\Big]dxdt \\
&\quad + \sum_{i=1}^2 \int_\Omega \Big(((u_i^{(\tau)})_+ + \eta)(\ln((u_i^{(\tau)})_+ + \eta) - 1) + \ln\eta(u_i^{(\tau)})_-\Big)\Big|_{t=t_k} dx \\
&\le C(T, \|u_i^0\|_{L_\Psi(\Omega)}).
\end{aligned}$$

From the elementary inequalities $x \le x(\ln x - 1) + C$ and $(1+x)\ln(1+x) - x \le x(\ln x - 1) + x + C$ for all $x \ge 0$ for some $C > 0$ and from (4.1) we obtain at $t = t_k$

$$\begin{aligned}
&\int_\Omega (u_i^{(\tau)})_+ dx \le \int_\Omega ((u_i^{(\tau)})_+ + \eta)(\ln((u_i^{(\tau)})_+ + \eta) - 1)dx + C|\Omega| \le C, \\
&\|(u_i^{(\tau)})_+\|_{L_\Psi(\Omega)} \le 1 + \int_\Omega \Psi\big((u_i^{(\tau)})_+\big)dx \le C.
\end{aligned} \tag{3.7}$$

Since the functions $u_i^{(\tau)}$ are piecewise constant with respect to t, we have

$$\begin{aligned}
&\int_0^T\int_\Omega \Big[\sum_{i=1}^2 \Big(4c_i\Big|\nabla\sqrt{(u_i^{(\tau)})_+ + \eta}\Big|^2 + a_i|\nabla(u_i^{(\tau)})_+|^2\Big) \\
&\quad + \chi_h \overline{u_1^{(\tau)}}\,\overline{u_2^{(\tau)}}\Big|D^h(\ln\big(((u_1^{(\tau)})_+ + \eta)((u_2^{(\tau)})_+ + \eta)\big)\Big|^2\Big]dxdt \\
&\quad + \sum_{i=1}^2 \sup_{0<t<T}\Big(\|(u_i^{(\tau)})_+(\cdot,t)\|_{L_\Psi(\Omega)} + \|\ln\eta(u_i^{(\tau)})_-(\cdot,t)\|_{L^1(\Omega)}\Big) \le C.
\end{aligned} \tag{3.8}$$

This gives a uniform bound for $\|\nabla(u_i^{(\tau)})_+\|_{L^2(Q_T)}$ and shows (3.2)–(3.3). An L^2 bound for $(u_i^{(\tau)})_+$ can be derived from this estimate, the Poincaré inequality, and (3.7):

$$\int_0^T \|(u_i^{(\tau)})_+\|^2_{L^2(\Omega)}dt \le C(|\Omega|,T)\int_0^T \|\nabla(u_i^{(\tau)})_+\|^2_{L^2(\Omega)}dt + C(|\Omega|,T). \tag{3.9}$$

For the proof of (3.4) we employ the estimate (2.11), rewritten at $t_k = k\tau$ as

$$\begin{aligned}
&\int_0^{t_k}\int_\Omega \sum_{i=1}^2 \Big(\frac{c_i}{2}|\nabla(u_i^{(\tau)})_-|^2 + 2a_i\eta|\nabla(u_i^{(\tau)})_-|^2\Big)dxdt + \frac{1}{2}\sum_{i=1}^2\int_\Omega |(u_i^{(\tau)})_-(\cdot,t_k)|^2dx \\
&\le C + \eta C\sum_{i=1}^2\int_0^{t_k}\int_\Omega (|(u_i^{(\tau)})_+|^2 + |(u_i^{(\tau)})_-|^2)dxdt \\
&\quad + \frac{C}{\eta^2}\int_0^{t_k}\int_\Omega \chi_h \overline{u_1^{(\tau)}}\,\overline{u_2^{(\tau)}}\Big|D^h\ln\big(((u_1^{(\tau)})_+ + \eta)((u_2^{(\tau)})_+ + \eta)\big)\Big|^2 dxdt.
\end{aligned}$$

Taking into account (3.3) and (3.9) and applying Gronwall's inequality, this proves (3.4).

Next we show the estimate (3.5). As the functions $u_i^{(\tau)}$ are piecewise constant with respect to t, we obtain, with the help of (3.8) and (3.9),

$$\int_0^t\int_\Omega\sum_{i=1}^2\Big(\frac{c_i}{2}|\nabla(u_i^{(\tau)})_-|^2+2a_i\eta|\nabla(u_i^{(\tau)})_-|^2\Big)dxdt+\frac12\sum_{i=1}^2\int_\Omega|(u_i^{(\tau)})_-(\cdot,t)|^2dx$$
$$\le C(|\Omega|,T,\|u_i^0\|_{L_\Psi(\Omega)},\eta)+C\int_0^t\int_\Omega\sum_{i=1}^2|(u_i^{(\tau)})_-|^2dxdt.$$

Thus, by Gronwall's inequality,

$$\sum_{i=1}^2\int_0^T\int_\Omega\Big(\frac{c_i}{2}|\nabla(u_i^{(\tau)})_-|^2+2a_i\eta|\nabla(u_i^{(\tau)})_-|^2\Big)dxdt$$
$$+\frac12\sum_{i=1}^2\sup_{0<t<T}\int_\Omega|(u_i^{(\tau)})_-(\cdot,t)|^2dx\le C(\eta). \tag{3.10}$$

This provides a uniform bound for $(u_i^{(\tau)})_-$ in $L^2(0,T;H^1(\Omega))$, and from (3.2), (3.8), (3.9), and (3.10) we infer

$$\|u_i^{(\tau)}\|_{L^2(0,T;H^1(\Omega))}+\|u_i^{(\tau)}\|_{L^\infty(0,T;L^1(\Omega))}\le C(\eta).$$

Applying the Gagliardo–Nirenberg inequality with $p=(2N+2)/N$ and $\theta=2N(p-1)/(p(N+2))$ (and thus $\theta p=2$) yields

$$\|u_i^{(\tau)}\|_{L^p(Q_T)}\le\left(\int_0^T\|u_i^{(\tau)}\|_{L^1(\Omega)}^{(1-\theta)p}\|u_i^{(\tau)}\|_{H^1(\Omega)}^{\theta p}dt\right)^{1/p}$$
$$\le\|u_i^{(\tau)}\|_{L^\infty(0,T;L^1(\Omega))}^{1-\theta}\left(\int_0^T\|u_i^{(\tau)}\|_{H^1(\Omega)}^{\theta p}dt\right)^{1/p}\le C(\eta).$$

Finally, we derive a bound for the discrete time derivative $D_t^\tau u_i^{(\tau)}$. Using (3.1), we obtain, for $r=(2N+2)/(2N+1)$, since $p>r$,

$$\begin{aligned}
&\|D_t^\tau u_i^{(\tau)}\|_{L^r(0,T;(W^{1,r'}(\Omega))')}\\
&\le\|c_i\nabla u_i^{(\tau)}+2a_i((u_i^{(\tau)})_++\eta)\nabla u_i^{(\tau)}\|_{L^r(Q_T)}\\
&\quad+\Big\|\chi_h\overline{u_1^{(\tau)}}\,\overline{u_2^{(\tau)}}D^h\ln\big(((u_1^{(\tau)})_++\eta)((u_2^{(\tau)})_++\eta)\big)+d_i(u_i^{(\tau)})_+q\Big\|_{L^r(Q_T)}\\
&\quad+\big\|f_i\big((u_1^{(\tau)})_++\eta,(u_2^{(\tau)})_++\eta\big)\big\|_{L^r(Q_T)}\\
&\le C(|\Omega|,T)\|\nabla u_i^{(\tau)}\|_{L^2(Q_T)}+2a_i\|u_i^{(\tau)}+\eta\|_{L^p(Q_T)}\|\nabla u_i^{(\tau)}\|_{L^2(Q_T)}\\
&\quad+\frac12(\|u_1^{(\tau)}\|_{L^p(Q_T)}+\|u_2^{(\tau)}\|_{L^p(Q_T)})\\
&\quad\times\left\|\chi_h\sqrt{\overline{u_1^{(\tau)}}\,\overline{u_2^{(\tau)}}}D^h\ln\big(((u_1^{(\tau)})_++\eta)((u_2^{(\tau)})_++\eta)\big)\right\|_{L^2(Q_T)}\\
&\quad+|d_i|\,\|u_i^{(\tau)}\|_{L^p(Q_T)}\|q\|_{L^2(Q_T)}+C(T,|\Omega|)(\|u_i^{(\tau)}\|_{L^p(Q_T)}+\|u_i^{(\tau)}\|_{L^p(Q_T)}).
\end{aligned}$$

Then (3.6) follows from (3.3) and (3.5). ∎

Now we are able to perform the limit $\tau,\, h \to 0$.

LEMMA 3.2. *As $\tau, h \to 0$ such that $32\tau \le h^2\eta^2$, there exists a pair (u_1^η, u_2^η) satisfying (up to a subsequence which is not relabeled), for $i = 1, 2$,*

$$\nabla u_i^{(\tau)} \rightharpoonup \nabla u_i^\eta \quad \text{weakly in } (L^2(Q_T))^N, \tag{3.11}$$

$$((u_i^{(\tau)})_+ + \eta)\nabla u_i^{(\tau)} \rightharpoonup ((u_i^\eta)_+ + \eta)\nabla u_i^\eta \quad \text{weakly in } (L^r(Q_T))^N, \tag{3.12}$$

$$\chi_h \overline{u_1^{(\tau)}}\,\overline{u_2^{(\tau)}} D^h \ln\big(((u_1^{(\tau)})_+ + \eta)((u_2^{(\tau)})_+ + \eta)\big)$$

$$\rightharpoonup \overline{u_1^\eta}\,\overline{u_2^\eta}\nabla \ln\big(((u_1^\eta)_+ + \eta)((u_2^\eta)_+ + \eta)\big) \quad \text{weakly in } (L^2(Q_T))^N, \tag{3.13}$$

$$d_i(u_i^{(\tau)})_+ q \rightharpoonup d_i(u_i^\eta)_+ q \quad \text{weakly in } (L^r(Q_T))^N, \tag{3.14}$$

$$f_i\big((u_1^{(\tau)})_+ + \eta, (u_2^{(\tau)})_+ + \eta\big)$$

$$\rightharpoonup f_i\big((u_1^\eta)_+ + \eta, (u_2^\eta)_+ + \eta\big) \quad \text{weakly in } L^{p/2}(\Omega), \tag{3.15}$$

$$D_t^\tau u_i^{(\tau)} \rightharpoonup \partial_t u_i^\eta \quad \text{weakly in } L^r(0,T;(W^{1,r'}(\Omega))'),$$

where $p = (2N+2)/N$, $r = (2N+2)/(2N+1)$, and $r' = 2N+2$.

Proof. The first and last convergences are direct consequences of (3.2) and (3.5). In order to treat the nonlinear terms, we need a strong convergence result. Taking into account (3.5) and (3.6), we can apply the version of Aubin's lemma in [26, Thm. 6] to obtain, for a subsequence which is not relabeled, as $\tau, h \to 0$,

$$u_i^{(\tau)} \to u_i^\eta \quad \text{strongly in } L^q(0,T;L^2(\Omega)), \quad 1 < q < 2. \tag{3.16}$$

In particular, (a subsequence of) $(u_i^{(\tau)})$ converges pointwise almost everywhere in Q_T to u_i^η. This, together with the bound $\|u_i^{(\tau)}\|_{L^p(Q_T)} \le C$ (which comes from (3.5)), implies

$$u_i^{(\tau)} \to u_i^\eta \quad \text{strongly in } L^\alpha(Q_T), \quad 2 < \alpha < p, \tag{3.17}$$

and $(u_i^{(\tau)})_+ \to (u_i^\eta)_+$ strongly in $L^\alpha(Q_T)$. By (3.11) we obtain for $s = 2\alpha/(2+\alpha) < r$

$$(u_i^{(\tau)})_+ \nabla u_i^{(\tau)} \rightharpoonup (u_i^\eta)_+ \nabla u_i^\eta \quad \text{weakly in } (L^s(Q_T))^N.$$

Since

$$\|(u_i^{(\tau)})_+ \nabla u_i^{(\tau)}\|_{L^r(Q_T)} \le \|u_i^{(\tau)}\|_{L^p(Q_T)} \|\nabla u_i^{(\tau)}\|_{L^2(Q_T)} \le C,$$

the above weak convergence also holds for $s = r$. In a similar way, since $q \in (L^2(Q_T))^N$, the convergences (3.14) and (3.15) can be proved.

Finally, we show (3.13). Using

$$\|\ln((u_i^{(\tau)})_+ + \eta)\|_{L^2(Q_T)} \le |\ln\eta| + \|(u_i^{(\tau)})_+ + \eta\|_{L^2(Q_T)} \le C(\eta),$$

$$\|D^h \ln((u_i^{(\tau)})_+ + \eta)\|_{L^2(Q_T)} \le \frac{C}{\eta} \|\nabla (u_i^{(\tau)})_+\|_{L^2(Q_T)} \le C(\eta),$$

and $(u_i^{(\tau)})_+ \to (u_i^\eta)_+$ almost everywhere in Q_T, we conclude that

$$D^h \ln((u_i^{(\tau)})_+ + \eta) \rightharpoonup \nabla \ln((u_i^\eta)_+ + \eta) \quad \text{weakly in } (L^2(Q_T))^N.$$

Then (3.13) follows from

$$\left\|\chi_h \overline{u_1^{(\tau)}}\,\overline{u_2^{(\tau)}}\right\|_{L^\infty(Q_T)} \le \frac{1}{\eta^2}.$$

This proves Lemma 3.2. □

Letting $\tau, h \to 0$ in the weak version of (3.1) such that $32\tau \le h^2\eta^2$, we obtain for all $\varphi \in L^{r'}(0,T;W^{1,r'}(\Omega))$

$$\begin{aligned}\int_0^T \langle \partial_t u_i^\eta, \varphi\rangle_{(W^{1,r'}(\Omega))',W^{1,r'}(\Omega)}\,dt + \int_{Q_T} (c_i\nabla u_i^\eta + 2a_i((u_i^\eta)_+ + \eta)\nabla u_i^\eta)\cdot\nabla\varphi dxdt \\ + \int_{Q_T}\Big[\overline{u_1^\eta}\,\overline{u_2^\eta}\nabla\ln\big(((u_1^\eta)_+ + \eta)((u_2^\eta)_+ + \eta)\big) + d_i(u_i)_+ q\Big]\cdot\nabla\varphi dxdt \\ = \int_{Q_T} f_i\big((u_1^\eta)_+ + \eta, (u_2^\eta)_+ + \eta\big)\varphi dxdt.\end{aligned} \tag{3.18}$$

By Lemma 3.2, the functions u_1^η and u_2^η are satisfying the properties

$$\begin{aligned}&u_i^\eta \in L^2(0,T;H^1(\Omega)) \cap L^p(Q_T),\\ &(u_i^\eta)_+ \in L^\infty(0,T;L_\Psi(\Omega)),\quad (u_i^\eta)_- \in L^\infty(0,T;L^2(\Omega)),\quad i=1,2.\end{aligned} \tag{3.19}$$

3.2. The limit $\eta \to 0$. The last step in the proof of Theorem 1.1 is to perform the limit $\eta \to 0$. First, we need some a priori estimates.

LEMMA 3.3. *Let $T > 0$. The following estimates hold for $i = 1, 2$:*

$$\|\nabla(u_i^\eta)_+\|_{L^2(Q_T)} + \|(u_i^\eta)_+\|_{L^\infty(0,T;L_\Psi(\Omega))} \le C, \tag{3.20}$$

$$\|\ln\eta(u_i^\eta)_-\|_{L^\infty(0,T;L^1(\Omega))} \le C, \tag{3.21}$$

$$\left\|\sqrt{\overline{u_1^\eta}\,\overline{u_2^\eta}}\nabla\ln\big(((u_1^\eta)_+ + \eta)((u_2^\eta)_+ + \eta)\big)\right\|_{L^2(Q_T)} \le C, \tag{3.22}$$

$$\|\nabla(u_i^\eta)_-\|_{L^2(Q_T)} + \|(u_i^\eta)_-\|_{L^\infty(0,T;L^2(\Omega))} \le C, \tag{3.23}$$

$$\|u_i^\eta\|_{L^2(0,T;H^1(\Omega))} + \|u_i^\eta\|_{L^p(Q_T)} \le C, \tag{3.24}$$

$$\|\partial_t u_i^\eta\|_{L^r(0,T;(W^{1,r'}(\Omega))')} \le C, \tag{3.25}$$

where $p = (2N+2)/N$, $r = (2N+2)/(2N+1)$, $r' = 2N+2$, and $C > 0$ is a constant independent of c_1, c_2, and η.

Proof. Let $i \in \{1,2\}$. Choose a sequence (v^ε) of smooth functions such that, as $\varepsilon \to 0$,

$$v^\varepsilon \to (u_i^\eta)_+ \quad \text{in } L^2(0,T;H^1(\Omega)) \cap L^\infty(0,T;L_\Psi(\Omega)) \cap L^p(Q_T) \tag{3.26}$$

and $v^\varepsilon = 0$ on $\{u_i \le 0\}$. Such a choice is possible in view of the regularity (3.19). We claim that

$$\begin{aligned}\int_0^t \langle \partial_t u_i^\eta, \ln(v^\varepsilon + \eta)\rangle dt \to \int_\Omega \big((u_i^\eta)_+ + \eta)(\ln(u_i^\eta)_+ + \eta) - 1) + \ln\eta(u_i^\eta)_-\big)dx \\ - \int_\Omega (u_i^0 + \eta)(\ln(u_i^0 + \eta) - 1)dx\end{aligned} \tag{3.27}$$

and

$$\int_{Q_T} a_i((u_i^\eta)_+ + \eta)\nabla u_i^\eta \cdot \nabla\ln(v^\varepsilon + \eta)dxdt \to \int_{Q_T} a_i|\nabla(u_i^\eta)_+|^2 dxdt. \tag{3.28}$$

In fact, in order to show the second claim (3.28), we need only to show

$$((u_i^\eta)_+ + \eta)\nabla \ln(v^\varepsilon + \eta) \rightharpoonup \nabla (u_i^\eta)_+ \quad \text{weakly in } (L^2(Q_T))^N.$$

This convergence follows from $((u_i^\eta)_+ + \eta)/(v^\varepsilon + \eta) \to 1$ almost everywhere in Q_T and $\nabla v^\varepsilon \to \nabla (u_i^\eta)_+$ in $(L_2(Q_T))^N$.

The proof of the first claim (3.27) is more delicate. By integration by parts, we have

(3.29)

$$\begin{aligned}\int_0^t \langle \partial_t u_i^\eta, \ln(v^\varepsilon + \eta)\rangle dt &= \int_0^t \langle \partial_t (u_i^\eta)_+, \ln(v^\varepsilon + \eta)\rangle dt + \int_0^t \langle \partial_t (u_i^\eta)_-, \ln \eta\rangle dt \\ &= \int_\Omega [((u_i^\eta)_+ + \eta)\ln(v^\varepsilon + \eta)]_0^t \, dx - \int_{Q_t} \frac{(u_i^\eta)_+ + \eta}{v^\varepsilon + \eta} \partial_t v^\varepsilon dx dt + \ln \eta \int_\Omega [(u_i^\eta)_-]_0^t \, dx.\end{aligned}$$

We consider the first term on the right-hand side. It holds for all $t \in (0,T)\backslash\mathcal{N}$, where $\mathcal{N}$ is a set of measure zero, that

$$\|\ln(v^\varepsilon(\cdot,t) + \eta) - \ln((u_i^\eta)_+(\cdot,t) + \eta)\|_{L^\infty(\Omega)} = \left\| \ln \frac{v^\varepsilon(\cdot,t) + \eta}{(u_i^\eta)_+(\cdot,t) + \eta} \right\|_{L^\infty(\Omega)} \le C$$

for some $C > 0$ and, as $\varepsilon \to 0$,

$$\ln(v^\varepsilon(\cdot,t) + \eta) - \ln((u_i^\eta)_+(\cdot,t) + \eta) \to 0 \quad \text{strongly in } L^1(\Omega),$$

uniformly in $t \in (0,T)\backslash\mathcal{N}$. In particular, this sequence converges in measure. Now let $\Phi(s) = e^s - s - 1$ be the complementary Young function to Ψ, and define $\Phi_2(s) = \exp(s^2) - 1$, $s \ge 0$. Then Φ_2 is a Young function, and

$$\lim_{t\to\infty} \frac{\Phi(kt)}{\Phi_2(t)} = 0 \quad \text{for all } k > 0.$$

Thus, by Theorem 4.1 of the appendix,

$$\ln(v^\varepsilon(\cdot,t) + \eta) - \ln((u_i^\eta)_+(\cdot,t) + \eta) \to 0 \quad \text{strongly in } L_\Phi(\Omega),$$

uniformly in $t \in (0,T)\backslash\mathcal{N}$. Therefore, as $(u_i^\eta)_+ + \eta \in L^\infty(0,T;L_\Psi(\Omega))$, Young's inequality (4.2) implies, for $t \in (0,T)\backslash\mathcal{N}$,

$$\begin{aligned}&\int_\Omega ((u_i^\eta)_+(\cdot,t) + \eta)\big(\ln(v^\varepsilon(\cdot,t) + \eta) - \ln((u_i^\eta)_+(\cdot,t) + \eta)\big)dx \\ &\le 2\|(u_i^\eta)_+(\cdot,t) + \eta\|_{L_\Psi(\Omega)} \|\ln(v^\varepsilon(\cdot,t) + \eta) - \ln((u_i^\eta)_+(\cdot,t) + \eta)\|_{L_\Phi(\Omega)} \to 0.\end{aligned}$$

We conclude that, for almost every $t \in (0,T)$, as $\varepsilon \to 0$,

$$\int_\Omega [((u_i^\eta)_+ + \eta)\ln(v^\varepsilon + \eta)]_0^t dx \to \int_\Omega [((u_i^\eta)_+ + \eta)\ln((u_i^\eta)_+ + \eta)]_0^t \, dx.$$

It remains to treat the second term in (3.29). Let $(0,t] = \cup_{k=0}^{K-1}(t_k, t_{k+1}]$, where $t_k \in (0,t]\backslash\mathcal{N}$ and $t_K := t$, be a partition of the interval $(0,t]$. Then we can write the term as follows:

$$\begin{aligned}&\lim_{\varepsilon\to 0} \int_{Q_t} \frac{(u_i^\eta)_+ + \eta}{v^\varepsilon + \eta} \partial_t v^\varepsilon dx dt \\ &= \lim_{\varepsilon\to 0} \lim_{K\to\infty} \sum_{k=0}^{K-1} \int_\Omega \frac{(u_i^\eta)_+(x,t_k) + \eta}{v^\varepsilon(x,t_k) + \eta} (v^\varepsilon(x,t_{k+1}) - v^\varepsilon(x,t_k))dx.\end{aligned}$$

The sequence $((u_i^\eta)_+(\cdot,t_k)+\eta)/(v^\varepsilon(\cdot,t_k)+\eta)$ converges to one weakly* in $L^\infty(\Omega)$ as $\varepsilon\to 0$ and $v^\varepsilon(\cdot,t_{k+1})-v^\varepsilon(\cdot,t_k)$ converges to $(u_i^\eta)_+(\cdot,t_{k+1})-{-(u_i^\eta)_+}(\cdot,t_k)$ strongly in $L^1(\Omega)$, uniformly in $t\in(0,T)\backslash\mathcal{N}$. Hence, we can exchange the limits $\varepsilon\to 0$ and $K\to\infty$ to obtain

$$\begin{aligned}
&\lim_{\varepsilon\to 0}\int_{Q_t}\frac{(u_i^\eta)_+ +\eta}{v^\varepsilon+\eta}\partial_t v^\varepsilon dxdt\\
&=\lim_{K\to\infty}\lim_{\varepsilon\to 0}\sum_{k=0}^{K-1}\int_\Omega\frac{(u_i^\eta)_+(x,t_k)+\eta}{v^\varepsilon(x,t_k)+\eta}(v^\varepsilon(x,t_{k+1})-v^\varepsilon(x,t_k))dx\\
&=\lim_{K\to\infty}\sum_{k=0}^{K-1}\int_\Omega((u_i^\eta)_+(x,t_{k+1})-(u_i^\eta)_+(x,t_k))dx\\
&=\int_\Omega((u_i^\eta)_+(x,t)-u_i^0(x))dx.
\end{aligned}$$

This proves (3.27).

Now we use $\varphi=\ln(v^\varepsilon+\eta)$ as a test function in (3.18) and perform the limit $\varepsilon\to 0$ by employing the above claims (3.27) and (3.28). This implies, after addition of the two equations for $i=1,2$ and estimating as above,

$$\begin{aligned}
&\sum_{i=1}^2\int_\Omega\big((u_i^\eta)_+ +\eta)(\ln(u_i^\eta)_+ +\eta)-1)+\ln\eta(u_i^\eta)_-\big)dx\\
&+\int_{Q_T}\sum_{i=1}^2\Big(c_i\frac{|\nabla(u_i^\eta)_+|^2}{(u_i^\eta)_+ +\eta}+a_i|\nabla(u_i^\eta)_+|^2\Big)dxdt\\
&+\int_{Q_T}\overline{u_1^\eta}\,\overline{u_2^\eta}\left|\nabla\ln\big(((u_1^\eta)_+ +\eta)((u_2^\eta)_+ +\eta)\big)\right|^2dxdt\le C,
\end{aligned}\tag{3.30}$$

where $C>0$ depends only on a_1, a_2, $\|q\|_{L^2(Q_T)}$, and $\|u_i^0\|_{L_\Psi(\Omega)}$. This shows (3.20)–(3.22).

In the next step, we choose $\varphi=w^\varepsilon$ as a test function in (3.18), where (w^ε) is a smooth sequence such that, as $\varepsilon\to 0$,

$$w^\varepsilon\to(u_i^\eta)_-\quad\text{in } L^2(0,T;H^1(\Omega))\cap L^\infty(0,T;L^2(\Omega))\tag{3.31}$$

and $w^\varepsilon=0$ in $\{u_i^\eta\ge 0\}$. Then we have

$$\begin{aligned}
&\int_0^t\langle\partial_t(u_i^\eta)_-,w^\varepsilon\rangle dt+\int_{Q_t}(c_i\nabla(u_i^\eta)_-\cdot\nabla w^\varepsilon+2a_i\eta\nabla(u_i^\eta)_-\cdot\nabla w^\varepsilon)dxdt\\
&=\int_{Q_t}\big[R_i-\beta_{i1}((u_1^\eta)_+ +\eta)-\beta_{i2}((u_2^\eta)_+ +\eta)\big]\eta w^\varepsilon dxdt.
\end{aligned}\tag{3.32}$$

We infer from (3.31) that

$$\begin{aligned}
&\lim_{\varepsilon\to 0}\int_0^t\langle\partial_t(u_i^\eta)_-,w^\varepsilon\rangle dt=\lim_{\varepsilon\to 0}\frac12\int_\Omega[|w^\varepsilon|^2]_0^t dx+\lim_{\varepsilon\to 0}\int_0^t\langle\partial_t((u_i^\eta)_- -w^\varepsilon),w^\varepsilon\rangle dt\\
&=\frac12\int_\Omega[|(u_i^\eta)_-|^2]_0^t dx+\lim_{\varepsilon\to 0}\int_\Omega[((u_i^\eta)_- -w^\varepsilon)w^\varepsilon]_0^t dx+\lim_{\varepsilon\to 0}\int_{Q_t}\partial_t w^\varepsilon((u_i^\eta)_- -w^\varepsilon)dxdt\\
&=\frac12\int_\Omega[|(u_i^\eta)_-|^2]_0^t dx
\end{aligned}$$

$$
\begin{aligned}
&+\lim_{\varepsilon\to 0}\lim_{K\to\infty}\sum_{k=0}^{K-1}\int_\Omega (w^\varepsilon(x,t_{k+1})-w^\varepsilon(x,t_k))((u_i^\eta)_-(x,t_k)-w^\varepsilon(x,t_k))dx\\
&=\frac{1}{2}\int_\Omega [|(u_i^\eta)_-|^2]_0^t dx,
\end{aligned}
$$

where similarly as above $(0,t]=\cup_{k=0}^{K-1}(t_k,t_{k+1}]$. The convergence of the other terms in (3.32) as $\varepsilon\to 0$ follows directly from (3.31). This yields

$$
\begin{aligned}
&\frac{1}{2}\sum_{i=1}^2\int_\Omega |(u_i^\eta)_-(\cdot,t)|^2dx+\sum_{i=1}^2\int_{Q_t}(c_i|\nabla(u_i^\eta)_-|^2+2a_i\eta|\nabla(u_i^\eta)_-|^2)dxdt\\
&=\sum_{i=1}^2\int_{Q_t}\big[R_i-\beta_{i1}((u_1^\eta)_+ +\eta)-\beta_{i2}((u_2^\eta)_+ +\eta)\big]\eta(u_i^\eta)_-dxdt\\
&\le C\sum_{i=1}^2\int_{Q_t}(|(u_i^\eta)_+|^2+|(u_i^\eta)_-|^2)+C,
\end{aligned}
\tag{3.33}
$$

where $C>0$ is independent of η. The estimate (3.20) and Gronwall's inequality then imply (3.23). Finally, the inequalities (3.24) and (3.25) can be derived similarly as in the proof of Lemma 3.1. ∎

LEMMA 3.4. *As $\eta\to 0$, there exist functions u_1, $u_2\ge 0$ such that the following convergences hold (up to subsequences which are not relabeled), for $i=1,2$:*

$$
\begin{aligned}
c_i\nabla u_i^\eta &\rightharpoonup c_i\nabla u_i \quad \text{weakly in } (L^2(Q_T))^N,\\
2a_i((u_i^\eta)_+ +\eta)\nabla u_i^\eta &\rightharpoonup 2a_iu_i\nabla u_i \quad \text{weakly in } (L^r(Q_T))^N,\\
\overline{u_1^\eta}\,\overline{u_2^\eta}\nabla\ln((u_1^\eta)_+ +\eta)((u_2^\eta)_+ +\eta) &\rightharpoonup \nabla(u_1u_2) \quad \text{weakly in } (L^r(Q_T))^N,\\
d_i(u_i^\eta)_+ q &\rightharpoonup d_iu_iq \quad \text{weakly in } (L^r(Q_T))^N,\\
f_i\big((u_1^\eta)_+ +\eta,(u_2^\eta)_+ +\eta\big) &\rightharpoonup f_i(u_1,u_2) \quad \text{weakly in } L^{p/2}(Q_T),\\
\partial_t u_i^\eta &\rightharpoonup \partial_t u_i \quad \text{weakly in } L^r(0,T;(W^{1,r'}(\Omega))').
\end{aligned}
$$

Proof. Similar to the discussion in the proof of Lemma 3.2, we conclude that there exist functions u_1 and u_2 such that $u_i^\eta\to u_i$ in $L^\alpha(Q_T)$ for all $2\le\alpha<p$. The estimate (3.21) implies

$$
\|(u_i^\eta)_-\|_{L^\infty(0,T;L^1(\Omega))}\le\frac{C}{|\ln\eta|}\to 0,
$$

from which we obtain $u_i\ge 0$ in Q_T, $i=1,2$.

Except for the third convergence, the discussion of the remaining convergence results are similar to those in the proof of Lemma 3.2. Observe that

$$
\overline{u_1^\eta}\,\overline{u_2^\eta}\nabla\ln((u_1^\eta)_+ +\eta)=\frac{1}{1+\eta(u_2^\eta)_+}\frac{1}{1+\eta(u_1^\eta)_+}\frac{(u_1^\eta)_+}{(u_1^\eta)_+ +\eta}(u_2^\eta)_+\nabla(u_1^\eta)_+.
$$

By similar arguments as above, it holds that

$$(u_2^\eta)_+\nabla(u_1^\eta)_+ \rightharpoonup u_2\nabla u_1 \quad \text{weakly in } (L^r(Q_T))^N.$$

Taking into account

$$\frac{1}{1+\eta(u_2^\eta)_+}\frac{1}{1+\eta(u_1^\eta)_+}\frac{(u_1^\eta)_+}{(u_1^\eta)_+ + \eta} \le 1,$$

we infer the desired convergence. □

Now, Theorem 1.1 is a consequence of the convergence results of Lemma 3.4 applied to (3.18).

Remark 3.5. Since the estimates in Lemma 3.3 are independent of c_1 and c_2, we obtain the existence of a weak solution even in the case $c_1 = 0$ or $c_2 = 0$. Indeed, we first obtain a weak solution for $c_1 > 0$, $c_2 > 0$, respectively. The a priori estimates of Lemma 3.3 allow us to perform the limit $c_1, c_2 \to 0$.

Remark 3.6. All the above estimates are true in *any* space dimension. The restriction $N \le 3$ is used only in the proof of Lemma 2.1. As mentioned in Remark 2.3, Lemma 2.1 holds in any space dimension if $\beta_{ij} = 0$ for all $i, j = 1, 2$. Therefore, Theorem 1.1 is true in *any* space dimension, provided that $\beta_{ij} = 0$ for all $i, j = 1, 2$.

4. Appendix. We recall the definition of an Orlicz space and some of its properties. For details, we refer the reader, e.g., to [1, 14].

A real-valued function $\Psi : [0,\infty) \to \mathbb{R}$ is called a *Young function* if $\Psi(t) = \int_0^t \psi(s)ds$ and $\psi : [0,\infty) \to [0,\infty)$ has the following properties:

- $\psi(0) = 0$, $\psi > 0$ on $(0,\infty)$, $\psi(t) \to \infty$ as $t \to \infty$;
- ψ is nondecreasing and right continuous at any point $s \ge 0$.

The function $\Phi(t) = \int_0^t \phi(s)ds$ with $\phi(s) = \sup_{\psi(t)\le s} t$ is called the *complementary Young function* of Ψ. For instance, $\Psi(s) = (1+s)\ln(1+s) - s$ and $\Phi(s) = e^s - s - 1$ are a pair of complementary Young functions.

Let Ψ be a Young function. The *Orlicz class* $K_\Psi(\Omega)$ is the set of (equivalence classes of) real-valued measurable functions u on Ω satisfying $\int_\Omega \Psi(|u(x)|)dx < \infty$. Then the *Orlicz space* $L_\Psi(\Omega)$ is the linear hull of $K_\Psi(\Omega)$ supplemented with the *Luxemburg norm*

$$\|u\|_{L_\Psi(\Omega)} := \inf\left\{k > 0 : \int_\Omega \Psi\Big(\frac{|u(x)|}{k}\Big) \le 1\right\}.$$

With this norm, the Orlicz space $L_\Psi(\Omega)$ is a Banach space.

We need some properties of Orlicz spaces. The first is the inequality [14, sections 3.6.3 and 3.8.5]

$$\|u\|_{L_\Psi(\Omega)} \le 1 + \int_\Omega \Psi(|u(x)|)dx, \quad u \in L_\Psi. \tag{4.1}$$

The second property is the *Hölder inequality* [14, sections 3.8.5 and 3.8.6]: Let Ψ and Φ be a pair of complementary Young functions and $u \in L_\Psi(\Omega)$, $v \in L_\Phi(\Omega)$. Then

$$\left|\int_\Omega uvdx\right| \le 2\|u\|_{L_\Psi(\Omega)}\|v\|_{L_\Phi(\Omega)}. \tag{4.2}$$

Finally, we need the following theorem [1, Thm. 8.22].

THEOREM 4.1. *Let $\Omega \in \mathbb{R}^N$ be bounded, and let Φ_1 and Φ_2 be two Young functions such that for all $k > 0$,*

$$\lim_{t\to\infty} \frac{\Phi_1(kt)}{\Phi_2(t)} = 0.$$

Then, any (u_n) sequence which is bounded in $L_{\Phi_2}(\Omega)$ and convergent in measure is convergent in $L_{\Phi_1}(\Omega)$.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces,* Academic Press, New York, 1975.

[2] L. CHEN AND A. JÜNGEL, work in preparation, 2004.

[3] Y. S. CHOI, R. LUI, AND Y. YAMADA, *Existence of global solutions for the Shigesada-Kawasaki-Teramoto model with weak cross-diffusion*, Discrete Contin. Dynam. Systems, 9 (2003), pp. 1193–1200.

[4] L. CORRIAS, B. PERTHAME, AND H. ZAAG, *A chemotaxis model motivated by angiogenesis*, C. R. Acad. Sci. Paris Sér. I Math., 336 (2003), pp. 141–146.

[5] P. DEGOND, S. GÉNIEYS, AND A. JÜNGEL, *Symmetrization and entropy inequality for general diffusion equations*, C. R. Acad. Sci. Paris Sér. I Math., 325 (1997), pp. 963–968.

[6] P. DEURING, *An initial-boundary value problem for a certain density-dependent diffusion system*, Math. Z., 194 (1987), pp. 375–396.

[7] H. GAJEWSKI AND K. ZACHARIAS, *Global behaviour of a reaction-diffusion system modelling chemotaxis*, Math. Nachr., 195 (1998), pp. 77–114.

[8] G. GALIANO, M. L. GARZÒN, AND A. JÜNGEL, *Semi-discretization and numerical convergence of a nonlinear cross-diffusion population model*, Numer. Math., 93 (2003), pp. 655–673.

[9] T. HILLEN AND K. PAINTER, *Global existence for a parabolic chemotaxis model with prevention of overcrowding*, Adv. in Appl. Math., 26 (2001), pp. 280–301.

[10] A. JÜNGEL, *Quasi-hydrodynamic Semiconductor Equations*, Birkhäuser, Basel, 2001.

[11] S. KAWASHIMA AND Y. SHUZITA, *On the normal form of the symmetric hyperbolic-parabolic systems associated with the conservation laws*, Tohoku Math. J. (2), 40 (1988), pp. 449–464.

[12] J. U. KIM, *Smooth solutions to a quasi-linear system of diffusion equations for a certain population model*, Nonlinear Anal., 8 (1984), pp. 1121–1144.

[13] S. KNIES, *Schwache Lösungen von Halbleitergleichungen im Falle von Ladungstransport mit Streueffekten*, Ph.D. thesis, Universität Bonn, Germany, 1997.

[14] A. KUFNER, O. JOHN, AND S. FUČÍK, *Function Spaces*, Nordhoff, Leyden, The Netherlands, 1977.

[15] D. LE, *Cross-diffusion systems on n spatial dimensional domains*, in Proceedings of the Fifth Mississippi State Conference on Differential Equations and Computational Simulations, Electron. J. Differ. Equ. Conf. 10, Southwest Texas State University, San Marcos, TX, 2003, pp. 193–210.

[16] Y. LOU, S. MARTÌNEZ, AND W.-M. NI, *On 3×3 Lotka-Volterra competition systems with cross-diffusion*, Discrete Contin. Dynam. Systems, 6 (2000), pp. 175–190.

[17] Y. LOU AND W.-M. NI, *Diffusion, self-diffusion and cross-diffusion*, J. Differential Equations, 131 (1996), pp. 79–131.

[18] Y. LOU, W.-M. NI, AND Y. WU, *The global existence of solutions for a cross-diffusion system*, Adv. in Math. (China), 25 (1996), pp. 283–284.

[19] A. MARROCCO, *Numerical simulation of chemotactic bacteria aggregation via mixed finite elements*, M2AN Math. Model. Numer. Anal., 37 (2003), pp. 617–630.

[20] M. MIMURA AND K. KAWASAKI, *Spatial segregation in competitive interaction-diffusion equations*, J. Math. Biol., 9 (1980), pp. 49–64.

[21] M. POZIO AND A. TESEI, *Global existence of solutions for a strongly coupled quasilinear parabolic system*, Nonlinear Anal., 14 (1990), pp. 657–689.

[22] R. REDLINGER, *Existence of the global attractor for a strongly coupled parabolic system arising in population dynamics*, J. Differential Equations, 118 (1995), pp. 219–252.

[23] W. RUAN, *Positive steady-state solutions of a competing reaction-diffusion system with large cross-diffusion coefficients*, J. Math. Anal. Appl., 197 (196), pp. 558–578.

[24] K. RYU AND I. AHN, *Positive steady-states for two interacting species models with linear self-cross-diffusions*, Discrete Contin. Dynam. Systems, 9 (2003), pp. 1049–1081.

[25] N. Shigesada, K. Kawasaki, and E. Teramoto, *Spatial segregation of interacting species*, J. Theoret. Biol., 79 (1979), pp. 83–99.
[26] J. Simon, *Compact sets in the Space $L^p(0,T;B)$*, Ann. Mat. Pura Appl. (4), 146 (1987), pp. 65–96.
[27] Y. Wanli, *Global solutions to some quasilinear parabolic systems in population dynamics*, J. Partial Differential Equations, 12 (1999), pp. 193–200.
[28] Y. Wu, *Qualitative studies of solutions for some cross-diffusion systems*, in China-Japan Symposium on Reaction-Diffusion Equations and Their Applications and Computational Aspects, T.-T. Li, M. Mimura, Y. Nishiura, and Q.-X. Ye, eds., World Scientific, Singapore, 1997, pp. 177–187.
[29] A. Yagi, *Global solution to some quasilinear parabolic system in population dynamics*, Nonlinear Anal., 21 (1993), pp. 603–630.

SIAM J. MATH. ANAL.
Vol. 36, No. 1, pp. 323–346

ASYMPTOTIC NORMALITY OF SCALING FUNCTIONS*

LOUIS H. Y. CHEN†, TIM N. T. GOODMAN‡, AND S. L. LEE†

Abstract. The Gaussian function $G(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, which has been a classical choice for multiscale representation, is the solution of the scaling equation

$$G(x) = \int_{\mathbb{R}} \alpha G(\alpha x - y) dg(y), \quad x \in \mathbb{R},$$

with scale $\alpha > 1$ and absolutely continuous measure

$$dg(y) = \frac{1}{\sqrt{2\pi}(\alpha^2 - 1)} e^{-y^2/2(\alpha^2-1)} dy.$$

It is known that the sequence of normalized B-splines (B_n), where B_n is the solution of the scaling equation

$$\phi(x) = \sum_{j=0}^{n} \frac{1}{2^{n-1}} \binom{n}{j} \phi(2x - j), \quad x \in \mathbb{R},$$

converges uniformly to G. The classical results on normal approximation of binomial distributions and the uniform B-splines are studied in the broader context of normal approximation of probability measures m_n, $n = 1, 2, \dots$, and the corresponding solutions ϕ_n of the scaling equations

$$\phi_n(x) = \int_{\mathbb{R}} \alpha \phi_n(\alpha x - y) dm_n(y), \quad x \in \mathbb{R}.$$

Various forms of convergence are considered and orders of convergence obtained. A class of probability densities are constructed that converge to the Gaussian function faster than the uniform B-splines.

Key words. normal approximation, probability measures, scaling functions, uniform B-splines, asymptotic normality

AMS subject classifications. 41A15, 41A25, 41A39, 42C40, 65T60

DOI. 10.1137/S0036141002406229

1. Introduction.

The Gaussian function, $G(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, and its derivatives have been widely used in scale-space representation (see [1], [11], [18]). The uniform B-spline, B_n, which is the solution of the scaling equation

$$\phi(x) = \sum_{j=0}^{n} \frac{1}{2^{n-1}} \binom{n}{j} \phi(2x - j), \quad x \in \mathbb{R}, \tag{1.1}$$

associated with the binomial distribution $\frac{1}{2^n}\binom{n}{j}$, $j = 0, 1, \dots, n$, approximates the Gaussian and provides fast computational algorithms for practical implementation of Gaussian scale-space representation (see [15], [16]). The B-spline, B_n, is the probability density function of the sum of n copies of independent identically distributed

*Received by the editors April 23, 2002; accepted for publication (in revised form) October 3, 2003; published electronically July 14, 2004. This research was supported by the Wavelets Strategic Research Programme, National University of Singapore, under a grant from the National Science and Technology Board and the Ministry of Education, Singapore.
http://www.siam.org/journals/sima/36-1/40622.html
†Department of Mathematics, University of Singapore, 10 Kent Ridge Road, Singapore 119260 (lhychen@ims.nus.edu.sg, matleesl@nus.edu.sg).
‡Department of Mathematics, University of Dundee, Dundee DD1 4HN, Scotland, UK (tgoodman@mcs.dundee.ac.uk).

uniform random variables on the interval $[0,1)$. It is well known that the binomial distributions converge to the normal distribution in the sense that

$$\lim_{n\to\infty} \sum_{k=0}^{[x_n]} \frac{1}{2^n} \binom{n}{k} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt, \tag{1.2}$$

where $x_n = \sqrt{n}x/2 + n/2$, and it is also known that

$$\lim_{n\to\infty} \int_{-\infty}^{x_n'} B_n(t)dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt, \tag{1.3}$$

where $x_n' := \sqrt{n}x/2\sqrt{3} + n/2$. Further, the normalized B-splines converge uniformly on $\mathbb{R}$ to the Gaussian function (see [5] and [13]). In fact, Curry and Schoenberg [5] considered the more general class of Polya frequency functions as limits of nonuniform B-splines with arbitrary knots. The Gaussian function satisfies the integral scaling equation

$$G(x) = \int_{\mathbb{R}} \alpha G(\alpha x - y) dg(y), \quad x \in \mathbb{R},$$

where $\alpha > 1$ is a scaling constant and g is the absolutely continuous measure given by

$$dg(y) = \frac{1}{\sqrt{2\pi}(\alpha^2 - 1)} e^{-y^2/2(\alpha^2-1)} dy.$$

The Gaussian function and its derivatives and the modulated Gaussian have been used extensively in many applications such as scale-space analysis and computer vision (see [1], [11], [18]). The normal approximation of the binomial distributions and the uniform B-splines enables the binomial coefficients and B-splines to replace the Gaussian function in the Gaussian scale-space representation and vice versa (see [11], [15], [16]). The Gaussian function is optimal in time-frequency localization, amenable to statistical analysis, and provides an accurate model of human vision (see [18]). While inheriting approximately many of the rich properties of the Gaussian, the binomial distributions and B-splines have the added advantage of providing fast algorithms for practical computations.

We shall consider a sequence of scaling equations

$$\phi_n(x) = \int_{\mathbb{R}} \alpha \phi_n(\alpha x - y) dm_n(y), \quad x \in \mathbb{R},\ n = 1, 2, \dots, \tag{1.4}$$

where $\alpha > 1$ and (m_n) is a sequence of probability measures with finite first and second moments. It will be shown in the next section that for each n, (1.4) has a unique solution, which is also a probability measure. We shall call ϕ_n the *m_n-scaling function* and m_n its *filter*. If m_n is a discrete measure concentrated on the integers $\mathbb{Z}$ with mass $h_n(j)$ at $j \in \mathbb{Z}$, then (1.4) becomes the discrete scaling equation

$$\phi_n(x) = \sum_{j\in\mathbb{Z}} \alpha h_n(j) \phi_n(\alpha x - j), \quad x \in \mathbb{R}. \tag{1.5}$$

In particular, if m_n is the discrete measure concentrated on the set $\{0, 1, \dots, n\}$ with mass $\frac{1}{2^n}\binom{n}{j}$ at $j = 0, 1, \dots, n$ and scale $\alpha = 2$, then (1.5) reduces to (1.1). The

object of this paper is to investigate the approximation of the Gaussian function by probability measures and the corresponding scaling functions in the same way as the normal approximation by binomial and B-spline distributions and to construct sequences of distributions that converge to the Gaussian faster than the binomial and B-spline distributions.

Suppose that (m_n) is a sequence of probability measures on $\mathbb{R}$ with mean $\mu(m_n) = \mu_n$ and standard deviation $\sigma(m_n) = \sigma_n$, and define

$$\widetilde{m}_n(S) = m_n(\sigma_n S + \mu_n) \quad \text{for measurable } S \subset \mathbb{R},$$

or, equivalently,

$$\widehat{\widetilde{m}}_n(u) = e^{iu\mu_n/\sigma_n}\widehat{m}_n(u/\sigma_n), \quad u \in \mathbb{R}. \tag{1.6}$$

We say that (m_n) is *asymptotically normal* if for all $x \in \mathbb{R}$,

$$\lim_{n\to\infty} \int_{-\infty}^{x} d\widetilde{m}_n(t) = \int_{-\infty}^{x} G(t)dt. \tag{1.7}$$

If m_n is absolutely continuous, then by the Radon–Nikodym theorem, $dm_n(t) = f_n(t)dt$ for a probability density function f_n, and then $d\widetilde{m}_n(t) = \widetilde{f}_n(t)dt$, where

$$\widetilde{f}_n(t) = \sigma_n f_n(\sigma_n t + \mu_n).$$

The central limit theorem tells us that if m_n is the probability distribution for the sum of n independent, identically distributed random variables, then (m_n) is asymptotically normal. In the case that each such random variable is uniformly distributed on the interval $[0, 1)$, m_n has density function B_n, and the asymptotic normality is also implied by the convergence of the normalized B-splines discussed earlier. Now it is well known that asymptotic normality can be stated in terms of convergence of characteristic functions, i.e., Fourier transforms of the probability density functions. To be precise, (1.7) is equivalent to

$$\widehat{\widetilde{m}}_n(u) \to e^{-u^2/2} \text{ locally uniformly on } \mathbb{R}, \tag{1.8}$$

where local uniform convergence means convergence that is uniform on compact subsets. This result is given in [7, p. 249], and more modern expositions are given in [10] and [17].

In section 2, we show that if m is a probability measure on $\mathbb{R}$ with finite first moment, then the solution of the scaling equation

$$\phi(x) = \int_{\mathbb{R}} \alpha\phi(\alpha x - y)dm(y), \quad x \in \mathbb{R}, \tag{1.9}$$

is also a probability measure. In (1.9), and throughout the paper, α is a number larger than 1, which we call the *scale*. We remark that if the solution is absolutely continuous, then its probability density satisfies (1.9). If the solution ϕ is not absolutely continuous, then it satisfies (1.9) in the weak sense, i.e.,

$$\widehat{\phi}(u) = \widehat{m}(u/\alpha)\widehat{\phi}(u/\alpha), \quad u \in \mathbb{R}. \tag{1.10}$$

The following result puts in perspective the asymptotic normality exhibited by the binomial coefficients and the uniform B-splines.

THEOREM 1.1. *Let (m_n) be a sequence of probability measures on $\mathbb{R}$ with finite first and second moments and $(\widehat{\widehat{m}}_n{}')$ be uniformly bounded in a neighborhood of the origin. Then (m_n) is asymptotically normal if and only if the corresponding sequence of m_n-scaling functions is asymptotically normal.*

In order to study the asymptotic normality of scaling functions, we need only to study the asymptotic normality of their filters, because of Theorem 1.1. The binomial coefficients, which are the filters for the uniform B-splines, define a sequence of discrete probability measures that is asymptotically normal. It follows from Theorem 1.1 that the coefficients $b_{n,k}$ in the expansion

$$\left(\frac{1+z+\cdots+z^{\alpha-1}}{\alpha}\right)^n = \sum_{k=0}^{n(\alpha-1)} b_{n,k} z^k, \tag{1.11}$$

where the scale α is here an integer, also define a sequence of probability measures that is asymptotically normal. This is because the uniform B-splines are also the solution of the scaling equations with measures $m_n(k) = b_{n,k}$, $k = 0, 1, \ldots, n(\alpha-1)$, for any integer scale $\alpha > 1$. For such α, the roots of the polynomials on the left of (1.11) that generate $b_{n,k}$ are the complex αth roots of unity that are not equal to 1. The next theorem gives a general result that holds for a large class of polynomials including those with negative roots as well as those in (1.11).

THEOREM 1.2. *Let $\gamma \in [0, \pi/2)$, and define $D_\gamma = \{z \in \mathbb{C} : \text{ satisfies } (1.12)\}$:*

$$\left|\operatorname{Im}\left\{\frac{z}{(1+z)^2}\right\}\right| \le \tan\gamma \operatorname{Re}\left\{\frac{z}{(1+z)^2}\right\}. \tag{1.12}$$

For $n = 1, 2, \ldots$, take $r_{n,1}, \ldots, r_{n,n}$ in D_γ and define

$$\sum_{k=0}^{n} a_{n,k} z^k = \prod_{j=1}^{n} (z + r_{n,j})/(1 + r_{n,j}). \tag{1.13}$$

We also assume that the $r_{n,j}$, $n = 1, 2, \ldots$, $j = 1, \ldots, n$, are bounded away from -1, that the coefficients $a_{n,k}$, $n = 1, 2, \ldots$, $k = 0, \ldots n$, are real, and that

$$\sigma_n^2 = \sum_{j=1}^{n} r_{n,j}/(1 + r_{n,j})^2 \to \infty \text{ as } n \to \infty. \tag{1.14}$$

If m_n, $n = 1, 2, \ldots$, denote the discrete measures defined by $m_n(\{k\}) = a_{n,k}$, $k = 0, 1, \ldots, n$, it follows that $\widehat{\widehat{m}}_n(u) \to e^{-u^2/2}$ locally uniformly as $n \to \infty$. If, in addition, $a_{n,k} \ge 0$, $k = 0, 1, \ldots, n$, for all sufficiently large n, then (m_n) is asymptotically normal.

Remark 1. We remark that the first part of Theorem 1.2 does not require m_n to be a probability measure; i.e., some of the coefficients $a_{n,k}$ could be negative.

After some preliminary results in the next section, we shall prove Theorem 1.1 in section 3. A proof of Theorem 1.2 is given in section 4. We note that a special case of this result, when all $r_{n,j} > 0$, was proved earlier using probabilistic techniques [3]. The completely different analytic techniques, which we employ here, give considerably more general results. These techniques also allow us to analyze, in the remainder of section 4, the order of convergence in the frequency domain for both the measures m_n and the corresponding scaling functions. In particular we shall prove the following theorem.

THEOREM 1.3. *We assume the conditions of Theorem* 1.2 *and that* $a_{n,k} \geq 0$, $k = 0, 1, \ldots, n$.

(a) *Then*

$$\left|\left|\widehat{\widetilde{\phi}}_n - e^{-(\cdot)^2/2}\right|\right|_\infty = O(\sigma_n^{-1}).$$

(b) *If* $\sum_{k=0}^n a_{n,k} z^k$ *is a reciprocal polynomial, i.e.,* $a_{n,0} \neq 0$ *and* $a_{n,k} = a_{n,n-k}$, $k = 0, 1, \ldots, n$, *then*

$$\left|\left|\widehat{\widetilde{\phi}}_n - e^{-(\cdot)^2/2}\right|\right|_\infty = O(\sigma_n^{-2}).$$

(c) *If, in addition to the condition in* (b),

$$\sigma_n^{-1} \sum_{j=1}^n r_{n,j}(r_{n,j}^2 - 4r_{n,j} + 1)/(1 + r_{n,j})^4 \text{ is bounded,} \tag{1.15}$$

then

$$\left|\left|\widehat{\widetilde{\phi}}_n - e^{-(\cdot)^2/2}\right|\right|_\infty = O(\sigma_n^{-3}).$$

Asymptotic normality entails weak convergence in the time domain. We show in section 5 that, under mild conditions on the shape of the filters and the scaling functions, both the measures and the corresponding scaling functions converge uniformly in the time domain. The shape conditions are satisfied if $r_{n,j}$ are restricted to certain sectors of the complex plane, reminiscent of total positivity. It is noted that for a special case of the choice, when all $r_{n,j} > 0$, Chui and Wang [4] consider convergence of the scaling functions. However, their approach is different, and they do not consider the related convergence of the measures m_n. Finally, in the same section, we consider the order of convergence in the time domain and prove the following results.

THEOREM 1.4. *We assume the conditions of Theorem* 1.2 *and that all* $r_{n,j}$ *lie in the sector* $|\arg z| \leq \frac{\pi}{3}$. *Then as* $n \to \infty$,

$$\max_{k=0,\ldots,n} \left|\sigma_n a_{n,k} - G\left(\frac{k - \mu_n}{\sigma_n}\right)\right| = O(\sigma_n^{-\frac{1}{2}}),$$

and if $\sum_{k=0}^n a_{n,k} z^k$ *is reciprocal,*

$$\max_{k=0,\ldots,n} \left|\sigma_n a_{n,k} - G\left(\frac{k - \mu_n}{\sigma_n}\right)\right| = O(\sigma_n^{-\frac{2}{3}}).$$

We remark that in [3], this problem is considered, using probabilistic techniques, for the special case when $a_{n,0}, \ldots, a_{n,n}$ are the Eulerian numbers. In this case $\sigma_n = \sqrt{\pi(n+1)/6}$. Thus our result gives order of convergence $O(\sigma_n^{-\frac{2}{3}}) = O(n^{-\frac{1}{3}})$, while [3] shows only convergence $O(n^{-\frac{1}{4}})$.

THEOREM 1.5. *We assume the conditions of Theorem* 1.2, *that* $r_{n,j}$ *include* 1 *and all* $\mathrm{Re}(r_{n,j}) \geq 0$. *For* $n = 1, 2, \ldots$, *let* ϕ_n *denote the scaling function corresponding to the measure* $m_n(\{k\}) = a_{n,k}$, $k = 0, 1, \ldots, n$, *with scale* 2, *and define*

$$\widetilde{\phi}_n(x) = \sigma(\phi_n)\phi_n(\sigma(\phi_n)x + \mu(\phi_n)), \quad x \in \mathbb{R}.$$

Then

$$\|\widetilde{\phi}_n - G\|_\infty = O(\sigma_n^{-\frac{1}{2}}).$$

If $\sum_{n=0}^{n} a_{n,k} z^k$ *is reciprocal for large enough* n, *then*

$$\|\widetilde{\phi}_n - G\|_\infty = O(\sigma_n^{-1}).$$

If, in addition, (1.15) *is satisfied, then*

$$\|\widetilde{\phi}_n - G\|_\infty = O(\sigma_n^{-\frac{3}{2}}).$$

It is noted that certain sequences of scaling functions give a faster rate of convergence to the Gaussian than the uniform B-splines. Also on considering Theorems 1.3 and 1.5, it might be expected that the second part of Theorem 1.4 should give order of convergence $O(\sigma_n^{-1})$ instead of $O(\sigma_n^{-2/3})$ and that under the additional condition (1.15) we should obtain order $O(\sigma_n^{-3/2})$. We have been unable to prove orders better than $O(\sigma_n^{-2/3})$ due to a technical restriction in Lemma 5.4, and we do not know whether this restriction can be removed.

2. Probability measures and scaling equations. Consider the scaling equation (1.9) where m is a probability measure and, as before, α is a number (not necessarily an integer) satisfying $\alpha > 1$. We shall show that (1.9) has a unique solution, which is a probability measure. Further, if m has finite first and second moments, then the solution of (1.9) also has finite first and second moments. Equation (1.10) suggests that, when $\widehat{\phi}(0) = 1$, $\widehat{\phi}(u)$ is given by the infinite product (2.1) below but with n replaced by ∞. We remark that products of the form (2.1) occur in the study of groups of transformations in Hilbert space (see, for example, [6, section 38]). For the case when ϕ is the B-spline B_n and $\alpha = 2$, this reduces to the classical formula of Viète:

$$\sin x/x = \prod_{j=1}^{\infty} \cos(x/2^j).$$

So as a preliminary result we need to consider the convergence of (2.1) in Lemma 2.1 below.

LEMMA 2.1. *Suppose that* m *is a probability measure with finite first moment. Then the products*

$$\prod_{j=1}^{n} \widehat{m}(u/\alpha^j), \quad u \in \mathbb{R}, \tag{2.1}$$

converge locally uniformly as $n \to \infty$.

Proof. Since m is a probability measure, $|\widehat{m}(u)| \le 1$ for all $u \in \mathbb{R}$. Then for every nonnegative integer n and all u,

$$\left| \prod_{j=1}^{n} \widehat{m}(u/\alpha^j) \right| \le 1 \quad \text{for all } u \in \mathbb{R}.$$

Also, since m has finite first moment, $\widehat{m}'$ is bounded, and so

$$\left|\widehat{m}(u/\alpha^j) - 1\right| \le C|u|/\alpha^j, \quad j = 1, 2, \dots,$$

for a constant $C > 0$. Thus for integers $n > \ell$,

$$\left| \prod_{j=1}^{\ell} \widehat{m}(u/\alpha^j) - \prod_{j=1}^{n} \widehat{m}(u/\alpha^j) \right| \le \sum_{j=1}^{n-\ell} |1 - \widehat{m}(u/\alpha^{\ell+j})| \le C|u|(\alpha^{-\ell} - \alpha^{-n})/(\alpha - 1),$$

which tends to zero uniformly on compact subsets of $\mathbb{R}$ as $\ell, n \to \infty$. Therefore, the product $\prod_{j=1}^{n} \widehat{m}(u/\alpha^j)$ converges uniformly on compact sets as $n \to \infty$. □

PROPOSITION 2.2. *If m is a probability measure with finite first and second moments, then the scaling equation* (1.9) *has a unique solution ϕ, which is also a probability measure with finite first and second moments. Further,*

$$\mu(\phi) = (\alpha - 1)^{-1}\mu(m) \quad \text{and} \quad \sigma(\phi)^2 = (\alpha^2 - 1)^{-1}\sigma(m)^2. \tag{2.2}$$

Proof. Choose a nonnegative initial function $f_0 \in C(\mathbb{R})$ with compact support and $\widehat{f_0}(0) = 1$, and for $n = 1, 2, \ldots$ define

$$f_n(x) = \int_{\mathbb{R}} \alpha f_{n-1}(\alpha x - y) dm(y), \quad x \in \mathbb{R}. \tag{2.3}$$

Then

$$\widehat{f_n}(u) = \widehat{f}_{n-1}(u/\alpha)\widehat{m}(u/\alpha) = \prod_{j=1}^{n} \widehat{m}(u/\alpha^j)\widehat{f_0}(u/\alpha^n), \quad u \in \mathbb{R}. \tag{2.4}$$

Further, f_n is nonnegative, and $\widehat{f_n}(0) = 1$ for $n = 0, 1, \ldots$. Therefore, f_n defines a sequence of probability measures $\mu_n \in C_0(\mathbb{R})^*$, where $d\mu_n(x) = f_n(x)dx$ and $C_0(\mathbb{R})^*$ is the dual of the space $C_0(\mathbb{R})$ of continuous functions that vanish at infinity. Therefore, $\widehat{\mu}_n = \widehat{f_n},\ n = 0, 1, \ldots$. Since the unit ball in $C_0(\mathbb{R})^*$ is weak* compact, there exist a subsequence μ_{n_ℓ} and a probability measure ϕ on $\mathbb{R}$ such that $\mu_{n_\ell} \to \phi$ as $\ell \to \infty$ in the weak* topology. It follows (see [7, p. 249]) that $\widehat{\mu}_{n_\ell}$ converges locally uniformly to $\widehat{\phi}$ as $n \to \infty$. By Lemma 2.1 and (2.4),

$$\widehat{\phi}(u) = \prod_{j=1}^{\infty} \widehat{m}(u/\alpha^j), \quad u \in \mathbb{R},$$

which satisfies (1.10).

Define

$$\Pi_n(u) := \prod_{j=1}^{n} \widehat{m}(u/\alpha^j), \quad u \in \mathbb{R}. \tag{2.5}$$

Then

$$\Pi_n(u) \to \widehat{\phi}(u) \ \text{ locally uniformly on } \mathbb{R}, \tag{2.6}$$

where ϕ is the solution of (1.9). We shall show that ${\Pi_n}'$ converges uniformly in a neighborhood of the origin. Since $\widehat{m}(0) = 1$, there exists a closed disc D centered at the origin such that $\widehat{m}(u) \neq 0$ for all $u \in D$. Differentiating (2.5) gives

$${\Pi_n}'(u) = \prod_{j=1}^{n} \widehat{m}(u/\alpha^j) \sum_{j=1}^{n} \frac{1}{\alpha^j} \frac{\widehat{m}\,'(u/\alpha^j)}{\widehat{m}(u/\alpha^j)}, \tag{2.7}$$

which shows that $\Pi_n{}'$ is uniformly convergent on D. It follows that $\widehat{\phi}\,'$ exists and $\Pi_n{}'$ converges uniformly to $\widehat{\phi}\,'$ on D. Hence $\widehat{\phi}\,'$ is continuous on D, and

$$\widehat{\phi}\,'(0) = (\alpha-1)^{-1}\widehat{m}\,'(0). \tag{2.8}$$

Differentiating (2.7) gives

$$\begin{aligned}\Pi_n{}''(u) = {} & \prod_{j=1}^{n}\widehat{m}(u/\alpha^j)\left(\sum_{j=1}^{n}\frac{1}{\alpha^j}\frac{\widehat{m}\,'(u/\alpha^j)}{\widehat{m}(u/\alpha^j)}\right)^2 \\ & + \prod_{j=1}^{n}\widehat{m}(u/\alpha^j)\sum_{j=1}^{n}\frac{1}{\alpha^{2j}}\frac{\widehat{m}\,''(u/\alpha^j)\widehat{m}(u/\alpha^j)-\widehat{m}\,'(u/\alpha^j)^2}{\widehat{m}(u/\alpha^j)^2},\end{aligned} \tag{2.9}$$

which shows that $\Pi_n{}''$ is uniformly convergent on D. Thus $\widehat{\phi}\,''$ exists and is continuous on D. A straightforward computation using (2.9) leads to

$$\widehat{\phi}\,''(0) = \frac{1}{(\alpha^2-1)}\left\{\widehat{m}\,''(0) + \frac{2\widehat{m}\,'(0)^2}{(\alpha-1)}\right\}. \tag{2.10}$$

It follows that ϕ has finite first and second moments, and the relationships (2.2) follow from (2.8) and (2.10). □

3. Proof of Theorem 1.1. We shall prove a slightly stronger result than that of Theorem 1.1. This result is contained in Theorem 3.1.

THEOREM 3.1. *Let (m_n) be a sequence of probability measures on $\mathbb{R}$ with finite first and second moments, and $(\widehat{\widehat{m}}_n{}')$ is uniformly bounded in a neighborhood of 0. Then the following are equivalent:*

(a) $\widehat{\widehat{m}}_n(u) \to e^{-u^2/2}$ *locally uniformly on* $\mathbb{R}$ *as* $n \to \infty$.
(b) $\widehat{\widehat{\phi}}_n(u) \to e^{-u^2/2}$ *locally uniformly on* $\mathbb{R}$ *as* $n \to \infty$.
(c) (m_n) *is asymptotically normal.*
(d) (ϕ_n) *is asymptotically normal.*

Further, if (a) *holds locally uniformly on* $\mathbb{R}$, *then* (b) *holds uniformly on* $\mathbb{R}$.

Proof. By Proposition 2.2, for each $n = 0, 1, \ldots$, (1.4) has a unique solution ϕ_n, which is also a probability measure with finite first and second order moments, and

$$\mu(m_n) = (\alpha-1)\mu(\phi_n) \quad \text{and} \quad \sigma(m_n)^2 = (\alpha^2-1)\sigma(\phi_n)^2. \tag{3.1}$$

By (1.6), (1.10), and (3.1),

$$\widehat{\widehat{\phi}}_n(u) = \widehat{\widehat{m}}_n(\alpha^{-1}\sqrt{\alpha^2-1}\,u)\widehat{\widehat{\phi}}_n(\alpha^{-1}u), \quad u \in \mathbb{R}. \tag{3.2}$$

Iterating (3.2) leads to

$$\widehat{\widehat{\phi}}_n(u) = \prod_{j=1}^{\infty}\widehat{\widehat{m}}_n(\alpha^{-j}\sqrt{\alpha^2-1}\,u), \quad u \in \mathbb{R}, \tag{3.3}$$

where the infinite product on the right converges locally uniformly on $\mathbb{R}$ and uniformly in n, since $(\widehat{\widehat{m}}_n{}')$ is uniformly bounded in a neighborhood of 0.

If (a) holds, then by (3.3) we have

$$\lim_{n\to\infty} \widehat{\widetilde{\phi}}_n(u) = \lim_{n\to\infty} \prod_{j=1}^{\infty} \widehat{\widetilde{m}}_n(\alpha^{-j}\sqrt{\alpha^2-1}\, u)$$
$$= \prod_{j=1}^{\infty} e^{-(\alpha^2-1)u^2/2\alpha^{2j}} = e^{-u^2/2}, \quad u \in \mathbb{R}.$$

Conversely, if $\lim_{n\to\infty} \widehat{\widetilde{\phi}}_n(u) = e^{-u^2/2}$, then by (3.2)

$$\widehat{\widetilde{m}}_n(u) = \frac{\widehat{\widetilde{\phi}}_n(\alpha\, u/\sqrt{\alpha^2-1})}{\widehat{\widetilde{\phi}}_n(u/\sqrt{\alpha^2-1})}, \quad u \in \mathbb{R},$$

for sufficiently large n. It follows that

$$\lim_{n\to\infty} \widehat{\widetilde{m}}_n(u) = \frac{e^{-\alpha^2 u^2/2(\alpha^2-1)}}{e^{-u^2/2(\alpha^2-1)}} = e^{-u^2/2}, \quad u \in \mathbb{R}.$$

A similar argument shows that (a) holds locally uniformly on $\mathbb{R}$ if and only if (b) holds locally uniformly on $\mathbb{R}$.

Now suppose that (a) holds uniformly on compact subsets of $\mathbb{R}$. Note that for any $u \in \mathbb{R}$ and $n \geq 1$,

$$|\widehat{\widetilde{m}}_n(u)| = \left| \int_{-\infty}^{-\infty} e^{-iux} dm_n(x) \right| \leq \int_{-\infty}^{-\infty} dm_n(x) = 1.$$

So for any $k \geq 1$,

$$\left| \widehat{\widetilde{\phi}}_n(u) \right| = \prod_{j=1}^{\infty} \left| \widehat{\widetilde{m}}_n(\alpha^{-j}\sqrt{\alpha^2-1}\, u) \right|$$
$$\leq \prod_{j=k+1}^{\infty} \left| \widehat{\widetilde{m}}_n(\alpha^{-j}\sqrt{\alpha^2-1}\, u) \right|$$
$$= \left| \widehat{\widetilde{\phi}}_n(\alpha^{-k}\, u) \right|, \quad u \in \mathbb{R}. \tag{3.4}$$

For any $\epsilon > 0$, we choose $A > 0$ and integer N so that $e^{-A^2/2} < \epsilon$ and

$$\left| \widehat{\widetilde{\phi}}_n(u) - e^{-u^2/2} \right| < \epsilon, \quad |u| \leq \alpha A,\ n > N.$$

Take any u with $|u| > A$. Then there is a nonnegative integer k such that $A < \alpha^{-k}|u| \leq \alpha A$, and so

$$e^{-(\alpha^{-k}u)^2/2} < e^{-A^2/2} < \epsilon.$$

Also for $n > N$, $|\widehat{\widetilde{\phi}}_n(\alpha^{-k}u) - e^{-(\alpha^{-k}u)^2/2}| < \epsilon$, and so

$$\left| \widehat{\widetilde{\phi}}_n(u) \right| \leq \left| \widehat{\widetilde{\phi}}_n(\alpha^{-k}\, u) \right| < 2\epsilon.$$

Since $e^{-u^2/2} < e^{-A^2/2} < \epsilon$, it follows that $|\widehat{\widetilde{\phi}}_n(u) - e^{-u^2/2}| < 3\epsilon$. Thus for all $n > N$ and $u \in \mathbb{R}$, $|\widehat{\widetilde{\phi}}_n(u) - e^{-u^2/2}| < 3\epsilon$, and hence (b) holds uniformly on $\mathbb{R}$.

Recall that the asymptotic normality of a sequence of distribution functions is equivalent to the local uniform convergence of their characteristic functions (see, for instance, [7, p. 249]). □

We remark that if (m_n) is a sequence of discrete probability measures on $\mathbb{Z}$ with finite first and second moments, then the condition that $(\widehat{\widetilde{m}}_n{}')$ be uniformly bounded in a neighborhood of 0 is automatically satisfied. The following lemma gives a slightly stronger result.

LEMMA 3.2. *If (m_n) is a sequence of discrete probability measures on $\mathbb{Z}$ with finite first and second moments, then $(\widehat{\widetilde{m}}_n{}')$ is uniformly bounded on any compact subset of $\mathbb{R}$.*

Proof. Let $m_n(\{k\}) = b_{n,k} \geq 0$, $n = 1, 2, \ldots$, $k \in \mathbb{Z}$, where $\sum_{k=-\infty}^{\infty} b_{n,k} = 1$. As before, we write

$$\mu_n := \sum_{k=-\infty}^{\infty} k b_{n,k} \quad \text{and} \quad \sigma_n^2 := \sum_{k=-\infty}^{\infty} (k - \mu_n)^2 b_{n,k}.$$

Then

$$\widehat{\widetilde{m}}_n(u) = \sum_{k=-\infty}^{\infty} b_{n,k} e^{i(\mu_n - k)u/\sigma_n},$$

and so

$$\begin{aligned} \widehat{\widetilde{m}}_n{}'(u) &= \frac{i}{\sigma_n} \sum_{k=-\infty}^{\infty} b_{n,k}(\mu_n - k) e^{i(\mu_n - k)u/\sigma_n} \\ &= \frac{i}{\sigma_n} \sum_{k=-\infty}^{\infty} b_{n,k}(\mu_n - k)(e^{i(\mu_n - k)u/\sigma_n} - 1). \end{aligned}$$

Since $|e^{iu} - 1| \leq 2|u|$ for all $u \in \mathbb{R}$,

$$\left| \widehat{\widetilde{m}}_n{}'(u) \right| \leq \frac{2|u|}{\sigma_n^2} \sum_{k=-\infty}^{\infty} (k - \mu_n)^2 b_{n,k} = 2|u|. \qquad \square$$

COROLLARY 3.3. *Let (m_n) be a sequence of discrete probability measures on $\mathbb{Z}$ with finite first and second moments. Then (m_n) is asymptotically normal if and only if the corresponding sequence of m_n-scaling functions with scale α is asymptotically normal.*

4. Convergence in the frequency domain. In order to apply Theorem 1.1 to study the asymptotic normality of scaling functions, we need first to study the asymptotic normality of their filters. We begin with a proof of Theorem 1.2.

Proof of Theorem 1.2. Let

$$\sum_{k=0}^{n} a_{n,k} z^k = \prod_{j=1}^{n} (p_{n,j} z + q_{n,j}), \tag{4.1}$$

where $q_{n,j} = 1 - p_{n,j}$. Then

$$\widehat{m}_n(u) = \prod_{j=1}^{n} (p_{n,j} e^{-iu} + q_{n,j})$$

and

$$\widehat{\widehat{m}}_n(u) = e^{iu\mu_n/\sigma_n} \prod_{j=1}^{n} (p_{n,j} e^{-iu/\sigma_n} + q_{n,j}),$$

where

$$\mu_n = \mu(m_n) = \sum_{j=1}^{n} p_{n,j}, \tag{4.2}$$

and

$$\sigma_n^2 = \sigma(m_n)^2 = \sum_{j=1}^{n} p_{n,j} q_{n,j}. \tag{4.3}$$

Therefore,

$$\log \widehat{\widehat{m}}_n(u) = \frac{iu\mu_n}{\sigma_n} + \sum_{j=1}^{n} F\left(p_{n,j}, \frac{-iu}{\sigma_n}\right), \tag{4.4}$$

where

$$F(p,t) = \log(pe^t + q), \quad q = 1 - p.$$

By induction, for $n = 2, 3, \ldots$,

$$F^{(n)}(p,t) := \frac{\partial^n}{\partial t^n} F(p,t) = (pe^t + q)^{-n} pq \sum_{j=0}^{n-2} (-1)^j c_n(j) p^j q^{n-2-j} e^{(j+1)t}, \tag{4.5}$$

where $c_2(j) = \delta_0(j)$, $j \in \mathbb{Z}$, and for $n = 2, 3, \ldots$, c_n satisfies the recursive relation

$$c_{n+1}(j) = (j+1)c_n(j) + (n-j)c_n(j-1), \quad j \in \mathbb{Z}. \tag{4.6}$$

From (4.6) we have $\sum_{j=-\infty}^{\infty} c_{n+1}(j) = n \sum_{j=-\infty}^{\infty} c_n(j)$, and since $\sum_{j=-\infty}^{\infty} c_2(j) = 1$, we have

$$\sum_{j=-\infty}^{\infty} c_n(j) = (n-1)!, \quad n = 2, 3. \ldots . \tag{4.7}$$

By (4.5) the Taylor series of $F(p,t)$ is given by

$$F(p,t) = \sum_{\nu=0}^{\infty} a_\nu(p) t^\nu, \tag{4.8}$$

where

$$a_0(p) = 0, \ a_1(p) = p, \ a_2(p) = \frac{1}{2} pq, \tag{4.9}$$

and for $\nu = 3, 4, \ldots$,

$$a_\nu(p) = \frac{pq}{\nu!} \sum_{k=0}^{\nu-2} (-1)^k c_\nu(k) p^k q^{\nu-2-k}. \tag{4.10}$$

By (4.4) and (4.8),

$$\log \widehat{\widetilde{m}}_n(u) = \frac{iu\mu_n}{\sigma_n} + \sum_{j=1}^{n} \sum_{\nu=0}^{\infty} a_\nu(p_{n,j}) \sigma_n^{-\nu} (-iu)^\nu. \tag{4.11}$$

By (4.2), (4.3), and (4.9),

$$\sum_{j=1}^{n} a_1(p_{n,j}) \sigma_n^{-1} (-iu) = -\frac{iu\mu_n}{\sigma_n},$$

$$\sum_{j=1}^{n} a_2(p_{n,j}) \sigma_n^{-2} (-iu)^2 = -\frac{u^2}{2},$$

so that (4.11) becomes

$$\log \widehat{\widetilde{m}}_n(u) = -\frac{u^2}{2} + \sum_{\nu=3}^{\infty} \sigma_n^{-\nu} (-iu)^\nu \sum_{j=1}^{n} a_\nu(p_{n,j}). \tag{4.12}$$

Now $r_{n,j} \in D_\gamma$ if and only if

$$\left| \mathrm{Im} \left\{ \frac{r_{n,j}}{(1+r_{n,j})^2} \right\} \right| \le \tan\gamma \ \mathrm{Re} \left\{ \frac{r_{n,j}}{(1+r_{n,j})^2} \right\}$$

or

$$|\mathrm{Im}\,(p_{n,j}q_{n,j})| \le \tan\gamma \ \mathrm{Re}\,(p_{n,j}q_{n,j})\,.$$

Therefore,

$$|p_{n,j}q_{n,j}| \le \sec\gamma \ \mathrm{Re}\,(p_{n,j}q_{n,j})\,. \tag{4.13}$$

On the other hand, $r_{n,j}$ being bounded away from -1 is equivalent to

$$|p_{n,j}| \le A - 1, \qquad n = 1, 2, \ldots, \quad j = 1, 2, \ldots, n, \tag{4.14}$$

for some constant A. By (4.10), (4.13), and (4.14),

$$\begin{aligned} |a_\nu(p_{n,j})| &\le \frac{|p_{n,j}q_{n,j}|}{\nu!} \sum_{k=0}^{\nu-2} c_\nu(k) |p_{n,j}|^k |q_{n,j}|^{\nu-2-k} \\ &\le \sec\gamma \ \mathrm{Re}(p_{n,j}q_{n,j}) A^{\nu-2}/\nu. \end{aligned} \tag{4.15}$$

By (4.12) and (4.15),

$$\begin{aligned} \left| \log \widehat{\widetilde{m}}_n(u) + \frac{u^2}{2} \right| &\le \sec\gamma \ \sum_{\nu=3}^{\infty} \frac{\sigma_n^{-\nu}|u|^\nu}{\nu} \sum_{j=1}^{n} \mathrm{Re}(p_{n,j}q_{n,j}) A^{\nu-2} \\ &\le \sec\gamma \ \sum_{\nu=3}^{\infty} \frac{|u|^\nu}{\nu} \left(\frac{A}{\sigma_n} \right)^{\nu-2} \\ &\le \sec\gamma \ \frac{A|u|^3}{\sigma_n} \left(1 - \frac{A|u|}{\sigma_n} \right)^{-1} \end{aligned} \tag{4.16}$$

whenever $A|u| < \sigma_n$. Since $\sigma_n \to \infty$ as $n \to \infty$, taking the limits as $n \to \infty$, (4.16) gives $\lim_{n\to\infty} \widehat{\widehat{m}}_n(u) = e^{-u^2/2}$ locally uniformly. □

Recall that the region D_γ in Theorem 1.2 comprises all $z \in \mathbb{C}$ satisfying

$$\left|\operatorname{Im}\left\{\frac{z}{(1+z)^2}\right\}\right| \le \tan\gamma \operatorname{Re}\left\{\frac{z}{(1+z)^2}\right\}.$$

It can be seen that D_γ contains the sector $|\arg z| \le \gamma$, and for $z = \pm re^{i\theta}$, $r > 0$, $\gamma \le \theta \le \pi$, (1.12) is equivalent to

$$\frac{\sin(\frac{\theta-\gamma}{2})}{\sin(\frac{\theta+\gamma}{2})} \le r \le \frac{\sin(\frac{\theta+\gamma}{2})}{\sin(\frac{\theta-\gamma}{2})}.$$

In particular D_γ contains the unit circle $r = 1$.

For the special case of Theorem 1.2, when all $r_{n,j} > 0$, the result was proved using probabilistic methods in [3] and [12]. Our analytic techniques allow us not only to prove asymptotic normality for a much larger class of measures but also, in the next result, to give information on the order of convergence in the frequency domain.

PROPOSITION 4.1. *We assume the conditions of Theorem* 1.2 *(except that we do not require* $a_{n,k} \ge 0$, $k = 0, 1, \dots, n$*). As before,*

$$\sigma_n^2 = \sum_{j=1}^{n} \frac{r_{n,j}}{(1+r_{n,j})^2}.$$

Then there is a constant $K > 0$ *so that for* $S_n := \{u : |u| \le K\sigma_n\}$ *the following hold.*

(a) *There is a constant* B *such that*

$$\left|\widehat{\widehat{m}}_n(u) - e^{-u^2/2}\right| \le B\sigma_n^{-1}, \quad u \in S_n,\ n = 1, 2, \dots . \tag{4.17}$$

(b) *If* $\sum_{k=0}^{n} a_{n,k} z^k$ *is a reciprocal polynomial, then there is a constant* C *such that*

$$\left|\widehat{\widehat{m}}_n(u) - e^{-u^2/2}\right| \le C\sigma_n^{-2}, \quad u \in S_n,\ n = 1, 2, \dots . \tag{4.18}$$

(c) *Finally, if in addition to the condition in* (b), (1.15) *is satisfied, then there is a constant* D *such that*

$$\left|\widehat{\widehat{m}}_n(u) - e^{-u^2/2}\right| \le D\sigma_n^{-3}, \quad u \in S_n,\ n = 1, 2, \dots . \tag{4.19}$$

Proof. (a) From (4.16) we see that for $|u| \le \frac{1}{4}A^{-1}\cos\gamma\ \sigma_n$,

$$\log \widehat{\widehat{m}}_n(u) + \frac{u^2}{2} \le \sec\gamma\ \frac{4A|u|^3}{3\sigma_n} \le \frac{1}{3}u^2,$$

and so $\log \widehat{\widehat{m}}_n(u) \le -\frac{1}{6}u^2$. By the mean value theorem,

$$\begin{aligned}\left|\widehat{\widehat{m}}_n(u) - e^{-u^2/2}\right| &\le e^{-u^2/6}\left|\log \widehat{\widehat{m}}_n(u) + \frac{u^2}{2}\right| \\ &\le \sec\gamma \left(\frac{4A|u|^3}{3\sigma_n}\right) e^{-u^2/6} \\ &\le B\sigma_n^{-1}\end{aligned}$$

for a constant B, which gives (4.17).

(b) We note from (4.6) and (4.10) that

$$a_3(p) = \frac{pq}{3!}(q-p), \tag{4.20}$$

$$a_4(p) = \frac{pq}{4!}(q^2 - 4pq + p^2). \tag{4.21}$$

Suppose that $P_n(z) = \sum_{k=0}^n a_{n,k} z^k$ is a reciprocal polynomial. Then $P_n(z) = 0$ if and only if $P_n(z^{-1}) = 0$. Noting that if $r_{n,j} = r_{n,k}^{-1}$, then $p_{n,j} = q_{n,k}$ and $q_{n,j} = p_{n,k}$, it follows that

$$\sum_{j=1}^n a_3(p_{n,j}) = 0. \tag{4.22}$$

So from (4.12) and (4.15),

$$\left| \log \widehat{\widetilde{m}}_n(u) + \frac{u^2}{2} \right| \le \sec\gamma \ \frac{A^2|u|^4}{\sigma_n^2} \ \left(1 - \frac{A|u|}{\sigma_n}\right)^{-1}$$

whenever $A|u| < \sigma_n$. Then (4.18) follows in a similar manner as before.

(c) Finally, we assume (1.15). Then (4.12), (4.21), (4.22), and (4.15) give (4.19). □

We note that $r^2 - 4r + 1 = 0$ when $r = 2 \pm \sqrt{3}$, and so (1.15) requires that in some sense the roots of $P_n(z) := \sum_{k=0}^n \ a_{n,k} z^k$ are close to $-2 \pm \sqrt{3}$. In particular, (1.15) will be satisfied if

$$P_n(z) = Q_{\ell_n}(z)(z^2 + 4z + 1)^{k_n},$$

where Q_{ℓ_n} is a reciprocal polynomial of degree $\ell_n = n - 2k_n$ and $n^{-1/2}\ell_n$ is bounded over n. In this case (4.19) takes the form

$$\left| \widehat{\widetilde{m}}_n(u) - e^{-u^2/2} \right| \le Cn^{-3/2}, \quad u \in S_n, \ n = 1, 2, \dots \ .$$

We now consider the order of convergence of the normalized m_n-scaling functions $\widetilde{\phi}_n$ as in Theorem 1.1, again in the frequency domain. From (3.3) it follows as in (4.4) that

$$\log \widehat{\widetilde{\phi}}_n(u) = \frac{iu\mu_n}{\sigma_n} + \sum_{j=1}^\infty \sum_{k=1}^n \ F\left(p_{n,k}, -\frac{iu}{\alpha^j \sigma_n}\right)$$

and as in (4.12) that

$$\log \widehat{\widetilde{\phi}}_n(u) = -\frac{u^2}{2} + \sum_{\nu=3}^\infty \ \frac{(-iu)^\nu}{(\alpha^2 - 1)\sigma_n^\nu} \ \sum_{j=1}^n a_\nu(p_{n,j}).$$

So as in (4.16) there is a constant A with

$$\left| \log \widehat{\widetilde{\phi}}_n(u) + \frac{u^2}{2} \right| \le \frac{A|u|^3}{\sigma_n} \left(1 - \frac{A|u|}{\sigma_n}\right)^{-1}$$

whenever $A|u| < \sigma_n$. By the mean value theorem, for $A|u| < \frac{1}{2}\sigma_n$,

$$\left|\widehat{\widetilde{\phi}}_n(u) - e^{-u^2/2}\right| \le \left\{e^{-u^2/2} + \left|\widehat{\widetilde{\phi}}(u) - e^{-u^2/2}\right|\right\} \frac{2A|u|^3}{\sigma_n},$$

and so

$$\begin{aligned}\left|\widehat{\widetilde{\phi}}(u) - e^{-u^2/2}\right| &\le e^{-u^2/2}\, \frac{2A|u|^3}{\sigma_n}\left(1 - \frac{2A|u|^3}{\sigma_n}\right)^{-1} \\ &\le e^{-u^2/2}\frac{4A|u|^3}{\sigma_n} \\ &\le B\sigma_n^{-1}\end{aligned}$$

if $|u|^3 < \sigma_n/4A$ for some constant B.

Similarly, if P_n is a reciprocal polynomial, then as in the derivation of (4.18), there are constants $A, B > 0$ such that

$$\left|\widehat{\widetilde{\phi}}_n(u) - e^{-u^2/2}\right| \le B\sigma_n^{-2}$$

whenever $|u| < A\sigma_n^{1/2}$. Finally, if (1.15) is satisfied, then there are constants $A, B > 0$ with

$$\left|\widehat{\widetilde{\phi}}_n(u) - e^{-u^2/2}\right| \le B\sigma_n^{-3}$$

whenever $|u| < A\sigma_n^{3/5}$.

To extend these estimates to all of $\mathbb{R}$ we need the following result.

LEMMA 4.2. *Suppose that m_n is a probability measure, $n = 1, 2, \dots$, and there is a sequence (β_n) with $\lim \beta_n = 0$ so that*

$$\left|\widehat{\widetilde{\phi}}_n(u) - e^{-u^2/2}\right| < \beta_n$$

whenever $|u| \le A|\log \beta_n|$ for some $A > 0$. Then

$$\overline{\lim}_{n\to\infty}\, \beta_n^{-1}\|\widehat{\widetilde{\phi}}_n - e^{-(\cdot)^2/2}\|_\infty \le 1.$$

Proof. Take $0 < \epsilon < 1$. Choose n large enough so that

$$2|\log(\beta_n\epsilon)| < \alpha^{-2}A^2\, |\log \beta_n|^2.$$

Take any u in $\mathbb{R}$ with $|u| > A|\log \beta_n|$. Then for some integer $k \ge 1$,

$$\alpha^{-1}A|\log \beta_n| < \alpha^{-k}|u| \le A|\log \beta_n|.$$

Putting $v = \alpha^{-k}|u|$, we have

$$v^2 > \alpha^{-2}A^2|\log \beta_n|^2 > 2|\log(\beta_n\epsilon)|,$$

and so

$$e^{-v^2/2} < \beta_n\epsilon.$$

Since $|\widehat{\widetilde{\phi}}_n(v) - e^{-v^2/2}| < \beta_n$, recalling (3.4) gives

$$\left|\widehat{\widetilde{\phi}}_n(u)\right| \le \left|\widehat{\widetilde{\phi}}(v)\right| < \beta_n(1+\epsilon).$$

Also $e^{-u^2/2} < e^{-v^2/2} < \beta_n\epsilon$, and so

$$\left|\widehat{\widetilde{\phi}}_n(u) - e^{-u^2/2}\right| < \beta_n(1+2\epsilon).$$

For any u with $|u| \le A|\log\beta_n|$ we have $|\widehat{\widetilde{\phi}}_n(u) - e^{-u^2/2}| < \beta_n$, and thus $\|\widehat{\widetilde{\phi}}_n - e^{-(\cdot)^2/2}\|_\infty \le \beta_n(1+2\epsilon)$ for all $u \in \mathbb{R}$. The result follows. □

Proof of Theorem 1.3. Theorem 1.3 follows from Lemma 4.2 and the preceding discussions. □

5. Convergence in the time domain. From Theorems 1.1 and 1.2 we can deduce the convergence of $\widetilde{m}_n$ and $\widetilde{\phi}_n$ to the Gaussian function G in the time domain only in the weak sense of (1.7). In this section we shall show that under mild assumptions on $(r_{n,j})$ in Theorem 1.2, both $\widetilde{m}_n$ and $\widetilde{\phi}_n$ have a "nice" shape, which ensures that the convergence is uniform. We consider two possibilities for the shape. For a continuous function ψ, we say ψ is *bell-shaped* if $\psi \ge 0$, $\lim_{x\to\pm\infty}\psi(x) = 0$, and there are two points $\alpha < \beta$ such that ψ is convex on $(-\infty, \alpha]$ and $[\beta, \infty)$ and concave on $[\alpha, \beta]$. We say that ψ is *logconcave* if it is supported on a closed interval, $\psi > 0$, and $\log\psi$ is concave on its interior. Neither of these properties implies the other. We note that in both cases there is a point γ such that ψ is increasing on $(-\infty, \gamma]$ and decreasing on $[\gamma, \infty)$. We also note that logconcavity is equivalent to *total positivity* of order 2, which says that for any $x_1 < x_2$ and $y_1 < y_2$,

$$\begin{vmatrix} \psi(x_1 - y_1) & \psi(x_1 - y_2) \\ \psi(x_2 - y_1) & \psi(x_2 - y_2) \end{vmatrix} \ge 0.$$

The following lemma shows that for a sequence of bell-shaped or logconcave functions, asymptotic normality implies uniform convergence. The result was stated in [5] for the case of logconcave functions, but no proof was given.

LEMMA 5.1. *Suppose that* (g_n) *is a sequence of continuous functions with* $\int_{-\infty}^{\infty} g_n = 1$, *which are either bell-shaped or logconcave, and for each* $x \in \mathbb{R}$,

$$\lim_{n\to\infty}\int_{-\infty}^{x} g_n = \int_{-\infty}^{x} G. \tag{5.1}$$

Then g_n *converges to* G *uniformly on* $\mathbb{R}$.

Proof. By (5.1), for any interval $I \subset \mathbb{R}$,

$$\lim_{n\to\infty}\int_I g_n = \int_I G. \tag{5.2}$$

Take $\epsilon > 0$. Then

$$\lim_{n\to\infty}\int_{-3\epsilon}^{-\epsilon} g_n = \int_{-3\epsilon}^{-\epsilon} G, \qquad \lim_{n\to\infty}\int_{-\epsilon}^{\epsilon} g_n = \int_{-\epsilon}^{\epsilon} G.$$

Since $\int_{-3\epsilon}^{-\epsilon} G < \int_{-\epsilon}^{\epsilon} G$, we have $\int_{-3\epsilon}^{-\epsilon} g_n < \int_{-\epsilon}^{\epsilon} G$ for large enough n. Similarly, for large enough n, $\int_{\epsilon}^{3\epsilon} g_n < \int_{-\epsilon}^{\epsilon} G$. So for large enough n, there are points $-3\epsilon < a_n < -\epsilon <$

$b_n < \epsilon < c_n < 3\epsilon$ with $g_n(a_n) < g_n(b_n) > g_n(c_n)$. For any such n, $\max_{x\in\mathbb{R}} g_n(x)$ occurs only for $x \in (-3\epsilon, 3\epsilon)$. For if $\max_{x\in\mathbb{R}} g_n(x) = g_n(\alpha)$ for $\alpha \le -3\epsilon$, then $g_n(\alpha) > g_n(a_n) < g_n(b_n) > g_n(c_n)$, which contradicts the shape of g_n. Similarly, $\max_{x\in\mathbb{R}} g_n(x)$ cannot occur for $x \ge 3\epsilon$.

Again take $\epsilon > 0$. Choose $\delta > 0$ such that $|G(x) - G(y)| < \epsilon$ whenever $|x - y| < \delta$. Take a function $B \ge 0$ with support in $[0, \delta]$, $\int_0^\delta B = 1$, and $||\widehat{B}||_1 < \infty$. Then

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} B(x-a) g_n(x)dx = \int_{-\infty}^{\infty} B(x-a)G(x)dx$$

uniformly in $a \in \mathbb{R}$. To see this, choose $A > 0$ so that $\int_{|u|>A} |\widehat{B}(u)|du < \epsilon$, and choose N so that

$$|\widehat{g}_n(u) - \widehat{G}(u)| < \epsilon \quad \text{for all } n > N,\ u \in [-A, A].$$

Then for all $n > N$,

$$\begin{aligned}
&\left| \int_{-\infty}^{\infty} B(x-a) g_n(x)dx - \int_{-\infty}^{\infty} B(x-a)G(x)dx \right| \\
&= \left| \int_{-\infty}^{\infty} e^{-iau} \widehat{B}(u)\widehat{g}_n(u)du - \int_{-\infty}^{\infty} e^{-iau} \widehat{B}(u)\widehat{G}(u)du \right| \\
&\le \int_{|u|>A} |\widehat{B}(u)|\ |\widehat{g}_n(u)|du + \int_{-A}^{A} |\widehat{g}_n(u) - \widehat{G}(u)|\ |\widehat{B}(u)|du \\
&+ \int_{|x|>A} |\widehat{B}(u)|\ |\widehat{G}(u)|du < \epsilon(2 + ||\widehat{B}||_1),
\end{aligned}$$

on noting that $|\widehat{g}_n(u)| \le \int_{-\infty}^{\infty} g_n(u)du = 1$.

Take $z < 0$. Choose N so that for all $n > N$, g_n is increasing on $(-\infty, z]$ and

$$\left| \int_{-\infty}^{\infty} B(x-a) g_n(x)dx - \int_{-\infty}^{\infty} B(x-a)G(x)dx \right| < \epsilon$$

for all $a \in \mathbb{R}$. For $y \le z$, $n > N$,

$$\begin{aligned}
\int_{-\infty}^{\infty} B(x-y+\delta) g_n(x)dx &= \int_{y-\delta}^{y} B(x-y+\delta) g_n(x)dx \\
&\le \int_{y-\delta}^{y} B(x-y+\delta) g_n(y)dx \\
&= g_n(y) \int_{-\infty}^{\infty} B = g_n(y).
\end{aligned}$$

Also for $n > N$,

$$\begin{aligned}
\int_{-\infty}^{\infty} B(x-y+\delta) g_n(x)dx &> \int_{-\infty}^{\infty} B(x-y+\delta) G(x)dx - \epsilon \\
&> \int_{-\infty}^{\infty} B(x-y+\delta) G(y)dx - 2\epsilon \\
&= G(y) - 2\epsilon.
\end{aligned}$$

Thus $g_n(y) > G(y) - 2\epsilon$ for all $n > N$. Similarly, for $y + \delta \le z$, $g_n(y) < G(y) + 2\epsilon$ for all $n > N$. Thus g_n converges to G uniformly on $(-\infty, z - \delta]$. A similar argument holds for $z > 0$, and so g_n converges to G uniformly outside any open interval containing 0.

Once again take $\epsilon > 0$ and choose $\delta > 0$ so that $|G(x) - G(y)| < \frac{\epsilon}{2}$ for $|x - y| \le 2\delta$. Choose N so that for $n > N$, $|g_n(x) - G(x)| < \frac{\epsilon}{2}$ for all $|x| \ge \delta$, and $\max g_n(x)$ occurs only for x in $(-\delta, \delta)$. Take any $n > N$ and x in $(-\delta, \delta)$. Then either $g_n(x) \ge g_n(-\delta)$ or $g_n(x) \ge g_n(\delta)$. Now $g_n(\delta) > G(\delta) - \frac{\epsilon}{2} > G(x) - \epsilon$ and similarly $g_n(-\delta) > G(x) - \epsilon$. Thus $g_n(x) > G(x) - \epsilon$. So we have shown that for any $\epsilon > 0$, there exists an integer N such that for all $n > N$ and all $x \in \mathbb{R}$, $g_n(x) > G(x) - \epsilon$.

Now suppose that g_n does not converge uniformly to G on $\mathbb{R}$. Then there is a number $k > 0$ and a sequence (x_n) with $\lim x_n = 0$ so that for arbitrarily large n,

$$g_n(x_n) > G(x_n) + k \quad \text{and} \quad \log g_n(x_n) > \log G(x_n) + k. \tag{5.3}$$

Choose points $0 < a < a + h < a + 2h < 1$. Then $2G(a + h) > G(a) + G(a + 2h)$ and $2G(-a - h) > G(-a) + G(-a - 2h)$. So for large enough n,

$$2g_n(a + h) > g_n(a) + g_n(a + 2h), \tag{5.4}$$

$$2g_n(-a - h) > g_n(-a) + g_n(-a - 2h). \tag{5.5}$$

Next choose $0 < 2\delta < a$ so that $|G(x) - G(y)| < k/3$ whenever $|x - y| \le \delta$. For large enough n, $x_n + 2\delta < a$ and $x_n + \delta/2 > 0$. Since $g_n \to G$ uniformly on $[\delta/2, \infty)$ and $|G(x_n + \delta) - G(x_n + 2\delta)| < k/3$, we have for large enough n,

$$|g_n(x_n + \delta) - g_n(x_n + 2\delta)| < \frac{k}{2}. \tag{5.6}$$

Also we have $G(x_n + \delta) < G(x_n) + k/3$, and so for large enough n,

$$g_n(x_n + \delta) < G(x_n) + \frac{k}{2}. \tag{5.7}$$

Hence for large enough n, by (5.6) and (5.7),

$$2g_n(x_n + \delta) < g_n(x_n + 2\delta) + G(x_n) + k.$$

Therefore, by (5.3), we see that for arbitrarily large n,

$$2g_n(x_n + \delta) < g_n(x_n) + g_n(x_n + 2\delta). \tag{5.8}$$

Now suppose g_n is bell-shaped. Choose n so that $x_n > -a$, $x_n + 2\delta < a$, and (5.4), (5.5), and (5.8) are satisfied. Let α, β be such that g_n is convex on $(-\infty, \alpha]$ and $[\beta, \infty)$ and concave on $[\alpha, \beta]$. By (5.4) and (5.5), $\beta > a$ and $\alpha < -a$. So g_n is concave on $[-a, a]$, which contradicts (5.8).

Next suppose that g_n is logconcave. A similar (but simpler) argument to that above shows that (5.8) can be replaced by

$$2 \log g_n(x + \delta) < \log g_n(x_n) + \log g_n(x_n + 2\delta),$$

which again gives a contradiction. □

We remark that the uniform convergence of g_n to G on $\mathbb{R}$ and the condition $\int_{-\infty}^{\infty} g_n = 1 = \int_{-\infty}^{\infty} G$ imply that $g_n \to G$ in $L^p(\mathbb{R})$ as $n \to \infty$ for all p, $1 \le p \le \infty$. Since convergence in $L^1(\mathbb{R})$ implies (5.1), the converse of Lemma 5.1 also holds.

From Lemma 5.1 we now derive the uniform convergence of $\widetilde{m}_n$ to G under an extra condition on the numbers $(r_{n,j})$ as in Theorem 1.4.

THEOREM 5.2. *We assume the conditions of Theorem* 1.4. *Then*

$$\lim_{n\to\infty}\left\{\sigma_n a_{n,k} - G\left(\frac{k-\mu_n}{\sigma_n}\right)\right\} = 0 \tag{5.9}$$

uniformly over k *in* $\mathbb{Z}$.

Proof. Since all $r_{n,j}$ lie in the sector $|\arg z| \le \frac{\pi}{3}$, it follows that the matrix $(a_{n,i-j})$ is totally positive of order 2. Hence $a_{n,k} \ge 0$ and

$$a_{n,k}^2 \ge a_{n,k-1}a_{n,k+1}, \quad k = 1,\dots,n-1, \ \ n = 1,2,\dots\ . \tag{5.10}$$

For $n = 1,2,\dots$, we define ψ_n as follows. Without loss of generality we may assume $a_{n,0}a_{n,n} \ne 0$, and it follows from (5.10) that $a_{n,k} > 0$, $k = 0,\dots,n$. We define ψ_n on $[-\mu_n/\sigma_n, (n-\mu_n)/\sigma_n]$ to be the piecewise linear function with knots $(j-\mu_n)/\sigma_n$, $j = 0,\dots,n$, satisfying

$$\psi_n\left(\frac{j-\mu_n}{\sigma_n}\right) = \log(\sigma_n a_{n,j}), \quad j = 0,1,\dots,n.$$

From (5.10), ψ_n is concave on $[-\mu_n/\sigma_n, (n-\mu_n)/\sigma_n]$. We now extend ψ_n to a continuous concave function on (α,β), where $\alpha = -(\mu_n+1)/\sigma_n$, $\beta = (n-\mu_n+1)/\sigma_n$, and

$$\lim_{x\to\alpha^+}\psi_n(x) = \lim_{x\to\beta^-}\psi_n(x) = -\infty.$$

For $n = 1,2,\dots$, we define

$$g_n(x) = \begin{cases} e^{\psi_n(x)}, & \alpha < x < \beta, \\ 0 & \text{otherwise.} \end{cases}$$

Clearly, g_n is logconcave, and

$$g_n\left(\frac{j-\mu_n}{\sigma_n}\right) = \sigma_n a_{n,j}, \quad j = 0,1,\dots,n.$$

As in Theorem 1.2, we define measures m_n, $n = 1,2,\dots$, by

$$m_n(\{k\}) = a_{n,k}, \quad k = 0,1,\dots,n,$$

and it follows that (m_n) is asymptotically normal. We note that for $k \in \mathbb{Z}$,

$$\int_{-\infty}^{\frac{k-\mu_n}{\sigma_n}} d\widetilde{m}_n = \sum_{j=0}^{k} a_{n,j},$$

where we put $a_{n,j} = 0$ for $j > n$. It follows from (5.15) that as $n \to \infty$,

$$\int_{-\infty}^{x} g_n - \int_{-\infty}^{x} d\widetilde{m}_n = O(\sigma_n^{-1})$$

uniformly in x. We can then apply Lemma 5.1 to the sequence of functions $g_n/\int_{-\infty}^{\infty} g_n$ to show that this sequence converges to G on $\mathbb{R}$. Hence g_n converges uniformly to G on $\mathbb{R}$, which by (5.15) gives (5.9). ∎

We now consider the uniform convergence of the normalized m_n-scaling functions $\widetilde{\phi}_n$ to G.

THEOREM 5.3. *Assume the conditions of Theorem* 1.5. *Then* $\widetilde{\phi}_n \to G$ *as* $n \to \infty$ *uniformly on* $\mathbb{R}$.

Proof. It follows from the work of Goodman and Micchelli (see [8]) and the properties of totally positive matrices (see [2]) that the functions ϕ_n, and hence $\widetilde{\phi}_n$, are bell-shaped. The result then follows from Theorem 1.1, Theorem 1.2, and Lemma 5.1. □

We remark that if the set of all $r_{n,j}$ lies in Re $z \geq 0$, then the condition that it also lies in D_γ for some $\gamma \in [0, \frac{\pi}{2})$ is equivalent to requiring that for some $\beta \in [0, \frac{\pi}{2})$ the set of all $r_{n,j}$ lying outside the sector $|\arg z| \leq \beta$ is bounded and bounded away from zero. In [4], Chui and Wang consider convergence of the sequence $(\widetilde{\phi}_n)$ as in Theorem 5.3 under the assumption that the polynomial $\sum_{k=0}^n a_{n,k} z^k$ is reciprocal and all $r_{n,j}$ are real and positive. They also assume that for $n = 1, 2, \dots$, $r_{n,j} = 1$ for at least Kn values of j for some fixed $K > 0$. They prove convergence in L^p, $1 \leq p < \infty$, which we have noted is weaker than uniform convergence.

We shall finish the paper by considering the order of uniform convergence for both the measures and the corresponding scaling functions. We first need to extend concepts of bell-shaped and logconcave to discrete measures. Suppose m is a probability measure on $\mathbb{Z}$ with $m(\{j\}) = a_j$, $j \in \mathbb{Z}$. We say m is *bell-shaped* if there are integers $k \leq \ell$ such that

$$\begin{aligned} 2a_j &\leq a_{j-1} + a_{j+1}, \quad j \leq k-1 \text{ and } j \geq \ell + 1, \\ 2a_j &\geq a_{j-1} + a_{j+1}, \quad k \leq j \leq \ell. \end{aligned}$$

We say m in *logconcave* if

$$a_j^2 \geq a_{j-1} a_{j+1}, \quad j \in \mathbb{Z}.$$

LEMMA 5.4. *For* $n = 1, 2, \dots$, *let* m_n *be a probability measure on* $\{0, 1, \dots, n\}$ *given by* $m_n(\{k\}) = a_{n,k}$, $k = 0, 1, \dots, n$, *which is either bell-shaped or logconcave, with mean* μ_n *and standard derivation* σ_n. *Suppose that for some* $K > 0$ *and* $r \geq 1$,

$$\left| \widehat{\widetilde{m}}_n(u) - e^{-u^2/2} \right| \leq K \sigma_n^{-r} \quad \text{for } |u| \leq K \sigma_n. \tag{5.11}$$

Then as $n \to \infty$,

$$\max_{k=0,\dots,n} \left| \sigma_n a_{n,k} - G\left(\frac{k - \mu_n}{\sigma_n} \right) \right| = O(\sigma_n^{-s}), \tag{5.12}$$

where $s = \min\{\frac{r}{2}, \frac{2}{3}\}$.

Proof. Take a nonnegative function N with support in $[-1, 1]$, $\int_{-\infty}^{\infty} N = 1$, $\|\widehat{N}\|_1 < \infty$, and for some $A > 0$,

$$|\widehat{N}(u)| \leq A(1 + |u|)^{-3r-1}, \quad u \in \mathbb{R}.$$

Take $0 < \delta < 1/2$. Let $B_1(x) := \delta^{-1} N(x/\delta)$ and $B_2(x) := \delta^{-1} N(x/\delta - 4)$. Then B_1 and B_2 have supports on $[-\delta, \delta]$ and $[3\delta, 5\delta]$, respectively, and $\int_{-\infty}^{\infty} B_1 = \int_{-\infty}^{\infty} B_2 = 1$. So

$$\int_{-\infty}^{\infty} B_1 G > G(\delta), \quad \int_{-\infty}^{\infty} B_2 G < G(3\delta).$$

and hence

$$\begin{aligned}\int_{-\infty}^{\infty} B_1 G - \int_{-\infty}^{\infty} B_2 G &> G(\delta) - G(3\delta)\\ &> |G'(\delta)|2\delta\\ &> |G''(\delta)|2\delta^2\\ &> |G''(1/2)|2\delta^2.\end{aligned}$$

Also for $j = 1, 2$,

$$\begin{aligned}\left|\int_{-\infty}^{\infty} B_j d\widetilde{m}_n - \int_{-\infty}^{\infty} B_j G\right| &= \left|\int_{-\infty}^{\infty} \widehat{B}_j(-u)(\widehat{\widetilde{m}}_n(u) - \widehat{G}(u))du\right|\\ &\le \int_{|u|\ge K\sigma_n} (|\widehat{\widetilde{m}}_n(u)| + \widehat{G}(u))|\widehat{B}_j(-u)|du\\ &\quad + \frac{K}{\sigma_n^r}\int_{-K\sigma_n}^{K\sigma_n} |\widehat{B}_j(u)|du\\ &\le 2\int_{|u|>K\sigma_n} |\widehat{N}(\delta u)|du + \frac{K}{\sigma_n^r}\int_{-\infty}^{\infty} |\widehat{N}(\delta u)|du\\ &= \frac{2}{\delta}\int_{|u|\ge K\delta\sigma_n} |\widehat{N}(u)|du + \frac{K}{\delta\sigma_n^r}\int_{-\infty}^{\infty} |\widehat{N}(u)|du\\ &\le \frac{C}{\delta(K\delta\sigma_n)^{3r}} + \frac{C}{\delta\sigma_n^r}\end{aligned}$$

for some $C > 0$. Choosing $\delta = c\sigma_n^{\beta-1}$ for some $\frac{1}{3} \le \beta < 1$, and $c > 1$, gives

$$\left|\int_{-\infty}^{\infty} B_j d\widetilde{m}_n - \int_{-\infty}^{\infty} B_j G\right| \le \frac{D}{c\sigma_n^{r+\beta-1}} \tag{5.13}$$

for some $D > 0$. Then

$$\begin{aligned}\int_{-\infty}^{\infty} B_1 d\widetilde{m}_n &> \int_{-\infty}^{\infty} B_1 G - \frac{D}{c\sigma_n^{r+\beta-1}}\\ &> \int_{-\infty}^{\infty} B_2 G + \left|G''\left(\frac{1}{2}\right)\right|2\delta^2 - \frac{D}{c\sigma_n^{r+\beta-1}}\\ &> \int_{-\infty}^{\infty} B_2 d\widetilde{m}_n + \frac{|G''(\frac{1}{2})|2c^2}{\sigma_n^{2-2\beta}} - \frac{2D}{c\sigma_n^{r+\beta-1}}\ .\end{aligned} \tag{5.14}$$

Now for $n = 1, 2, \dots$, choose a continuous function g_n, which is bell-shaped or logconcave as m_n is bell-shaped or logconcave, respectively, and satisfies

$$g_n\left(\frac{j-\mu}{\sigma_n}\right) = \sigma_n a_{n,j}, \quad j = 0, 1, \dots, n. \tag{5.15}$$

If m_n is logconcave, then this can be done as in the proof of Theorem 5.2, while if m_n is bell-shaped we can take g_n to be simply the piecewise linear interpolant. Note that if, for some constant b, $g_n \ge b$ on the support of B_j, $j = 1$ or 2, then $\int_{-\infty}^{\infty} B_j d\widetilde{m}_n$ bounds the product of b and a Riemann sum for B_j over its support with

interval length σ_n^{-1}. This Riemann sum equals a Riemann sum for N over $[0,1]$ with interval length $\delta^{-1}\sigma_n^{-1}$, which differs from $\int_0^1 N$ by $O(\delta^{-2}\sigma_n^{-2})$. Thus, by the uniform boundedness of g_n, we have

$$\int B_j d\widetilde{m}_n \geq b + O(\delta^{-2}\sigma_n^{-2}),$$

and similarly the result holds with $\geq$ replaced by $\leq$. Thus if $g_n(x) \leq g_n(y)$ for all $x \in [-\delta, \delta]$, $y \in [3\delta, 5\delta]$, we have

$$\int_{-\infty}^{\infty} B_1 d\widetilde{m}_n \leq \int_{-\infty}^{\infty} B_2 d\widetilde{m}_n + \frac{a}{\delta^2\sigma_n^2} = \int_{-\infty}^{\infty} B_2 d\widetilde{m}_n + \frac{a}{c^2\sigma_n^{2\beta}}$$

for a fixed constant a. Choosing $\beta = 2/3$ and c large enough, this would contradict (5.14), and so there are points $-\delta < b_n < \delta$, $3\delta < c_n < 5\delta$ with $g_n(b_n) > g_n(c_n)$. Similarly, we can choose b_n so that there is a point a_n in $(-5\delta, -3\delta)$ with $g_n(a_n) < g_n(b_n)$. As in the proof of Lemma 5.1, it follows from the shape of g_n that the maximum of $g_n(x)$ occurs only for x in $(-5\delta, 5\delta)$. So we have shown that for a constant a, maximum of $g_n(x)$ occurs for x in $(-a\sigma_n^{-1/3}, a\sigma_n^{1/3})$ for $n = 1, 2, \ldots$.

Now take $\delta = \sigma_n^{\beta-1}$ for some $1/3 \leq \beta < 1$ and $\gamma \geq a\sigma_n^{-1/3} + \delta$. Let $B(x) = \delta^{-1}N(\delta^{-1}(x-\gamma))$ so that B has support on $[\gamma - \delta, \gamma + \delta]$. As in (5.13)

$$\left| \int_{-\infty}^{\infty} B d\widetilde{m}_n - \int_{-\infty}^{\infty} BG \right| \leq \frac{D}{\sigma_n^{r+\beta-1}}$$

for some $D > 0$. Since g_n is decreasing on $[a\sigma_n^{-1/3}, \infty)$, for a constant $b > 0$,

$$\begin{aligned} g_n(\gamma - \delta) &\geq \int_{-\infty}^{\infty} B d\widetilde{m}_n - \frac{b}{\delta^2\sigma_n^2} \\ &\geq \int_{-\infty}^{\infty} BG - \frac{b}{\delta^2\sigma_n^2} - \frac{D}{\sigma_n^{r+\beta-1}} \\ &\geq G(\gamma + \delta) - \frac{b}{\sigma_n^{2\beta}} - \frac{D}{\sigma_n^{r+\beta-1}} \ . \end{aligned}$$

Since $|G'(\tau)| < 1$ for all τ in $\mathbb{R}$, $|G(x) - G(y)| \leq |x - y|$ for all $x, y \in \mathbb{R}$. So $G(\gamma + \delta) \geq G(\gamma - \delta) - 2\delta$, and so

$$g_n(\gamma - \delta) \geq G(\gamma - \delta) - \frac{b}{\sigma_n^{2\beta}} - \frac{D}{\sigma_n^{r+\beta-1}} - 2\delta.$$

Similarly,

$$g_n(\gamma + \delta) \leq G(\gamma + \delta) + \frac{b}{\sigma_n^{2\beta}} + \frac{D}{\sigma_n^{r+\beta-1}} + 2\delta.$$

Thus for all $x \geq a\sigma_n^{-1/3} + 2\delta$,

$$|g_n(x) - G(x)| \leq \frac{b}{\sigma_n^{2\beta}} + \frac{D}{\sigma_n^{r+\beta-1}} + \frac{2}{\sigma_n^{1-\beta}}.$$

For $r \geq 4/3$, put $\beta = 1/3$ to give

$$|g_n(x) - G(x)| = 0(\sigma_n^{-\frac{2}{3}}).$$

For $1 \le r \le 4/3$, put $\beta = 1 - r/2$ to give

$$|g_n(x) - G(x)| = 0(\sigma_n^{-\frac{r}{2}}).$$

Similarly, the result holds for $x \le -a\sigma_n^{-\frac{1}{3}} - 2\delta$. Thus for a constant $b > a$,

$$\sup\{|g_n(x) - G(x)| : |x| \ge b\sigma_n^{-\frac{1}{3}}\} = O(\sigma_n^{-s}) \tag{5.16}$$

for s as in the statement of Lemma 5.4. Note that for any $\delta > 0$ and $x, y \in (-\delta, \delta)$,

$$|G(x) - G(y)| \le |G''(0)|\delta|x - y| \le \delta|x - y|. \tag{5.17}$$

Take any $x \in (-b\sigma_n^{-\frac{1}{3}}, b\sigma_n^{\frac{-1}{3}})$. Then either

$$g_n(x) \ge g_n(b\sigma_n^{-\frac{1}{3}}) \quad \text{or} \quad g_n(x) \ge g_n(-b\sigma_n^{-\frac{1}{3}}).$$

Suppose the former. Then

$$\begin{aligned} g_n(x) \ge g_n(b\sigma_n^{-\frac{1}{3}}) &> G(b\sigma_n^{-\frac{1}{3}}) - O(\sigma_n^{-s}) \\ &> G(x) - O(\sigma_n^{-s}) - 2(b\sigma_n^{-\frac{1}{3}})^2. \end{aligned}$$

The same holds similarly for the latter case. Thus

$$\sup\{g_n(x) - G(x) : |x| \le b\sigma_n^{-\frac{1}{3}}\} = O(\sigma_n^{-s}). \tag{5.18}$$

Now note, as in the proof of Lemma 5.1, that if g_n is bell-shaped, then for all large enough n, g_n is concave on $[-\frac{2}{3}, \frac{2}{3}]$. Since concavity implies logconcavity, g_n is logconcave on $[-\frac{2}{3}, \frac{2}{3}]$ for all large enough n. By (5.16) and the mean value theorem,

$$\sup\left\{|\log g_n(x) - \log G(x)| : b\sigma_n^{-\frac{1}{3}} \le |x| \le \frac{2}{3}\right\} = O(\sigma_n^{-s}).$$

Take $0 \le x \le b\sigma_n^{-\frac{1}{3}}$ and n so large that $b\sigma_n^{-\frac{1}{3}} \le \frac{2}{9}$. Then

$$\begin{aligned} \log g_n(x) &\le 2\log g_n(x + b\sigma_n^{-\frac{1}{3}}) - \log g_n(x + 2b\sigma_n^{-\frac{1}{3}}) \\ &\le 2\log G(x + b\sigma_n^{-\frac{1}{3}}) - \log G(x + 2b\sigma_n^{-\frac{1}{3}}) + O(\sigma_n^{-s}) \\ &\le \log G(x) + |\log G(x + b\sigma_n^{-\frac{1}{3}}) - \log G(x)| \\ &\qquad + |\log G(x + b\sigma_n^{-\frac{1}{3}}) - \log G(x + 2b\sigma_n^{-\frac{1}{3}})| + O(\sigma_n^{-s}) \\ &\le \log G(x) + O(\sigma_n^{-\frac{2}{3}}) + O(\sigma_n^{-s}). \end{aligned}$$

A similar argument holds for $-b\sigma_n^{-\frac{1}{3}} \le x \le 0$, and applying the mean value theorem gives

$$\sup\{G(x) - g_n(x) : |x| \le b\sigma_n^{-\frac{1}{3}}\} = O(\sigma_n^{-3}). \tag{5.19}$$

Combining (5.16), (5.18), and (5.19) and recalling (5.15) then gives the result. □

Proof of Theorem 1.4. Theorem 1.4 follows from Proposition 4.1 and Lemma 5.4. □

To consider the order of uniform convergence for the scaling functions, we need the following analogue of Lemma 5.4. This can be proved in a similar manner to Lemma 5.4, but the proof is simpler, in particular because there is no restriction on the range of u as in (5.11).

LEMMA 5.5. *Suppose that (g_n) is a sequence of continuous functions, which are either bell-shaped or logconcave with $\int_{-\infty}^{\infty} g_n = 1$ and $\|\widehat{g}_n(u) - e^{-u^2/2}\|_\infty < \alpha_n$ for $n = 1, 2, \dots$, where $\lim_{n\to\infty} \alpha_n = 0$. Then as $n \to \infty$,*

$$\|g_n - G\|_\infty = O(\alpha_n^{\frac{1}{2}}).$$

Proof of Theorem 1.5. Theorem 1.5 follows from Theorem 1.3 and Lemma 5.5. □

REFERENCES

[1] J. BABAUB, A. P. WITKIN, M. BAUDIN, AND R. O. DUDA, *Uniqueness of Gaussian kernel for scale-space filtering*, IEEE Trans. Pattern Anal. Machine Intell., 8 (1986), pp. 26–33.

[2] J. M. CARNICER, T. N. T. GOODMAN, AND J. M. PEÑA, *A generalisation of the variation diminishing property*, Adv. Comput. Math., 3 (1995), pp. 375–394.

[3] L. CARLITZ, D. C. KURTZ, R. SCOVILLE, AND O. P. STACKELBERG, *Asymptotic properties of Eulerian numbers*, Z. Wahrsch. Verw. Gebiete, 23 (1972), pp. 47–54.

[4] C. C. K. CHUI AND J. Z. WANG, *A study of asymptotic optimal time-frequency localization of scaling functions and wavelets*, Ann. Numer. Math., 4 (1997), pp. 143–216.

[5] H. B. CURRY AND I. J. SCHOENBERG, *On Pólya frequency functions* IV. *The fundamental spline functions and their limits*, J. Analyse Math., 17 (1966), pp. 71–107.

[6] W. F. DONOGHUE, JR., *Distributions and Fourier Transforms*, Academic Press, New York, 1969.

[7] W. FELLER, *An Introduction to Probability and Its Applications*, Vol. II, John Wiley, New York, 1971.

[8] T. N. T. GOODMAN AND C. A. MICCHELLI, *On refinement equations determined by Pólya frequency sequences*, SIAM J. Math. Anal., 23 (1992), pp. 766–784.

[9] L. H. HARPER, *Stirling behaviour is asymptotically normal*, Ann. Math. Statist., 38 (1967), pp. 410–414.

[10] K. ITÔ, *Introduction to Probability*, Cambridge University Press, Cambridge, UK, 1984.

[11] T. A. POGGIO, V. TORRE, AND C. KOCH, *Computational vision and regularization theory*, Nature, 317 (1985), pp. 314–319.

[12] G. J. SZEKELY, *On the coefficients of polynomials with negative zeros*, Studia Sci. Math. Hungar., 8 (1973), pp. 123–124.

[13] M. UNSER, A. ALDROUBI, AND M. EDEN, *On the asymptotic convergence of B-spline wavelets to Gabor functions*, IEEE Trans. Inform. Theory, 38 (1992), pp. 864–872.

[14] M. UNSER, A. ALDROUBI, AND M. EDEN, *Polynomial spline pyramid*, IEEE Trans. Pattern Anal. Machine Intell., 15 (1993), pp. 364–378.

[15] M. UNSER, A. ALDROUBI, AND S. J. SCHIFF, *Fast implementation of continuous wavelet transforms with integer scale*, IEEE Trans. Signal Process., 42 (1994), pp. 3519–3523.

[16] Y. P. WANG AND S. L. LEE, *Scale-space derived from B-spline*, IEEE Trans. Pattern Anal. Machine Intell., 20 (1998), pp. 1040–1055.

[17] D. WILLIAMS, *Probability with Martingales*, Cambridge University Press, Cambridge, UK, 1991.

[18] R. A. YOUNG AND R. M. LESPERANCE, *The Gaussian derivative model for spatial-temporal vision*: II. *Cortical data*, Spatial Vision, 14 (2001), pp. 321–389.

SIAM J. MATH. ANAL.
Vol. 36, No. 2, pp. 347–383

WIGNER MEASURES IN THE DISCRETE SETTING: HIGH-FREQUENCY ANALYSIS OF SAMPLING AND RECONSTRUCTION OPERATORS*

FABRICIO MACIÀ†

Abstract. The goal of this article is to determine how the oscillation and concentration effects developed by a sequence of functions in $\mathbb{R}^d$ are modified by the action of sampling and reconstruction operators on regular grids. Our analysis is performed in terms of Wigner and defect measures, which provide a quantitative description of the high-frequency behavior of bounded sequences in $L^2(\mathbb{R}^d)$. We actually present explicit formulas that make possible the computation of such measures for sampled/reconstructed sequences. As a consequence, we are able to characterize sampling and reconstruction operators that preserve or filter the high-frequency behavior of specific classes of sequences. The proofs of our results rely on the construction and manipulation of Wigner measures associated to sequences of discrete functions.

Key words. Wigner measures, sampling and reconstruction, high-frequency analysis, concentration and oscillation, weak convergence, weak compactness, shift-invariant spaces

AMS subject classifications. 42C15, 94A12, 65D05, 46E35, 46E39

DOI. 10.1137/S0036141003431529

1. Introduction.

1.1. Statement of the problem: Oscillation and concentration under the effect of sampling and reconstruction. A central problem in numerical analysis and signal theory is that of reconstructing a function $u(x)$ defined in $\mathbb{R}^d$ from a discrete set of measurements taken on an uniform grid of step size h. These discrete values are typically obtained by applying to the function u a *sampling operator* S_φ^h of the type

$$S_\varphi^h u(n) := \frac{1}{h^d} \int_{\mathbb{R}^d} u(x) \overline{\varphi\left(\frac{x}{h} - n\right)} dx$$

for some *sampling function* φ. One then tries to recover u by means of a *reconstruction* (or *interpolation*) *operator* T_ψ^h, which associates to a sequence of discrete values $U := (U_n)$ a function

$$T_\psi^h U(x) := \sum_{n \in \mathbb{Z}^d} U_n \psi\left(\frac{x}{h} - n\right);$$

here ψ is some fixed *reconstruction function*. This process usually provides only an approximation $u^h := T_\psi^h S_\varphi^h u$ of the original function u, with an error that vanishes as h tends to zero. Such reconstruction schemes have been the subject of intensive study from the point of view of both approximation theory and numerical analysis.

*Received by the editors July 15, 2003; accepted for publication (in revised form) March 5, 2004; published electronically July 29, 2004. This work was supported by projects BFM02-03345 of MCyT (Spain) and HYKE (ref. HPRN-CT-2002-00282), HMS2000 (ref. HPRN-CT-2000-00109) of the European Union.

http://www.siam.org/journals/sima/36-2/43152.html

†Departamento de Matemática Aplicada, Universidad Complutense de Madrid, Fac. CC. Matemáticas, Avda. Complutense s/n, 28040 Madrid, Spain (fabricio_macia@mat.ucm.es).

Here we shall be concerned with the *high-frequency* approximation properties of those operators; that is, we shall study how the scheme $T_\psi^h S_\varphi^h$ is able to capture (or filter) *oscillation* and *concentration*-like phenomena on the functions it is intended to approximate. More generally, we shall be interested in clarifying how the high-frequency behavior of a sequence of reconstructed functions depends on the profiles φ, ψ and the sampling rate h chosen.

Before giving a more precise statement of our objectives, let us first illustrate the above discussion with two specific examples: consider $f_k(x) := k^{d/2}\rho(k(x-x_0))$ and $g_k(x) := \rho(x)e^{ikx\cdot\xi^0}$ with $\rho \in L^2(\mathbb{R}^d)$; the sequence (f_k) concentrates around the point x_0 as $k \to \infty$, whereas (g_k) oscillates in the direction ξ^0. The results we shall present in this paper are aimed at understanding to what extent the sequences $(T_\psi^{h_k} S_\varphi^{h_k} f_k)$ and $(T_\psi^{h_k} S_\varphi^{h_k} g_k)$ reproduce the same behavior as (f_k) and (g_k) (i.e., if concentration and oscillation persist) for a given sequence (h_k) of positive reals that tends to zero (the sampling steps) and some choice of φ and ψ.

Perhaps the simplest convenient setting in which to formulate our results is provided by the notion of *defect measure*, an object that gives a quantitative description of what we shall understand by concentration and oscillation effects and whose definition we next recall. Let (u_k) be a weakly converging sequence in the space $L^2(\mathbb{R}^d)$; denote by u its weak limit and remark that the densities $|u_k - u|^2$ are uniformly bounded in $L^1(\mathbb{R}^d)$. Helly's compactness theorem then ensures that some subsequence $(|u_{k_n} - u|^2)$ weakly converges in the set of positive Radon measures,[1] or, in other words, that there exists a positive measure ν on $\mathbb{R}^d$ such that

$$\int_{\mathbb{R}^d} \phi(x)|u_{k_n}(x) - u(x)|^2 dx \to \int_{\mathbb{R}^d} \phi(x) d\nu(x) \quad \text{as } n \to \infty$$

for every $\phi \in C_c(\mathbb{R}^d)$. When the above convergence takes place without extracting a subsequence we say that ν is the *defect measure* of the sequence (u_k).

Immediately from this definition one deduces the following general principle: if ν is the defect measure of a sequence (u_k) and $\omega \subset \mathbb{R}^d$ is an open subset, then there is an equivalence between $\nu(\omega) = 0$ and the fact that $u_k|_\omega$ converges strongly to $u|_\omega$ in $L^2_{\text{loc}}(\omega)$. Thus, the support of ν is precisely the set where strong convergence fails, that is, the set where oscillations and concentrations take place.

But defect measures are also able to detect concentration and oscillatory phenomena and give quantitative information about them. Consider the sequences (f_k), (g_k) previously defined; they both weakly converge to zero in $L^2(\mathbb{R}^d)$, and it is easy to check that their respective defect measures are $\|\rho\|^2_{L^2(\mathbb{R}^d)}\delta_{x_0}$ and $|\rho(x)|^2 dx$. Notice that, in the first case, the defect measure actually captures the concentration of the sequence around the point $x = x_0$. In the complementary of that point, where the sequence converges strongly to zero, the measure vanishes. In the second example, the defect measure is uniformly distributed on $\mathbb{R}^d$, this being consistent with the fact that strong convergence does not take place in any subset of $\mathbb{R}^d$.

Let us point out that the analysis of concentration and oscillation effects developed by a sequence of functions is a central issue in many problems of the calculus of variations and partial differential equations. A number of applications of defect

[1] From now on, we shall use the term *measure* as an abbreviation of the longer *Radon measure*. Recall that the space of Radon measures $\mathcal{M}(\mathbb{R}^d)$ is identified, by Riesz's theorem, with the space of continuous linear functionals on $C_c(\mathbb{R}^d)$.

measures may be found in the analysis of variational problems with loss of compactness performed by Lions in [11, 12].[2]

Consider a sequence (u_k), weakly converging to zero in $L^2(\mathbb{R}^d)$; sample it using a profile φ and form the reconstructed sequence

$$v_k := T_\psi^{h_k} S_\varphi^{h_k} u_k$$

for some given ψ and some sequence (h_k) of positive reals tending to zero. The functions v_k are bounded in $L^2(\mathbb{R}^d)$ and tend weakly to zero, provided φ and ψ satisfy suitable hypotheses (see Lemma 3.1 in section 3 below). Suppose furthermore that the densities $|v_k|^2$ weakly converge to the defect measure $\nu_{\varphi,\psi}$.

One of the main issues addressed in this article is that of understanding the relations existing between the defect measure $\nu_{\varphi,\psi}$, the profiles φ, ψ, and the sequences (u_k), (h_k). Among these, we point out the following:

A. Is there a formula, valid for any sequence (u_k), relating $\nu_{\varphi,\psi}$ to the defect measure ν only in terms of the profiles φ and ψ?

B. Given (u_k), characterize the profiles φ and ψ such that $\nu_{\varphi,\psi} = 0$. This is the problem of *filtering* since, as we have discussed before, $\nu_{\varphi,\psi} = 0$ is equivalent to the strong local convergence to zero of the sequence $(T_\psi^{h_k} S_\varphi^{h_k} u_k)$.

C. Similarly, characterize the profiles φ and ψ such that $\nu_{\varphi,\psi} = \nu$ for a given (u_k).

D. Finally, characterize the profiles that give $\nu_{\varphi,\psi} = \nu$ *for every* (u_k).

We shall prove that the answer to question A is negative. This is due to the fact that the measure $\nu_{\varphi,\psi}$ is sensitive to the characteristic directions of oscillation of the sequence (u_k), whereas ν is unable to distinguish them. As we have seen above, the defect measure of the oscillating sequence (g_k) equals $|\rho(x)|^2 dx$ independently of the vector ξ^0; that is not the case for $\nu_{\varphi,\psi}$. Indeed, under additional assumptions on φ and ψ we prove (see Theorem 1.3 and Corollary 1.4) that

$$\nu_{\varphi,\psi}(x) = \sum_{k\in\mathbb{Z}^d} |\widehat{\psi}(\xi^0 + 2\pi k)|^2 |\widehat{\varphi}(\xi^0)|^2 |\rho(x)|^2 dx.$$

Thus the measure $\nu_{\varphi,\psi}$ is ξ^0-dependent and cannot be expressed solely in terms of ν, φ, and ψ. Note that $\nu_{\varphi,\psi}$ is identically zero as soon as any of $\sum_{k\in\mathbb{Z}^d} |\widehat{\psi}(\xi^0 + 2\pi k)|^2$ or $\widehat{\varphi}(\xi^0)$ is null. Analogously, the profiles that give $\nu_{\varphi,\psi} = \nu$ are precisely those which satisfy

$$\sum_{k\in\mathbb{Z}^d} |\widehat{\psi}(\xi^0 + 2\pi k)|^2 |\widehat{\varphi}(\xi^0)|^2 = 1.$$

Therefore, in order to understand how $\nu_{\varphi,\psi}$ is built, we must have at our disposal an object that is able to distinguish between oscillatory phenomena at different directions.

1.2. Wigner measures. This refinement is provided by the theory of *Wigner measures*.[3] Given a bounded sequence in $L^2(\mathbb{R}^d)$ one associates to it a measure $\mu(x,\xi)$

[2] We also refer to Evans's notes [4] for an exposition of some additional applications as well as a discussion of other measure-theoretical objects (such as, for example, Young measures) designed to study the failure of strong convergence.

[3] This object is present in the work of Wigner on semiclassical quantum mechanics [20]. Recently, Wigner measures have gained interest since the works of, for example, Gérard [6], Lions and Paul [13], and Markowich, Mauser, and Poupaud [14]. Related objects are the *microlocal defect measures* or *H-measures*, introduced independently by Gérard [5] and Tartar [18].

on $\mathbb{R}^d \times \mathbb{R}^d$ which describes the concentration and oscillation effects (these are the respective roles of the variables x and ξ) occurring at some characteristic length-scale. This measure takes into account the characteristic speeds as well as the directions of propagation of oscillations. One way of defining them consists in replacing the density $|u(x)|^2$ involved in the definition of the defect measure by the phase space (microlocal) density,

$$m^\varepsilon[u](x,\xi) := \frac{1}{(2\pi\varepsilon)^d}\overline{u(x)}\widehat{u}(\xi/\varepsilon)e^{ix\cdot\xi/\varepsilon}, \tag{1.1}$$

where $\widehat{u}$ is the Fourier transform of u and ε is a positive constant. The $(2\pi)^{-d}$ factor in the definition of $m^\varepsilon[u]$ is placed to have

$$\int_{\mathbb{R}^d} m^\varepsilon[u](x,\xi)d\xi = |u(x)|^2, \qquad \int_{\mathbb{R}^d} m^\varepsilon[u](x,\xi)dx = \frac{|\widehat{u}(\xi/\varepsilon)|^2}{(2\pi\varepsilon)^d}. \tag{1.2}$$

Thus, the function $m^\varepsilon[u]$ may be looked at as a joint physical space–Fourier space "density," in spite of the fact that $m^\varepsilon[u]$ is not positive in general. However, limits of these quantities are positive measures.

THEOREM 1.1. *Let (u_k) be a bounded sequence in $L^2(\mathbb{R}^d)$ and let (ε_k) be a sequence of positive numbers tending to zero. Then it is possible to extract a subsequence (u_{k_n}) such that, for every test function $a \in \mathcal{S}(\mathbb{R}^d \times \mathbb{R}^d)$,*

$$\lim_{n\to\infty} \int_{\mathbb{R}^d\times\mathbb{R}^d} a(x,\xi)m^{\varepsilon_{k_n}}[u_{k_n}](x,\xi)dxd\xi = \int_{\mathbb{R}^d\times\mathbb{R}^d} a(x,\xi)d\mu(x,\xi), \tag{1.3}$$

where μ is a finite positive measure on $\mathbb{R}^d \times \mathbb{R}^d$.

A measure $\mu \in \mathcal{M}_+(\mathbb{R}^d\times\mathbb{R}^d)$ is called the *Wigner measure* of the sequence (u_k) at scale (ε_k) whenever the limit (1.3) holds without extracting a subsequence. Different proofs of Theorem 1.1 may be found in [8, 13, 7]. Let us point out that other quadratic densities may be used to define Wigner measures. For instance, in [13] μ is obtained by replacing $m^\varepsilon[u]$ in the limit (1.3) by the more familiar *Wigner transform*:

$$w^\varepsilon[u](x,\xi) := \int_{\mathbb{R}^d} u\left(x-\varepsilon\frac{p}{2}\right)\overline{u\left(x+\varepsilon\frac{p}{2}\right)}e^{ip\cdot\xi}\frac{dp}{(2\pi)^d}. \tag{1.4}$$

It is also possible to consider *wave-packet (Husimi) transforms*. Of course, all of these methods are equivalent (the same limit is obtained); see the discussion in [8].

The Wigner measure encodes all the information contained in the defect measure, provided the sequence (u_k) oscillates at frequencies of the order of ε_k^{-1}. We state this more precisely in the following proposition (see [8, 13]).

PROPOSITION 1.2. *If μ is the Wigner measure at scale (ε_k) of a sequence (u_k) and ν is the measure obtained as the weak limit in $\mathcal{M}_+(\mathbb{R}^d)$ of the densities $|u_k|^2dx$, then the identity*

$$\nu(x) = \int_{\mathbb{R}^d} \mu(x,d\xi)$$

*holds, provided (u_k) is ε_k-*oscillatory:

$$\limsup_{k\to\infty} \int_{|\xi|>R/\varepsilon_k} |\widehat{u_k}(\xi)|^2 d\xi \to 0 \quad \text{as } R\to\infty. \tag{1.5}$$

Notice that condition (1.5) actually expresses that the energy of the Fourier transform of u_k is concentrated in a ball of radius R/ε_k, which should be understood as the requirement that the sequence (u_k) does not oscillate at length scales finer than ε_k.

To illustrate this discussion it may be helpful to look at explicit computations. The Wigner measure at scale (ε_k) of the concentrating sequence (f_k) defined at the beginning of this section is given by

$$\mu(x,\xi) = \begin{cases} \|\rho\|^2_{L^2(\mathbb{R}^d)}\delta_{x_0}(x)\otimes\delta_0(\xi) & \text{if } \varepsilon_k k\to 0,\\ \delta_{x_0}(x)\otimes|\widehat{\rho}(\xi)|^2\dfrac{d\xi}{(2\pi)^d} & \text{if } \varepsilon_k = k^{-1},\\ 0 & \text{if } \varepsilon_k k\to\infty, \end{cases} \tag{1.6}$$

while for the oscillating sequence (g_k) it can be checked to be

$$\mu(x,\xi) = \begin{cases} |\rho(x)|^2dx\otimes\delta_0(\xi) & \text{if } \varepsilon_k k\to 0,\\ |\rho(x)|^2dx\otimes\delta_{\xi^0}(\xi) & \text{if } \varepsilon_k = k^{-1},\\ 0 & \text{if } \varepsilon_k k\to\infty. \end{cases} \tag{1.7}$$

These examples show the importance of the choice of the scale (ε_k). When this scale is taken to be coarser than the characteristic length-scale k^{-1} of oscillation/concentration, it is no longer true that the projection on the first component of their Wigner measures coincides with the defect measure. On the other hand, in the case $\varepsilon_k k \to 0$ (the scale chosen is much smaller than the actual oscillation scale) the Wigner measure is not able to capture the direction of oscillation. Hence, to obtain a complete description, the scale (ε_k) must be taken of the same order as that of the oscillations.

Wigner measures turn out to be the correct tools for comparing the high-frequency behavior of the sequences (u_k) and $(T^{h_k}_\psi S^{h_k}_\varphi u_k)$.

1.3. Computation of Wigner and defect measures. Given a sequence of sampling steps (h_k), it seems clear that the functions $T^{h_k}_\psi S^{h_k}_\varphi u_k$ will not develop oscillation and concentration effects of characteristic sizes asymptotically smaller than h_k. Most commonly, these functions will form an h_k-oscillatory sequence;[4] consequently, only Wigner measures at scales coarser than or of the same order as (h_k) will be considered.

In order to establish explicit formulas, we shall require additional hypotheses on φ, ψ and on the Wigner measures involved. Nevertheless, in order to simplify the statement of our results, in this introduction we shall impose the following (more restrictive) condition on the admissible profiles:

$$|\gamma(x)| \le C(1+|x|)^{-d-\varepsilon} \quad \text{for every } x\in\mathbb{R}^d \text{ and some } C,\varepsilon>0. \tag{1.8}$$

More general results may be found in section 7.

We prove the following.

THEOREM 1.3. *Let φ, ψ satisfy* (1.8). *Suppose (u_k) is a bounded sequence in $L^2(\mathbb{R}^d)$ and that μ is its Wigner measure at scale (h_k). Suppose, moreover, that the measures*

$$|\widehat{\varphi}(\xi+2\pi n)|^2\mu(x,\xi+2\pi n) \tag{1.9}$$

[4]However, this may fail for some pathological examples (see section 5.3).

are mutually singular for $n \in \mathbb{Z}^d$.

Then the Wigner measure at scale (h_k) *of the sequence* $(T_\psi^{h_k} S_\varphi^{h_k} u_k)$ *is given by*

$$\mu_{\varphi,\psi}(x,\xi) = |\widehat{\psi}(\xi)|^2 \sum_{k\in\mathbb{Z}^d} |\widehat{\varphi}(\xi+2\pi n)|^2 \mu(x,\xi+2\pi n).$$

From this, one deduces the following corollary.

COROLLARY 1.4. *If, moreover,* $|T_\psi^{h_k} S_\varphi^{h_k} u_k|^2 dx$ *weakly converges to a measure* $\nu_{\varphi,\psi}$, *then*

$$\nu_{\varphi,\psi}(x) = \int_{\mathbb{R}^d} \sum_{k\in\mathbb{Z}^d} |\widehat{\psi}(\xi+2\pi k)|^2 |\widehat{\varphi}(\xi)|^2 \mu(x,d\xi).$$

This shows, in particular, that a formula relating $\nu_{\varphi,\psi}$ and the weak limit ν of $|u_k|^2 dx$ does not exist unless (u_k) is h_k-oscillatory and μ is of the form $\nu(x)\otimes\sigma(\xi)$. It also shows that $\nu = \nu_{\varphi,\psi}$ if and only if $\sum_{k\in\mathbb{Z}^d} |\widehat{\psi}(\xi+2\pi k)|^2 |\widehat{\varphi}(\xi)|^2 = 1$ for $\tilde{\mu}$-almost every $\xi \in \mathbb{R}^d$, where $\tilde{\mu} := \int \mu(\cdot, d\xi)$. Consequently, there do not exist profiles φ, ψ satisfying (1.8) such that ν equals $\nu_{\varphi,\psi}$ for every h_k-oscillatory sequence (u_k).

On the other hand, Theorem 1.3 implies that question A above does have a positive answer in terms of Wigner measures, at least when restricted to the class of sequences which satisfy (1.9). That condition, roughly speaking, imposes a restriction on the size of the region in frequency space where an admissible sequence fails to converge strongly to zero. Below, we shall compare it with that appearing in Shannon's sampling theorem.

The above results will be obtained as corollaries of the more general Theorems 7.1 and 7.3. Profiles that belong to negative-order Sobolev spaces or that fail to satisfy the localization hypothesis (1.8) are allowed. However, this will require us to impose compatibility conditions on the Wigner measure μ.

As an illustration of the range of results that will be obtained in this more general setting, we present an *asymptotic version* of *Shannon's sampling theorem*.[5] It corresponds to taking as sampling profile $\varphi = \delta_0$, the Dirac delta at the origin, and as reconstruction function $\widehat{\psi} := \mathbf{1}_Q$, where $Q := [-\pi,\pi)^d$. Notice that $S_{\delta_0}^h u(n) = u(hn)$ is the discretization operator, whereas the T_ψ^h corresponds to band-limited reconstruction.

THEOREM 1.5. *Let* (u_k) *be a bounded sequence in* $L^2(\mathbb{R}^d)$ *and denote by* μ *its Wigner measure at scale* (h_k). *Suppose, in addition, that* $u_k \in H^s(\mathbb{R}^d)$ *for some* $s > d/2$ *and*

$$\begin{aligned} &\text{(i)} && (1-h_k^2\Delta_x)^{s/2} u_k \quad \textit{are uniformly bounded in } L^2(\mathbb{R}^d).\\ &\text{(ii)} && \mu(\mathbb{R}^d \times (\partial Q + 2\pi n)) = 0 \quad \textit{for } n \in \mathbb{Z}^d.\\ &\text{(iii)} && \mu(x,\xi+2\pi n), \quad n\in\mathbb{Z}^d, \quad \textit{are mutually singular measures.} \end{aligned} \tag{1.10}$$

Then the Wigner measure at scale (h_k) *of* $(T_\psi^{h_k} S_{\delta_0}^{h_k} u_k)$ *is*

$$\mu_{\delta_0,\psi}(x,\xi) = \mathbf{1}_Q(\xi) \sum_{n\in\mathbb{Z}^d} \mu(x,\xi+2\pi n). \tag{1.11}$$

Moreover, if $|T_\psi^{h_k} S_{\delta_0}^{h_k} u_k|^2 dx$ *and* $|u_k|^2 dx$ *weakly converge to* ν_S *and* ν, *respectively, then*

$$\nu_S(x) = \int_{\mathbb{R}^d} \mu(x,d\xi) = \nu(x).$$

[5]See section 3.1 for a statement of Shannon's original sampling theorem.

Thus, unlike the operators considered in Theorem 1.3, the composition of discretization and band-limited reconstruction preserves the defect measure for a large class of sequences.

Notice that, by the Sobolev imbedding theorem, $S_{\delta_0}^{h_k} u_k$ is well-defined. Actually, (1.10.i) ensures that the sequence of discretizations is square-summable and, consequently, that $(T_\psi^{h_k} S_{\delta_0}^{h_k} u_k)$ is bounded in $L^2(\mathbb{R}^d)$ and h_k-oscillatory (for a more complete result, we refer to Lemma 3.1). Condition (1.10.ii) appears because $\widehat{\psi}$ is not continuous; we shall discuss its necessity in section 4.4. Finally, (1.10.iii) should be understood as the analogue of Shannon's original band-limited condition in this context.

To conclude this short description, let us present how the above results may be refined when the sequence (u_k) is known to be ε_k-oscillatory and the sampling rate (h_k) is taken to satisfy $h_k/\varepsilon_k \to 0$. As can be expected, much more precision is gained.

THEOREM 1.6. *Suppose φ, ψ satisfy* (1.8) *and (u_k) is an ε_k-oscillatory, bounded sequence in $L^2(\mathbb{R}^d)$. If μ is its Wigner measure at scale (ε_k), then the corresponding measure of the sequence $(T_\psi^{h_k} S_\varphi^{h_k} u_k)$ is*

$$\mu_{\varphi,\psi} = |\widehat{\psi}(0)|^2 |\widehat{\varphi}(0)|^2 \mu.$$

Moreover, if the densities a $|T_\psi^{h_k} S_\varphi^{h_k} u_k|^2 dx$ and $|u_k|^2 dx$ weakly converge to $\nu_{\varphi,\psi}$ and ν, respectively, then

$$\nu_{\varphi,\psi}(x) = \sum_{n\in\mathbb{Z}^d} |\widehat{\psi}(2\pi n)|^2 |\widehat{\varphi}(0)|^2 \nu(x).$$

This theorem holds under much more general conditions on φ and ψ (see Theorem 7.6) and gives a positive answer to question A, provided we consider only ε_k-oscillatory sequences.

An immediate consequence of the above result is that zero-mean sampling profiles φ (i.e., with $\widehat{\varphi}(0) = 0$, as a wavelet, for instance) completely filter any oscillations that occur at scales much coarser than the sampling rate h_k. For such a profile, $\nu_{\varphi,\psi} = 0$ for every ε_k-oscillatory sequence. An analogous phenomenon occurs for reconstruction profiles satisfying $\widehat{\psi}(2\pi n) = 0$ for every $n \in \mathbb{Z}^d$.

On the other hand, a sufficient condition to have equality between $\nu_{\varphi,\psi}$ and ν is that $|\widehat{\varphi}(0)| = |\widehat{\psi}(0)| = 1$ and $|\widehat{\psi}(2\pi n)| = 0$ for $n \neq 0$.

1.4. Strategy of proof: Wigner measures in the discrete setting. The proof of the results we have presented above will be achieved by analyzing separately the sampling and reconstruction operators S_φ^h and T_ψ^h. In order to develop this strategy, it is necessary to deal with the concept of *Wigner measure associated to a sequence of discrete functions*. We shall introduce it by means of a discrete analogue of the transform $m^\varepsilon[\cdot]$. We detail this in the following paragraph.

To a discrete square-summable function $U \in L^2(h\mathbb{Z}^d)$, where $L^2(h\mathbb{Z}^d)$ stands for the space of the functions U defined on $\mathbb{Z}^d$ with values in $\mathbb{C}$ such that the norm

$$\|U\|_h := \left(h^d \sum_{n\in\mathbb{Z}^d} |U_n|^2 \right)^{1/2}$$

is finite, we associate

$$M^\varepsilon[U](x,\xi) := \frac{h^{2d}}{(2\pi\varepsilon)^d} \sum_{m\in\mathbb{Z}^d} \overline{U_m} \widehat{U}\left(\frac{h}{\varepsilon}\xi\right) e^{im\cdot(h/\varepsilon)\xi} \delta_{hm}(x). \tag{1.12}$$

Here, δ_{hm} is the Dirac mass centered at the point hm, and $\widehat{U}$ denotes the discrete Fourier transform

$$\widehat{U}(\xi) := \sum_{n\in\mathbb{Z}^d} U_n e^{-in\cdot\xi},$$

which, as is well known, is a $2\pi\mathbb{Z}^d$-periodic function in $L^2_{\rm loc}(\mathbb{R}^d)$. The discrete transform $M^\varepsilon[U]$ may be related to the continuous $m^\varepsilon[u]$ by noticing that

$$M^\varepsilon[U] = m^\varepsilon[T^h_{\delta_0}U], \quad \text{where } T^h_{\delta_0}U(x) = h^d \sum_{n\in\mathbb{Z}^d} U^h_n \delta_{hn}(x). \tag{1.13}$$

This is meaningful, since $m^\varepsilon[u]$ is well-defined for any tempered distribution $u \in \mathcal{S}'(\mathbb{R}^d)$.

In order to simplify our language we make the following definition.

DEFINITION 1.7. *Let $h = (h_k)$ be a scale. We shall call a sequence (U^{h_k}) h_k-bounded if and only if $U^{h_k} \in L^2(h_k\mathbb{Z}^d)$ and $\|U^{h_k}\|_{h_k} \le C$ for every $k \in \mathbb{N}$.*

One has the following convergence result (which is not a direct consequence of Theorem 1.1).

PROPOSITION 1.8. *Let (h_k), (ε_k) be scales such that (h_k/ε_k) is bounded and let (U^{h_k}) be an h_k-bounded sequence of discrete functions. Then $(M^{\varepsilon_k}[U^{h_k}])$ is bounded in $\mathcal{S}'(\mathbb{R}^d\times\mathbb{R}^d)$, and given any of its convergent subsequences $(U^{h_{k_n}})$, there exists a positive measure μ such that*

$$\lim_{n\to\infty} \langle M^{\varepsilon_{k_n}}[U^{h_{k_n}}], a\rangle_{\mathcal{S}'\times\mathcal{S}} = \int_{\mathbb{R}^d\times\mathbb{R}^d} a(x,\xi)d\mu(x,\xi) \tag{1.14}$$

for every $a \in \mathcal{S}(\mathbb{R}^d\times\mathbb{R}^d)$.

This will be proved as a corollary of the more general Proposition 3.4, which in turn follows from the analysis of Wigner measures associated to functions in negative-order Sobolev spaces that is performed in section 8. As in the continuous setting, we say that a measure μ is the *Wigner measure at scale* (ε_k) of a sequence of discrete functions (U^{h_k}) if the limit (1.14) holds for the whole sequence.

Remark 1.9. (i) When (h_k/ε_k) is unbounded, it may happen that $M^{\varepsilon_k}[U^{h_k}]$ is not bounded in $\mathcal{S}'(\mathbb{R}^d\times\mathbb{R}^d)$.

(ii) If $h_k/\varepsilon_k \to c > 0$, then μ is not finite. Indeed, it is periodic (with respect to the lattice $(2\pi/c)\mathbb{Z}^d$) in the ξ variable.

(iii) However, when $h_k/\varepsilon_k \to 0$, the Wigner measure μ is finite, as in the continuous case.

With this tool at our disposal, we are able to compare the Wigner measure of a sequence of discrete functions (U^{h_k}) with that of a reconstructed sequence $(T^{h_k}_\psi U^{h_k})$. Analogously, we may compute Wigner measures of sequences of sampled discrete functions $(S^{h_k}_\varphi u_k)$ in terms of those corresponding to the original sequence (u_k). These are, respectively, the contents of Theorems 4.6 and 4.2.

1.5. Plan of the article. Results and assumptions concerning the operators S^h_φ and T^h_ψ are collected in section 3.

In section 4, the problem of computing Wigner measures for sequences of sampled or reconstructed functions is addressed. Formulas for Wigner measures at scales of the same order as the sampling/reconstruction step (h_k) are presented in Theorems 4.6 and 4.2. Theorems 1.3 and 1.5 then easily follow from those two results. We also

point out the relationships existing between these Wigner measures and the concept of *Wigner series* introduced in [14, 9].

The problem of the computation of defect measures of sequences of the form $(T_\psi^{h_k} U^{h_k})$ is considered in section 5; the main results are presented in Proposition 5.8 and Corollary 5.9.

In section 6 we investigate Wigner measures at scales (ε_k) satisfying $h_k/\varepsilon_k \to 0$. Explicit formulas are presented in Theorems 6.1 and 6.2, from which Theorem 1.6 immediately follows.

The composition of sampling and reconstruction is studied in section 7, where the main results of this article are proved.

Finally, section 8 contains the elements from the theory of Wigner measures on which the proofs of most of the results of this article are based. Propositions 8.1 and 8.3, which extend the theory of Wigner measures to sequences in Sobolev spaces of negative order, are systematically used throughout this paper.

2. Notation and conventions. We briefly present some notation that will be used throughout this article.

$B(x;R)$ will denote the open ball with radius R of $\mathbb{R}^d$ centered at the point x. $\mathbf{1}_A$ will denote the characteristic function of a set $A \subseteq \mathbb{R}^d$.

We write Γ to denote the lattice $2\pi\mathbb{Z}^d$. A function f defined on $\mathbb{R}^d$ is Γ-periodic if $f(x+\gamma) = f(x)$ for every $\gamma \in \Gamma$ and every $x \in \mathbb{R}^d$.

We adopt the following convention for the Fourier transform:

$$\widehat{u}(\xi) := \int_{\mathbb{R}^d} u(x) e^{-ix\cdot\xi} dx.$$

Given a measurable function $\varphi(\xi)$, the *Fourier multiplier* of symbol φ is the operator $\varphi(D_x)$ formally defined by

$$\varphi(D_x)u(x) := \int_{\mathbb{R}^d} \varphi(\xi)\widehat{u}(\xi) e^{ix\cdot\xi} \frac{d\xi}{(2\pi)^d} = \check{\varphi} * u(x),$$

$\check{\varphi}$ being the inverse Fourier transform of φ.

A particularly important Fourier multiplier is the *Bessel potential* $\langle D_x \rangle$ of symbol

$$\langle \xi \rangle := (1 + |\xi|^2)^{1/2}.$$

Next, we recall the definition of some function spaces.

As usual, $\mathcal{S}(\mathbb{R}^d)$ denotes the space of *rapidly decreasing functions* and $\mathcal{S}'(\mathbb{R}^d)$ stands for its dual, the space of *tempered distributions*.

Given $r \in \mathbb{R}$, $H^r(\mathbb{R}^d)$, the *Sobolev space* of order r, consists of the distributions $u \in \mathcal{S}'(\mathbb{R}^d)$ such that $\langle D_x \rangle^r u \in L^2(\mathbb{R}^d)$.

The weighted space $L^2(\mathbb{R}^d; \langle x \rangle^r)$ is that of the functions $u \in L^1_{\mathrm{loc}}(\mathbb{R}^d)$ such that

$$\|u\|_{L^2(\mathbb{R}^d;\langle x\rangle^r)} := \left(\int_{\mathbb{R}^d} |u(x)|^2 \langle x \rangle^r dx \right)^{1/2} < \infty.$$

The analogous definition is understood for $L^\infty(\mathbb{R}^d; \langle x \rangle^r)$.

By $C^\infty(\mathbb{R}^d; \langle x \rangle^r)$ we intend the space of functions $u \in C^\infty(\mathbb{R}^d)$ such that

$$\|\partial_x^\alpha u\|_{L^\infty(\mathbb{R}^d;\langle x\rangle^r)} < \infty \quad \text{for every multi-index } \alpha \in \mathbb{N}^d.$$

$C_0(\mathbb{R}^d)$ denotes the spaces of continuous functions on $\mathbb{R}^d$ vanishing at infinity.

Given an open set $\Omega \subseteq \mathbb{R}^d$, $\mathcal{M}_+(\Omega)$ is the set of *positive Radon measures* on Ω, which can be identified through Riesz's theorem to the set of positive functionals on $C_c(\Omega)$, the space of continuous functions on Ω with compact support.

In order to lighten our writing, we shall write $\mathcal{S}$ and $\mathcal{S}'$ instead of $\mathcal{S}(\mathbb{R}^d_x \times \mathbb{R}^d_\xi)$ and $\mathcal{S}'(\mathbb{R}^d_x \times \mathbb{R}^d_\xi)$, respectively.

For a measurable function $f : \mathbb{R}^d \to \mathbb{C}$, we use the notation

$$D_f := \{x \in \mathbb{R}^d : f \text{ is not continuous at } x\}.$$

An important, perhaps nonstandard, definition is that of a scale.

DEFINITION 2.1. *A* scale (ε_k) *is a sequence of positive numbers that tends to zero as* $k \to \infty$.

Given two scales (h_k) and (ε_k), the notation $h_k \ll \varepsilon_k$ and $h_k \sim \varepsilon_k$ will be used to indicate that $\lim_{k\to\infty} h_k/\varepsilon_k = 0$ and $\lim_{k\to\infty} h_k/\varepsilon_k = c > 0$, respectively.

Finally, we shall always denote

$$Q := [-\pi, \pi)^d.$$

3. Sampling and reconstruction.

3.1. Definitions and examples. We now describe the sampling and reconstruction operators we are going to consider. Given a distribution $\varphi \in \mathcal{S}'(\mathbb{R}^d)$ we set, for every $n \in \mathbb{Z}^d$ and $h > 0$,

$$\varphi_n^h(x) := \varphi\left(\frac{x}{h} - n\right).$$

The *reconstruction* (or *synthesis*) *operator* T_φ^h, acting on discrete functions U of $\mathbb{Z}^d$, is defined to be

$$T_\varphi^h U(x) := \sum_{n\in\mathbb{Z}^d} U_n \varphi_n^h(x). \tag{3.1}$$

This expression is well-defined for finitely supported discrete functions. When φ is a continuous function such that $\varphi(0) = 1$ and $\varphi(k) = 0$ for $k \in \mathbb{Z}^d \setminus \{0\}$, then $T_\varphi^h U$ is actually a function that *interpolates* the discrete values U_n on the grid $h\mathbb{Z}^d$, i.e., $T_\varphi^h U(hn) = U_n$ for all $n \in \mathbb{Z}^d$.

Analogously, the *sampling* (or *analysis*) *operator* S_φ^h, a priori only acting on functions $u \in \mathcal{S}(\mathbb{R}^d)$, is defined as follows: $S_\varphi^h u$ is the discrete function given by

$$S_\varphi^h u(n) := h^{-d} \langle \overline{\varphi_n^h}, u \rangle_{\mathcal{S}'(\mathbb{R}^d)\times\mathcal{S}(\mathbb{R}^d)}.$$

When $\varphi = \delta_0$, we obtain the usual *discretization* operator: $S_{\delta_0}^h u(n) = u(hn)$ for every $n \in \mathbb{Z}^d$.

Indeed, these sampling/reconstruction schemes include several well-known procedures on regular grids. Among many others we may cite the following:

- *Cardinal B-splines.* The B-spline of order zero is the function $\varphi(x) := \mathbf{1}_{[-1/2,1/2]^d}(x)$; the function $T_\varphi^h U$ is just the piecewise constant interpolation of the discrete function U on the grid $h\mathbb{Z}^d$. The B-spline of order 1,

$$\varphi(x) = \mathbf{1}_{[-1/2,1/2]^d} * \mathbf{1}_{[-1/2,1/2]^d} = \prod_{j=1}^d (1 - |x_j|)_+,$$

gives rise to the piecewise linear interpolation operator. Analogously, B-splines of order $r \in \mathbb{N}$ are defined iterating this convolution r times. These are $C^{r-1}(\mathbb{R}^d)$ functions supported in $[-r/2, r/2]^d$, taking the value 1 at the origin. More details may be found, for instance, in [2].

- *Band-limited sampling/reconstruction.* This corresponds to the profile

$$\varphi(\xi) := \prod_{j=1}^{d} \operatorname{sinc}(\xi_j),$$

where the *cardinal sine function* is defined by

$$\operatorname{sinc}(t) := \frac{\sin \pi t}{\pi t}.$$

It is easy to check that $\widehat{\varphi}(\xi) = \mathbf{1}_Q(\xi)$. This profile is relevant because of *Shannon's sampling theorem*: *a function u belongs to the space*

$$V^h := \{u \in L^2(\mathbb{R}^d) : \operatorname{supp} \widehat{u} \subset [-\pi/h, \pi/h)^d\} = \operatorname{range}(T_\varphi^h)$$

if and only if

$$u = \sum_{n \in \mathbb{Z}^d} u(hn) \varphi_n^h.$$

In particular, such functions are determined by their values on the grid $h\mathbb{Z}^d$.

- *Wavelets.* Take $h_k := 2^{-k}$ for every $k \in \mathbb{Z}$. A function $\psi \in L^2(\mathbb{R}^d)$ is a wavelet, provided $\{\psi_n^{h_k} : n \in \mathbb{Z}^d,\ k \in \mathbb{Z}\}$ is an orthonormal basis of $L^2(\mathbb{R}^d)$. For more details on wavelets and the closely related *multiresolution analyses*, the reader may see [10, 16].

Additional examples and references (from the viewpoint of signal theory) may be found in the survey [19].

3.2. Boundedness properties. In order to ensure that the sampling and reconstruction operators are bounded, we shall make the assumption (BP_s) below:

$$\text{(BP_s)} \qquad \begin{gathered} \varphi \in H^s(\mathbb{R}) \text{ and, for some } B > 0, \\ \tau_{\langle D_x \rangle^s \varphi}(\xi) := \sum_{k \in \mathbb{Z}^d} |\langle \xi + 2\pi k \rangle^s \widehat{\varphi}(\xi + 2\pi k)|^2 \le B \quad \text{for a.e. } \xi \in \mathbb{R}^d. \end{gathered}$$

LEMMA 3.1. *Suppose $\varphi \in \mathcal{S}'(\mathbb{R}^d)$. Then the following are equivalent:*

(i) *φ satisfies* (BP_s).

(ii) *There exists $B > 0$ such that*

$$\|\langle hD_x \rangle^s T_\varphi^h U\|_{L^2(\mathbb{R}^d)} \le \sqrt{B} \|U\|_{L^2(h\mathbb{Z}^d)} \tag{3.2}$$

holds uniformly for $h > 0$ and $U \in L^2(h\mathbb{Z}^d)$.

(iii) *There exists $B > 0$ such that*

$$\|S_\varphi^h u\|_{L^2(h\mathbb{Z}^d)} \le \sqrt{B} \|\langle hD_x \rangle^{-s} u\|_{L^2(\mathbb{R}^d)} \tag{3.3}$$

holds uniformly for $h > 0$ and $u \in H^{-s}(\mathbb{R}^d)$.

Moreover, whenever (i), (ii), *or* (iii) *is fulfilled, the smallest constant B for which any of the above assertions holds is precisely $\|\tau_{\langle D_x\rangle^s\varphi}\|_{L^\infty(Q)}$.*

Proof. To see why (i) and (ii) are equivalent, first observe that, given any $\varphi \in H^s(\mathbb{R}^d)$, the following identity holds:

$$T^h_\varphi = \langle hD_x\rangle^{-s} T^h_{\langle D_x\rangle^s\varphi}. \tag{3.4}$$

To check this, simply notice that

$$\widehat{T^h_\varphi U}(\xi) = h^d\widehat{\varphi}(h\xi)\sum_{n\in\mathbb{Z}^d} U_n e^{-ihn\cdot\xi} = \widehat{\varphi}(h\xi)h^d\widehat{U}(h\xi),$$

and hence

$$\widehat{T^h_\varphi U}(\xi) = \langle h\xi\rangle^{-s}\langle h\xi\rangle^s\widehat{\varphi}(h\xi)h^d\widehat{U}(h\xi) = \langle hD_x\rangle^{\widehat{-s}}\widehat{T^h_{\langle D_x\rangle^s\varphi}U}(\xi).$$

Since $\langle D_x\rangle^s\varphi \in L^2(\mathbb{R}^d)$ and

$$\|\langle hD_x\rangle^s T^h_\varphi U\|_{L^2(\mathbb{R}^d)} = \|T^h_{\langle D_x\rangle^{-s}\varphi}U\|_{L^2(\mathbb{R}^d)}$$

it suffices to deal with the case $s = 0$. But it is a well-known result (see, for instance, [3, 17]) that for $\varphi \in L^2(\mathbb{R}^d)$, (i) and (ii) are equivalent and that $\|T^h_\varphi\| = \|\tau_\varphi\|_{L^\infty(\mathbb{R}^d)}$ whenever T^h_φ is bounded.

Statements (ii) and (iii) are equivalent because of the following duality relation:

$$(\langle hD_x\rangle^s T^h_\varphi U, \langle hD_x\rangle^{-s}u)_{L^2(\mathbb{R}^d)} = (U, S^h_\varphi u)_{L^2(h\mathbb{Z}^d)},$$

which holds for every $u \in H^{-s}(\mathbb{R}^d)$ and $U \in L^2(h\mathbb{Z}^d)$. This is simple to check:

$$\begin{aligned}(\langle hD_x\rangle^s T^h_\varphi U, \langle hD_x\rangle^{-s}u)_{L^2(\mathbb{R}^d)} &= \sum_{n\in\mathbb{Z}^d} U_n\int_{\mathbb{R}^d}\langle hD_x\rangle^s\varphi^h_n(x)\overline{\langle hD_x\rangle^{-s}u(x)}dx\\ &= \sum_{n\in\mathbb{Z}^d} U_n\langle\varphi^h_n,\overline{u}\rangle_{H^s(\mathbb{R}^d)\times H^{-s}(\mathbb{R}^d)}\\ &= h^d\sum_{n\in\mathbb{Z}^d} U^h_n\overline{S^h_\varphi u(n)}. \qquad \square\end{aligned}$$

Remark 3.2. For $s \le 0$, estimate (3.2) implies that

$$\|\langle\varepsilon D_x\rangle^s T^h_\varphi U^h\|_{L^2(\mathbb{R}^d)} \le \sqrt{B}\|U^h\|_{L^2(h\mathbb{Z}^d)}, \tag{3.5}$$

as soon as $h/\varepsilon \le 1$, as can be easily checked by taking Fourier transforms.

A sufficient condition for (BP_s) in terms of decay on φ is given next.

LEMMA 3.3. *Suppose $\varphi \in H^s(\mathbb{R}^d)$ satisfies, for some $\varepsilon > 0$,*

$$\int_{\mathbb{R}^d}|\langle D_x\rangle^s\varphi(x)|^2(1+|x|)^{d+\varepsilon}dx < \infty. \tag{3.6}$$

Then $\widehat{\varphi}$ and $\tau_{\langle D_x\rangle^s\varphi}$ are continuous functions. In particular, (BP_s) *always holds for such a φ.*

Proof. It follows along the lines of [16, Lemma II.7]. Under condition (3.6), $\langle\xi\rangle^s\widehat{\varphi} \in H^{d/2+\varepsilon/2}(\mathbb{R}^d)$; Sobolev's imbedding theorem then ensures that $\langle\xi\rangle^s\widehat{\varphi}$ is a

continuous function, and hence so is $\widehat{\varphi}$. The continuity of $\tau_{\langle D_x\rangle^s\varphi}$ is a consequence of the fact that, whenever $\chi \in C_c^\infty(\mathbb{R}^d)$ satisfies $\sum_{n\in\mathbb{Z}^d} |\chi(\xi+2\pi n)| \geq 1$, the expression

$$\left[\sum_{n\in\mathbb{Z}^d} \|u\chi(\cdot+2\pi n)\|^2_{H^s(\mathbb{R}^d)}\right]^{1/2}$$

defines an equivalent norm in $H^s(\mathbb{R}^d)$, $s \geq 0$. This actually proves that

$$\sum_{n\in\mathbb{Z}^d} \sup_{\xi\in\mathbb{R}^d} |\langle\xi\rangle^s \widehat{\varphi}(\xi)\chi(\xi+2\pi n)|^2 < \infty.$$

In particular, the series defining $\tau_{\langle D_x\rangle^s\varphi}$ is uniformly convergent and the claim then follows. □

Condition (3.6) automatically holds for profiles φ such that

$$|\langle D_x\rangle^s\varphi(x)| \leq C(1+|x|)^{-d-\varepsilon} \quad \text{for every } x\in\mathbb{R}^d \text{ and some } C,\varepsilon>0; \tag{3.7}$$

in particular, the hypothesis (1.8) we assumed in the introduction implies (BP_s) for $s=0$.

Now we can prove a general result from which Proposition 1.8 immediately follows.

PROPOSITION 3.4. *Suppose φ satisfies* (BP_s) *and we are given scales (h_k), (ε_k) such that (h_k/ε_k) is bounded. If (U^{h_k}) is an h_k-bounded sequence of discrete functions, then the distributions $m^{\varepsilon_k}[T^{h_k}_\varphi U^{h_k}]$ are uniformly bounded in $\mathcal{S}'$. Moreover, the limit of any weakly convergent subsequence is a positive measure.*

The proof of this is a direct consequence of Remark 3.2 and the general result established in Proposition 8.1.

3.3. Bases and projections. Below, we recall some results from approximation theory that will be needed in what follows. These results deal with the range in $H^s(\mathbb{R}^d)$ of the reconstruction operator T^h_φ; we denote this space by V^h_φ and assume that it is equipped with the (equivalent) norm $\|\langle hD_x\rangle^s\cdot\|_{L^2(\mathbb{R}^d)}$.

The space V^h_φ is a *principal shift invariant* (PSI) space. When any of the conditions of Lemma 3.1 are satisfied, the family $\{h^{-d/2}\varphi^h_n : n\in\mathbb{Z}^d\}$ is said to form a *Bessel system* for V^h_φ.

The next lemma clarifies how the function $\tau_{\langle D_x\rangle^s\varphi}$ characterizes further basis properties of the functions φ^h_n.

LEMMA 3.5. *Let $\varphi\in\mathcal{S}'(\mathbb{R}^d)$ satisfy* (BP_s). *Then*

(i) *$\{h^{-d/2}\varphi^h_n : n\in\mathbb{Z}^d\}$ is an orthonormal basis of V^h_φ if and only if*

$$\tau_{\langle D_x\rangle^s\varphi}(\xi) = 1 \quad \textit{for a.e. } \xi\in\mathbb{R}^d;$$

(ii) *$\{h^{-d/2}\varphi^h_n : n\in\mathbb{Z}^d\}$ is a* Riesz basis[6] *of V^h_φ if and only if there exist constants A, $B>0$ such that*

$$A \leq \tau_{\langle D_x\rangle^s\varphi}(\xi) \leq B \quad \textit{for a.e. } \xi\in\mathbb{R}^d.$$

[6]This means that there exist constants $A,B>0$ such that

$$A\|U\|^2_{L^2(h\mathbb{Z}^d)} \leq \|T^h_\varphi U\|^2_{H^s(\mathbb{R}^d)} \leq B\|U\|^2_{L^2(h\mathbb{Z}^d)}$$

for all $U\in L^2(h\mathbb{Z}^d)$. This is equivalent to the existence of a linear isomorphism $R: V^h \to V^h$ such that $\{h^{-d/2}R\varphi^h_n : n\in\mathbb{Z}^d\}$ forms an orthonormal basis of $H^s(\mathbb{R}^d)$. This property is sometimes also referred to as $(\varphi^h_n)_{n\in\mathbb{Z}^d}$ forming a *stable frame* in $H^s(\mathbb{R}^d)$.

Proof. It follows from (3.4) that the operator $\langle hD_x\rangle^s$ is a unitary isomorphism from V_φ^h onto the range of $T^h_{\langle D_x\rangle^s\varphi}$. Hence $\{h^{-d/2}\varphi_n^h : n\in\mathbb{Z}^d\}$ is an orthonormal (resp., Riesz) basis of V_φ^h if and only if $\{h^{-d/2}(\langle D_x\rangle^s\varphi)_n^h : n\in\mathbb{Z}^d\}$ is an orthonormal (resp., Riesz) basis of $V^h_{\langle D_x\rangle^s\varphi}$. Thus, the lemma need only be proved for profiles $\varphi\in L^2(\mathbb{R}^d)$; this is a well-known result (see, for instance, [17]). ∎

We shall also need the following expression for the orthogonal projection onto V_φ^h.

LEMMA 3.6. *Let $\varphi\in\mathcal{S}'(\mathbb{R}^d)$ satisfy* (BP_s). *The orthogonal projection $P_\varphi^h : H^s(\mathbb{R}^d)\to V_\varphi^h$ equals $P_\varphi^h = T_\varphi^h S^h_{\widetilde{\langle D_x\rangle^s\varphi}}\langle hD_x\rangle^s$, where, for $f\in L^2(\mathbb{R}^d)$, $\tilde f\in L^2(\mathbb{R}^d)$ is defined by*

$$\widehat{\tilde f}(\xi) := \begin{cases} \dfrac{\widehat f(\xi)}{\tau_f(\xi)} & \text{if } \tau_f(\xi)\neq 0,\\ 0 & \text{otherwise.}\end{cases}$$

Proof. The proof of the result for $s=0$ may be found in [3, Theorem 2.9]. We limit ourselves to this case by noticing that

$$P_\varphi^h = \langle hD_x\rangle^{-s}P^h_{\langle hD_x\rangle^s\varphi}\langle hD_x\rangle^s,$$

since, as we have seen in (3.4), the range of T_φ^h equals that of $\langle hD_x\rangle^{-s}T^h_{\langle hD_x\rangle^s\varphi}$, and $\langle hD_x\rangle^s$ is an orthogonal mapping. Using the L^2-result we obtain

$$P_\varphi^h = \langle hD_x\rangle^{-s}T^h_{\langle hD_x\rangle^s\varphi}S^h_{\widetilde{\langle D_x\rangle^s\varphi}}\langle hD_x\rangle^s = T_\varphi^h S^h_{\widetilde{\langle D_x\rangle^s\varphi}}\langle hD_x\rangle^s,$$

as claimed. ∎

4. High-frequency analysis: $h\sim\varepsilon$.

4.1. Reduction to the case $h=\varepsilon$. In this section we analyze the effect of sampling and reconstruction on Wigner measures at scales (ε_k) of the same order of the sampling/reconstruction rate (h_k) (i.e., such that (h_k/ε_k) is bounded).

First notice that it suffices to treat the case $\varepsilon_k = h_k$; the more general case can be obtained by a proper rescaling. This is due to the identity

$$m^\varepsilon[u](x,\xi) = (h/\varepsilon)^d m^h[u](x,(h/\varepsilon)\xi),$$

which clearly implies the following lemma.

LEMMA 4.1. *Suppose $h_k/\varepsilon_k\to c>0$. Then $m^{\varepsilon_k}[u_k]$ converges in $\mathcal{S}'$ if and only if $m^{h_k}[u_k]$ does. Their respective limits μ_c and μ are related through*

$$\mu_c(x,\xi) = c^d\mu(x,c\xi). \tag{4.1}$$

When $h_k=\varepsilon_k$, the transforms $M^{h_k}[U^{h_k}]$ are Γ-periodic in the variable ξ; hence, so are their limiting Wigner measures.

4.2. Sampling. We start by exploring the effect of sampling on the structure of Wigner measures. The computation of the Wigner measure at scale (h_k) of a sequence of samples $(S_\varphi^{h_k}u_k)$ is done in the following theorem; it is applicable whenever the hypothesis (D) below is fulfilled:

$$\operatorname*{essup}_{\xi\in Q}\sum_{|n|\geq R}|\langle\xi+2\pi n\rangle^s\widehat\varphi(\xi+2\pi n)|^2\to 0 \quad\text{as } R\to\infty. \tag{D}$$

Notice that profiles with the property (3.6) immediately verify (D).

Before stating our result, it is important to notice that the Fourier transform of a profile φ satisfying condition (BP_s) is an element of $L^2_{\mathrm{loc}}(\mathbb{R}^d)$. In particular, it is only defined modulo sets of zero Lebesgue measure. Thus, when dealing with pointwise properties of $\widehat{\varphi}$, we shall systematically assume that a precise representative of the class of $\widehat{\varphi}$ has, once and for all, been chosen.

For instance, the Wigner measures μ of the sequences (u_k) in Theorem 4.2 below will be assumed to satisfy conditions (MS) and (ND):

(MS) $\quad |\widehat{\varphi}(\xi+2\pi n)|^2\mu(x,\xi+2\pi n), \quad n\in\mathbb{Z}^d, \quad$ are mutually singular measures.

(ND) $$\mu(\mathbb{R}^d\times\overline{D_{\widehat{\varphi}}})=0,$$

where, recall, $D_{\widehat{\varphi}}$ stands for the set of discontinuity points of $\widehat{\varphi}$. These conditions must be understood to hold for the same representative of $\widehat{\varphi}$.

THEOREM 4.2. *Let (h_k) be a scale and take φ satisfying* (BP_s) *and* (D). *Let (u_k) be a sequence in $H^{-s}(\mathbb{R}^d)$ such that $(\langle h_kD_x\rangle^{-s}u_k)$ is bounded, and suppose that $m^{h_k}[u_k]$ converges to a Wigner measure μ that fulfills* (ND), (MS).

Then $M^{h_k}[S^{h_k}_{\varphi}u_k]$ converges to the Wigner measure μ^{φ} given by

$$\mu^{\varphi}(x,\xi)=\sum_{n\in\mathbb{Z}^d}|\widehat{\varphi}(\xi+2\pi n)|^2\mu(x,\xi+2\pi n). \tag{4.2}$$

Remark 4.3. (i) As pointed out above, formula (4.6) holds for the same precise representative of the Fourier transform $\widehat{\varphi}$ which was chosen in (ND) and (MS).

(ii) The necessity of hypotheses (ND) and (MS) will be discussed in section 4.4.

(iii) Condition (D) may be replaced by the assumption that $(\langle h_kD_x\rangle^{-s}u_k)$ is h_k-oscillatory. This will be made clear in the proof of the theorem.

The proof of this theorem is postponed to the end of this section.

The expression (4.2) may be related to the concept of *Wigner series* introduced in [14, 9]. Recall that given $u\in\mathcal{S}'(\mathbb{R}^d)$, the *Wigner series* of u at scale ε is defined by

$$w^{\varepsilon}_S[u](x,\xi):=\frac{1}{(2\pi)^d}\sum_{n\in\mathbb{Z}^d}u(x-\varepsilon\pi n)\overline{u}(x+\varepsilon\pi n)e^{in\cdot\xi}.$$

It is easy to check that $w^{\varepsilon}_S[u](x,\xi)=\sum_{n\in\mathbb{Z}^d}w^{\varepsilon}[u](x,\xi+2\pi n)$.[7]

When (u_k) is bounded in $L^2(\mathbb{R}^d)$, is ε_k-oscillatory, and possesses a Wigner measure at scale (ε_k), then the following relation holds:

$$\lim_{k\to\infty}\int_{\mathbb{R}^d\times\mathbb{R}^d}a(x,\xi)w^{\varepsilon_k}_S[u_k](x,\xi)dxd\xi=\int_{\mathbb{R}^d\times\mathbb{R}^d}\sum_{n\in\mathbb{Z}^d}a(x,\xi+2\pi n)d\mu(x,\xi) \tag{4.3}$$

for $a\in\mathcal{S}$; see [1].

Theorem 4.2 has a simple interpretation in terms of Wigner series: the measure μ^{φ} may be obtained as the limit of the Wigner series

$$w^{h_k}_S[\widehat{\varphi}(h_kD_x)u_k].$$

[7] See (1.4) for the definition of the Wigner transform $w^{\varepsilon}[u]$.

This is due to the fact that, under any of the hypotheses (D), $(\widehat{\varphi}(h_k D_x)u_k)$ is h_k-oscillatory. Besides, as a consequence of Proposition 8.3, the Wigner measure at scale (h_k) of $(\widehat{\varphi}(h_k D_x)u_k)$ is given by $|\widehat{\varphi}(\xi)|^2\mu(x,\xi)$. The assertion then follows from (4.3).

As was already mentioned in the introduction, condition (MS) is a restriction on the support of the measure $|\widehat{\varphi}(\xi)|^2\mu(x,\xi)$. Two extremal cases in which it is trivially satisfied are the following:

(i) $\widehat{\varphi}|_{\mathbb{R}^d\setminus Q} \equiv 0$; in this case (MS) holds independently of what μ is.

(ii) The sequence (u_k) is *asymptotically band-limited*; i.e., its Wigner measures at scale (h_k) are concentrated on the cube $\overline{Q}$. For those sequences, condition (MS) only involves the behavior of μ on the boundary ∂Q: it essentially expresses that the restrictions of μ to parallel sides of ∂Q do not overlap (i.e., are mutually singular). A sufficient condition for this is, for instance,

$$\limsup_{k\to\infty} \int_{\mathbb{R}^d\setminus Q_R} \left|\widehat{u_k}\left(\frac{\xi}{h_k}\right)\right|^2 \frac{d\xi}{(2\pi h_k)^d} \to 0 \quad \text{as } R\to\infty, \tag{4.4}$$

where $Q_R := [-\pi, \pi - 1/R)^d$.

Remark 4.4. In any of the above cases, we have

$$\mathbf{1}_Q(\xi)\mu^{\varphi}(x,\xi) = |\widehat{\varphi}(\xi)|^2\mu(x,\xi).$$

Hence, the restriction of μ^{φ} to $\mathbb{R}^d \times Q$ coincides with μ if and only if $|\widehat{\varphi}(\xi)|^2 = 1$ for μ-almost every $\xi \in \overline{Q}$.

The specific choice $\varphi = \delta_0$ corresponds to the analysis of *discretization*, for then $S^h_{\delta_0}u(n) = u(hn)$. Theorem 4.2 takes the following simple form.

COROLLARY 4.5. *Let (h_k) be a scale and let (u_k) be a sequence in $H^s(\mathbb{R}^d)$, for some $s > d/2$, such that $(\langle h_k D_x\rangle^s u_k)$ is bounded. If μ is its Wigner measure at scale (h_k) and the measures $\mu(x,\xi+2\pi n)$ are mutually singular, then Wigner measure μ^{δ_0} corresponding to the sequence of discretizations is the periodization:*

$$\mu^{\delta_0}(x,\xi) = \sum_{n\in\mathbb{Z}^d} \mu(x,\xi+2\pi n).$$

In other words, μ^{δ_0} is the limit of the Wigner series $w^{h_k}_S[u_k]$.

This corollary is particularly useful in the explicit computation of Wigner measures for discrete functions. As an example, consider the concentrating and oscillating sequences we defined in the introduction, $f_k(x) = k^{d/2}\rho(k(x-x_0))$ and $g_k(x) := \rho(x)e^{ikx\cdot\xi^0}$ with $\rho \in L^2(\mathbb{R}^d)$. Using identities (1.6) and (1.7) we obtain, for (f_k) and (g_k), respectively,

$$\mu^{\delta_0}(x,\xi) = \delta_{x_0}(x) \otimes \sum_{n\in\mathbb{Z}} |\widehat{\rho}(\xi+2\pi n)|^2 \frac{d\xi}{(2\pi)^d}$$

if, for instance, $\operatorname{supp}\widehat{\rho} \subset Q$, and

$$|\rho(x)|^2 dx \otimes \sum_{n\in\mathbb{Z}^d} \delta_{\xi^0+2\pi n}(\xi), \tag{4.5}$$

with no assumption on ρ.

4.3. Reconstruction. Now we deal with the reconstruction operator T^h_φ; it modifies the high-frequency behavior of a sequence of discrete functions in the following way.

THEOREM 4.6. *Let (h_k) be a scale and (U^{h_k}) an h_k-bounded sequence; take φ satisfying* (BP_s). *If $M^{h_k}[U^{h_k}]$ converges to the Wigner measure μ which verifies* (ND), *then $m^{h_k}[T^{h_k}_\varphi U^{h_k}]$ converges to a Wigner measure μ_φ given by*

$$\mu_\varphi(x,\xi) = |\widehat{\varphi}(\xi)|^2 \mu(x,\xi). \tag{4.6}$$

The proof of Theorem 4.6 is based on explicit formulas for the Fourier transforms of $T^h_\varphi U$. As we have already seen,

$$\widehat{T^h_\varphi U}(\xi) = \widehat{\varphi}(h\xi) h^d \widehat{U}(h\xi) \tag{4.7}$$

for any $U \in L^2(h\mathbb{Z}^d)$. The following remark ensures that Proposition 8.3 can be applied in the proof below.

Remark 4.7. If $\varphi \in H^s(\mathbb{R}^d)$ satisfies (BP_s), then $\widehat{\varphi} \in L^\infty(\mathbb{R}^d; \langle\xi\rangle^s)$.

Proof of Theorem 4.6. Just notice that (4.7) can be rewritten as

$$T^h_\varphi U^h = \widehat{\varphi}(hD_x) T^h_{\delta_0} U^h.$$

The hypotheses made on φ and μ allow us to apply Proposition 8.3 (see Remark 4.7) and conclude the proof. □

Identity (4.6) expresses how the measure μ is modulated by the profile φ; the necessity of the hypothesis (ND) for this result is discussed in section 4.4 as well.

Since μ is Γ-periodic in ξ, formula (4.6) suggests that μ may be compared to the periodization of μ_φ with respect to the variable ξ.

COROLLARY 4.8. *Let φ, (U^{h_k}), μ, and μ_φ be as in Theorem* 4.6. *Then the periodization*

$$\mu_{\varphi,s}(x,\xi) := \sum_{n\in\mathbb{Z}^d} \langle \xi + 2\pi n\rangle^{2s} \mu_\varphi(x, \xi + 2\pi n) \tag{4.8}$$

is a well-defined[8] *measure, Γ-periodic in ξ, that satisfies*

$$\mu_{\varphi,s}(x,\xi) = \tau_{\langle D_x\rangle^s \varphi}(\xi)\mu(x,\xi). \tag{4.9}$$

In particular,

(i) *if $\tau_{\langle D_x\rangle^s\varphi}(\xi) = 1$ except for ξ in a set of zero μ-measure, then $\mu_{\varphi,s} = \mu$;*

(ii) *$\tau_{\langle D_x\rangle^s\varphi} \equiv 1$ if and only if the identity $\mu_{\varphi,s} = \mu$ holds for every sequence (U^{h_k}).*

Proof. Since $|\langle\xi\rangle^s \widehat{\varphi}(\xi)|^2$ is a nonnegative continuous function, the series defining $\tau_{\langle D_x\rangle^s\varphi}(\xi)$ converges absolutely for every ξ on the support of μ (which consists of continuity points for $\widehat{\varphi}(\xi)$). Thus, by the dominated convergence theorem,

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} a(x,\xi)\tau_{\varphi,s}(\xi)d\mu(x,\xi) = \sum_{n\in\mathbb{Z}^d} \int_{\mathbb{R}^d\times\mathbb{R}^d} a(x,\xi)|\langle\xi+2\pi n\rangle^s \widehat{\varphi}(\xi+2\pi n)|^2 d\mu(x,\xi)$$

[8] The limit defining the sum (4.8) is understood to exist for the weak convergence of measures in $\mathcal{M}_+(\mathbb{R}^d\times\mathbb{R}^d)$.

for every $a \in C_c(\mathbb{R}^d \times \mathbb{R}^d)$. Now, taking into account (4.6) and the fact that μ is Γ-periodic in ξ, we find that

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} a(x,\xi)\tau_{\varphi,s}(\xi)d\mu(x,\xi) = \sum_{n\in\mathbb{Z}^d}\int_{\mathbb{R}^d\times\mathbb{R}^d} a(x,\xi)\langle\xi+2\pi n\rangle^{2s}d\mu_\varphi(x,\xi+2\pi n),$$

and the first part of the result follows.

Statement (i) as well as the "only if" part of (ii) are trivial. To obtain the necessity in (ii), just consider sequences of discrete functions whose Wigner measures are of the form $\mu(x,\xi) = \nu(x)\otimes\sum_{n\in\mathbb{Z}^d}\delta_{\xi^0+2\pi n}$ (as (4.5), for instance). Clearly, for $\mu_{\varphi,s} = \mu$ to hold for such a measure, we must have $\tau_{\langle D_x\rangle^s\varphi}(\xi^0) = 1$. □

Remark 4.9. (i) Because of Lemma 3.5, if relation $\mu_{\varphi,s} = \mu$ holds for every h_k-bounded sequence of discrete functions, then the profile φ has the following property: $\{h^{-d/2}\varphi_n^h : n \in \mathbb{Z}^d\}$ is an orthonormal family in $H^s(\mathbb{R}^d)$ for every $h > 0$.

(ii) However, the converse is not true: if φ gives rise to an orthonormal family, then $\tau_{\langle D_x\rangle^s\varphi}(\xi) = 1$ holds outside a set of null Lebesgue measure. If μ is supported on that set, identity $\mu_{\varphi,s} = \mu$ may not hold.

As in the preceding section, our result has an interpretation in terms of Wigner series. Under the conditions of Theorem 4.6, the measure $\mu_{\varphi,s}$ may be obtained as the limit of the functions

$$w_S^{h_k}[\langle h_k D_x\rangle^s T_\varphi^{h_k} U^{h_k}],$$

as $k\to\infty$, provided $(\langle h_k D_x\rangle^s T_\varphi^{h_k} U^{h_k})$ is h_k-oscillatory. Note, however, that this may not be the case for certain profiles φ (see section 5.3).

In particular, Corollary 4.8 shows that the limits of $w_S^{h_k}[T_\varphi^{h_k}U^{h_k}]$ and $M^{h_k}[U^{h_k}]$ coincide if we choose, for instance, $\varphi := \mathbf{1}_{[-1/2,1/2)^d}$.

4.4. The necessity of the hypotheses of Theorems 4.2 and 4.6. Formulas (4.6) and (4.2) may not hold when $\widehat{\varphi}$ is not continuous and the Wigner measure μ does not vanish on the closure of the set of discontinuity points $D_{\widehat{\varphi}}$. We illustrate this with two one-dimensional examples where

$$\varphi(x) = \frac{\sin\pi x}{\pi x}.$$

We will chose $\mathbf{1}_Q$ as the representative of $\widehat{\varphi}$ for which the counterexamples will be built.

1. *Necessity of condition* (ND) *in Theorem* 4.6. Take U^h to be the sequence discrete function of $L^2(h\mathbb{Z}^d)$ given by their Fourier transforms:

$$\widehat{U^h}(\xi) := \frac{1}{h}\sum_{n\in\mathbb{Z}}\mathbf{1}_{(-1,1)}\left(\frac{\xi-(2n+1)\pi}{h}\right).$$

Then, denoting by μ the Wigner measure at scale h of (U^h),

$$|\widehat{\varphi}(\xi)|^2\mu(x,\xi) = \frac{\sin^2(x)}{\pi^2x^2}dx\otimes\delta_{-\pi}(\xi).$$

This measure differs from μ_φ, which is given by

$$\mu_\varphi(x,\xi) = \frac{\sin^2(x/2)}{\pi^2x^2}dx\otimes[\delta_\pi(\xi)+\delta_{-\pi}(\xi)].$$

Remark 4.10. (i) The particular choice of the representative of $\widehat{\varphi}$ does not play a role. Theorem 4.6 still fails if we take as representative of $\widehat{\varphi}$ the characteristic functions of $(-\pi,\pi)^d$ or $[-\pi,\pi]^d$.

(ii) In particular, this example shows that even the two projections on x and ξ of the measures μ and μ_φ may differ.

(iii) This also shows that the periodization in ξ of μ_φ does not necessarily coincide with μ, even when $\tau_\varphi = 1$, as is the case here. Thus the conclusion of Corollary 4.8 may fail when $\widehat{\varphi}$ is not continuous.

Our following counterexample to Theorem 4.2 is based on the same principle.

2. *Necessity of condition* (ND) *in Theorem* 4.2. Define

$$\widehat{v^h}(\xi) := \mathbf{1}_{(-1,1)}(\xi + \pi/h).$$

Then, denoting by μ the Wigner measure at scale h of (v^h),

$$\sum_{n\in\mathbb{Z}} |\widehat{\varphi}(\xi+2\pi n)|^2 \mu(x,\xi+2\pi n) = \frac{\sin^2(x)}{\pi^2 x^2} dx \otimes \sum_{n\in\mathbb{Z}} \delta_{(2n+1)\pi}(\xi),$$

and this is different from μ^φ, which is precisely

$$\mu^\varphi(x,\xi) = \frac{\sin^2(x/2)}{\pi^2 x^2} dx \otimes \sum_{n\in\mathbb{Z}} \delta_{(2n+1)\pi}(\xi).$$

Finally, we investigate hypothesis (MS). Now we set $\varphi := \delta_0$.

3. *Necessity of condition* (MS) *in Theorem* 4.2. Define

$$\widehat{v^h}(\xi) := \mathbf{1}_Q(h\xi) \sum_{n\in\mathbb{Z}} \mathbf{1}_{(-1,1)}(\xi - (2n+1)\pi).$$

Clearly, as in our first example, the periodization of the Wigner measure of (v^h) is

$$\sum_{k\in\mathbb{Z}^d} \mu(x,\xi+2\pi n) = \frac{\sin^2(x/2)}{\pi^2 x^2} dx \otimes \sum_{k\in\mathbb{Z}} \delta_{(2n+1)\pi}(\xi).$$

However, the sequence of discretizations $(S^h_{\delta_0} v^h)$ has the following one:

$$\mu^{\delta_0}(x,\xi) = \frac{\sin^2(x)}{\pi^2 x^2} dx \otimes \sum_{n\in\mathbb{Z}} \delta_{(2n+1)\pi}(\xi).$$

The verification of these statements easily follows from (1.6), identity (8.5), and Lemma 8.13.

4.5. A Poisson summation formula and proof of Theorem 4.2. The computation of the Fourier transform of $S^h_\varphi u$ is given by the following identity.

LEMMA 4.11. *Let φ satisfy* (BP_s) *and let $u \in H^{-s}(\mathbb{R}^d)$. Then the Fourier transform of $S^h_\varphi u$ is*

$$h^d \sum_{n\in\mathbb{Z}^d} S^h_\varphi u(n) e^{-ihn\cdot\xi} = \sum_{n\in\mathbb{Z}^d} \overline{\widehat{\varphi}(h\xi+2\pi n)}\, \widehat{u}\left(\xi + \frac{2\pi}{h} n\right), \tag{4.10}$$

the convergence of the first series being in $L^2_{loc}(\mathbb{R}^d)$*, while the second takes place in* $L^1_{loc}(\mathbb{R}^d)$.

Proof. Begin by noticing that $\overline{\widehat{\varphi}}\widehat{u} \in L^1(\mathbb{R}^d)$, and thus

$$\Pi^h(\xi) := \sum_{n\in\mathbb{Z}^d} \overline{\widehat{\varphi}(h\xi + 2\pi n)}\widehat{u}(\xi + 2\pi/hn)$$

is a well-defined $(2\pi/h)\mathbb{Z}^d$-periodic $L^1_{\rm loc}(\mathbb{R}^d)$ function, the series defining it being absolutely convergent in $L^1_{\rm loc}(\mathbb{R}^d)$. We can compute its Fourier coefficients:

$$\begin{aligned}\int_{[-\pi/h,\pi/h)^d} \Pi^h(\xi)e^{ihn\cdot\xi}\frac{h^d d\xi}{(2\pi)^d} &= \sum_{k\in\mathbb{Z}^d}\int_Q \overline{\widehat{\varphi}(\xi+2\pi k)}\widehat{u}\left(\frac{\xi+2\pi k}{h}\right)e^{in\cdot\xi}\frac{d\xi}{(2\pi)^d}\\ &= \int_{\mathbb{R}^d}\overline{\widehat{\varphi}(\xi)}\widehat{u}\left(\frac{\xi}{h}\right)e^{in\cdot\xi}\frac{d\xi}{(2\pi)^d}\\ &= \int_{\mathbb{R}^d}\overline{h^d\widehat{\varphi}(h\xi)e^{-ihn\cdot\xi}}\widehat{u}(\xi)\frac{d\xi}{(2\pi)^d}\\ &= \langle\overline{\varphi_n^h}, u\rangle_{\mathcal{S}'\times\mathcal{S}} = h^d S^h_\varphi u(n).\end{aligned}$$

Lemma 3.1 proves that $S^h_\varphi u$ is square-summable and, consequently,

$$\Pi^h(\xi) = \sum_{n\in\mathbb{Z}^d} h^d S^h_\varphi u(n)e^{-ihn\cdot\xi},$$

the sum being understood in the L^2-sense. This is precisely formula (4.10). □

Remark 4.12. Identity (4.10) may be viewed as a generalization of the *Poisson summation formula*. Taking as φ the Dirac delta δ_0, we obtain

$$h^d\sum_{n\in\mathbb{Z}^d} u(hn)e^{-ihn\cdot\xi} = \sum_{n\in\mathbb{Z}^d}\widehat{u}\left(\xi+\frac{2\pi}{h}n\right)$$

for every $u\in H^s(\mathbb{R}^d)$ with $s > d/2$.

Proof of Theorem 4.2. The proof will be done in two steps.

Step 1. We first establish the result for sequences such that $\widehat{\varphi}(\xi)\widehat{u_k}(\xi/h_k)$ has support in a ball $B(0;R)$ for every $k\in\mathbb{N}$. We claim that the following formula holds:

$$T^{h_k}_{\delta_0} S^{h_k}_\varphi u_k(x) = \sum_{|n|\le R+\pi\sqrt{d}} e^{-2\pi i n\cdot x/h_k}\overline{\widehat{\varphi}}(h_k D_x)u_k(x).$$

This is obtained by applying the inverse Fourier transform to both sides of identity (4.10) and remarking that only summands satisfying $|n|\le R+\pi\sqrt{d}$ must be considered because of the condition on the support of $\widehat{\varphi}\,\widehat{u_k}(\cdot/h_k)$. The Wigner measures of the functions

$$e^{-2\pi i n\cdot x/h_k}\overline{\widehat{\varphi}}(h_k D_x)u_k(x)$$

are precisely (cf. Proposition 8.3 and Remark 4.7)

$$|\widehat{\varphi}(\xi+2\pi n)|^2\mu(x,\xi+2\pi n).$$

By hypothesis, they are mutually singular so, by Lemma 8.13, we deduce that the measure μ^φ obtained as the limit of $m^{h_k}[T^{h_k}_{\delta_0}S^{h_k}_\varphi u_k]$ is given by (4.2).

Step 2. We prove the result in the general case by taking advantage of hypothesis (D). Let $\chi \in C_c^\infty(\mathbb{R}^d)$ be a cut-off function identically equal to one in the unit ball $B(0;1)$. Denote by $S^{h_k}_{\varphi,R}u_k$ the truncation given by

$$\begin{aligned}\widehat{S^{h_k}_{\varphi,R}u_k}(\xi) &:= S^{h_k}_{\varphi}\chi\left(\widehat{\frac{h_kD_x}{R}}\right)u_k(\xi)\\ &= \frac{1}{(h_k)^d}\sum_{n\in\mathbb{Z}^d}\overline{\widehat{\varphi}(\xi+2\pi n)}\chi\left(\frac{\xi+2\pi n}{R}\right)\widehat{u_k}\left(\frac{\xi+2\pi n}{h_k}\right).\end{aligned}$$

Then, by the first step we have just proved, $M^{h_k}[S^{h_k}_{\varphi,R}u]$ converges to

$$\mu^{\varphi}_R(x,\xi) := \sum_{n\in\mathbb{Z}^d}\left|\widehat{\varphi}(\xi+2\pi n)\chi\left(\frac{\xi+2\pi n}{R}\right)\right|^2\mu(x,\xi+2\pi n). \tag{4.11}$$

We claim that (D) implies the following:

$$\limsup_{k\to\infty}\|S^{h_k}_{\varphi}u_k - S^{h_k}_{\varphi,R}u_k\|^2_{L^2(h_k\mathbb{Z}^d)} \to 0 \quad \text{as } R\to\infty. \tag{4.12}$$

It is sufficient to realize that

$$S^{h_k}_{\varphi}u_k - S^{h_k}_{\varphi,R}u_k = S^{h_k}_{\psi_R}u_k$$

for $\widehat{\psi_R} := \overline{\chi(\cdot/R)}\widehat{\varphi}$. The norm of $S^{h_k}_{\psi_R}$ is precisely (cf. Lemma 3.1)

$$\operatorname*{essup}_{\xi\in Q}\sum_{n\in\mathbb{Z}^d}\left|\langle\xi+2\pi n\rangle^s\widehat{\varphi}(\xi+2\pi n)\chi\left(\frac{\xi+2\pi n}{R}\right)\right|^2,$$

which tends to zero as $R\to 0$.

Lemma 8.12 then ensures that μ^{φ}_R weakly converge to μ^{φ}. Identity (4.11) means that

$$\int_{\mathbb{R}^d\times\mathbb{R}^d}a(x,\xi)d\mu^{\varphi}_R(x,\xi) = \int_{\mathbb{R}^d\times\mathbb{R}^d}\sum_{k\in\mathbb{Z}^d}a(x,\xi+2\pi n)|\widehat{\varphi}(\xi)|^2|\chi(\xi/R)|^2d\mu(x,\xi)$$

for every test function $a\in\mathcal{S}$. Passing to limits as $R\to\infty$ in the above identity, we obtain the claimed result.

Notice that the same argument may be applied if, instead of condition (D), we have that $(\langle h_kD_x\rangle^{-s}u_k)$ is h_k-oscillatory. This is because (4.12) may be estimated from above by

$$\limsup_{k\to\infty}\left\|\langle h_kD_x\rangle^{-s}\left(1-\chi\left(\frac{h_kD_x}{R}\right)\right)u_k\right\|^2_{L^2(\mathbb{R}^d)} \to 0 \quad \text{as } R\to\infty$$

because of Lemma 3.1 and the h_k-oscillation hypothesis. □

5. Computation of defect measures.

5.1. Relations between defect and Wigner measures in the discrete setting. In this paragraph, we establish the analogue of Proposition 1.2 in the discrete setting. In particular, we present conditions that ensure that the projection on the x-component of a Wigner measure may be obtained as the limit of quadratic densities of the type

$$E^h[U^h](x) := h^d \sum_{n\in\mathbb{Z}^d} |U_n^h|^2 \delta_{hn}(x).$$

PROPOSITION 5.1. *Let (h_k) be a scale and (U^{h_k}) an h_k-bounded sequence. Suppose that $(M^{h_k}[U^{h_k}])$ converges to μ as $k\to\infty$. Then, for every $\phi\in C_c(\mathbb{R}^d)$,*

$$\int_{\mathbb{R}^d\times Q} \phi(x)d\mu(x,\xi) = \lim_{k\to\infty} (h_k)^d \sum_{n\in\mathbb{Z}^d} \phi(h_k n)|U_n^{h_k}|^2. \tag{5.1}$$

If (ε_k) is a scale such that $h_k \ll \varepsilon_k$ and the transforms $M^{\varepsilon_k}[U^{h_k}]$ converge to μ, then (5.1) *holds, provided (U^{h_k}) is* ε_k-oscillatory, *i.e.,*

$$\limsup_{k\to\infty} (h_k)^d \int_{Q\setminus B(0;h_k/\varepsilon_k R)} |\widehat{U^{h_k}}(\xi)|^2 d\xi \to 0 \quad \text{as } R\to\infty. \tag{5.2}$$

In view of Proposition 5.1, one might think that Wigner measures at scales coarser than h_k are unnecessary. However, as the next result shows, if (U^{h_k}) is ε_k-oscillatory for such a scale, then the Wigner measure at scale (h_k) does not give any information about the oscillation effects.

PROPOSITION 5.2. *Let (h_k) and (ε_k) be scales such that $h_k \ll \varepsilon_k$. For every ε_k-oscillatory, h_k-bounded sequence (U^{h_k}) such that $M^{h_k}[U^{h_k}] \rightharpoonup \mu$ as $k\to\infty$, we have*

$$\mu(x,\xi) = \nu(x) \otimes \sum_{k\in\mathbb{Z}^d} \delta_{2\pi k}(\xi),$$

where ν is the weak limit in $\mathcal{M}_+(\mathbb{R}^d)$ of the measures $E^{h_k}[U^{h_k}]$.

The Wigner measure also gathers the information on the densities $|\mathcal{F}^{\varepsilon_k}U^{h_k}(\xi)|^2$; indeed, these converge to the projection on ξ of the Wigner measure, provided that no energy is lost at infinity.

PROPOSITION 5.3. *Let (h_k) and (ε_k) be scales such that h_k/ε_k is bounded. Suppose that (U^{h_k}) is* compact at infinity*,*

$$\limsup_{k\to\infty} (h_k)^d \sum_{|h_k n|>R} |U_n^{h_k}|^2 \to 0 \quad \text{as } R\to\infty, \tag{5.3}$$

and that $M^{\varepsilon_k}[U^{h_k}] \rightharpoonup \mu$ as $k\to\infty$. Then

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} \psi(\xi)d\mu(x,\xi) = \lim_{k\to\infty} \int_{\mathbb{R}^d} \psi(\xi)|\mathcal{F}^{\varepsilon_k}U^{h_k}(\xi)|^2 d\xi$$

for every $\psi\in C_c(\mathbb{R}^d)$.

The proof of Propositions 5.1 and 5.3 requires the following preliminary result, which explains how the transform $M^{\varepsilon}[U^h]$ of a discrete function U^h can be localized.

LEMMA 5.4. *Let $U^h \in L^2(h\mathbb{Z}^d)$ and let φ, $\phi \in C_c^\infty(\mathbb{R}^d)$. Then for every $a \in \mathcal{S}(\mathbb{R}^d \times \mathbb{R}^d)$ the following holds:*

$$\lim_{k\to\infty} \left| \langle M^{\varepsilon_k}[U^{h_k}], |\phi(x)|^2\varphi(\xi)\rangle_{\mathcal{S}'\times\mathcal{S}} - (h_k)^d \int_{\mathbb{R}^d} |\widehat{\phi U^{h_k}}(\xi)|^2 \varphi\left(\frac{\varepsilon_k}{h_k}\xi\right) \frac{d\xi}{(2\pi)^d} \right| = 0.$$

Proof. First note that, as a consequence of relation (1.13) and Lemma 8.5, we have

$$\lim_{k\to\infty} |\langle M^{\varepsilon_k}[U^{h_k}], |\phi(x)|^2\varphi(\xi)\rangle_{\mathcal{S}'\times\mathcal{S}} - \langle M^{\varepsilon_k}[\phi U^{h_k}], \psi(x)\varphi(\xi)\rangle_{\mathcal{S}'\times\mathcal{S}}| = 0 \tag{5.4}$$

for every test function $\psi \in C_c^\infty(\mathbb{R}^d)$ such that $\psi(x) = 1$ for $x \in \operatorname{supp}\phi$. Now, (8.2.i) and (1.13), together with Plancherel's formula for the discrete Fourier transform, yield

$$\begin{aligned}\langle M^{\varepsilon_k}[\phi U^{h_k}], \psi(x)\varphi(\xi)\rangle_{\mathcal{S}'\times\mathcal{S}} &= \langle \phi(x) T_{\delta_0}^{h_k} U^{h_k}, \varphi(\varepsilon_k D_x)\phi(x) T_{\delta_0}^{h_k} U^{h_k}\rangle_{\mathcal{S}'\times\mathcal{S}} \\ &= (h_k)^{2d} \int_{\mathbb{R}^d} |\widehat{\phi U^{h_k}}(h_k\xi)|^2 \varphi(\varepsilon_k\xi) \frac{d\xi}{(2\pi)^d},\end{aligned}$$

and the result follows. □

Proof of Proposition 5.1. Identity (5.1) in the case $h_k = \varepsilon_k$ is a direct consequence of the identity

$$\int_Q M^h[U](x,\xi)d\xi = E^h[U^h](x)$$

and the fact that, due to the Γ-periodicity in ξ of $M^{h_k}[U^{h_k}]$ and μ, one has

$$\lim_{k\to\infty} \int_{\mathbb{R}^d\times Q} \phi(x) M^{h_k}[U^{h_k}](x,\xi)dxd\xi = \int_{\mathbb{R}^d\times Q} \phi(x)d\mu(x,\xi)$$

for every $\phi \in C_c^\infty(\mathbb{R}^d)$.

Next we analyze the case $h_k/\varepsilon_k \to 0$. Given functions $\phi, \chi \in C_c^\infty(\mathbb{R}^d)$, and using Lemma 5.4 and periodization in the variable ξ, we find

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} |\phi(x)|^2\chi(\xi)d\mu(x,\xi) = \lim_{k\to\infty} (h_k)^d \int_Q |\widehat{\phi U^{h_k}}(\xi)|^2 \sum_{n\in\mathbb{Z}^d} \chi\left(\frac{\varepsilon_k}{h_k}(\xi + 2\pi n)\right) \frac{d\xi}{(2\pi)^d}. \tag{5.5}$$

Choose a function $\chi \in C_c^\infty(\mathbb{R}^d)$ such that

$$\begin{aligned}\chi(\xi) &= 1 \quad \text{for } |\xi| \le 1, \\ \chi(\xi) &= 0 \quad \text{for } |\xi| \ge 2, \\ 0 \le \chi(\xi) &\le 1 \quad \text{for } \xi \in \mathbb{R}^d,\end{aligned}$$

and set $\chi_R(\xi) := \chi(\xi/R)$ for every $R > 0$. With such a test function and $h_k/\varepsilon_k < \pi/R$ we have $\chi_R(\frac{\varepsilon_k}{h_k}(\xi + 2\pi n)) = \chi_R(\frac{\varepsilon_k}{h_k}\xi) \le 1$ for every $\xi \in Q$. Then, taking this into account in (5.5) and using Plancherel's formula, the following is obtained:

$$\lim_{k\to\infty} \left| (h_k)^d \int_Q |\widehat{\phi U^{h_k}}(\xi)|^2 \frac{d\xi}{(2\pi)^d} - \int_{\mathbb{R}^d\times\mathbb{R}^d} |\phi(x)|^2\chi_R(\xi)d\mu(x,\xi) \right| \le M(R),$$

where

$$M(R) := \limsup_{k\to\infty} (h_k)^d \int_{\mathbb{R}^d} \left(1 - \chi_R\left(\frac{\varepsilon_k}{h_k}(\xi + 2\pi n)\right)\right) |\widehat{\phi U^{h_k}}(\xi)|^2 \frac{d\xi}{(2\pi)^d}.$$

Identity (5.1) is obtained by letting R tend to ∞, noticing that (5.2) implies that $M(R) \to 0$ as $R \to \infty$. □

Proof of Proposition 5.2. Since $(U^{h_k})_{k\in\mathbb{N}}$ is ε_k-oscillatory, we have

$$\limsup_{k\to\infty} \int_{Q\setminus B(0;\delta)} |\widehat{\phi U^{h_k}}(\xi)|^2 d\xi = 0$$

for every $\delta > 0$ and $\phi \in C_c^\infty(\mathbb{R}^d)$. Using Lemma 5.4 below, we obtain, for every $\varphi \in C_c^\infty(Q \setminus \{0\})$,

$$0 = \lim_{k\to\infty} \int_Q \varphi(\xi) |\widehat{\phi U^{h_k}}(\xi)|^2 \frac{d\xi}{(2\pi)^d} = \int_{\mathbb{R}^d\times Q} |\phi(x)|^2 \varphi(\xi) d\mu(x,\xi).$$

In particular, μ is concentrated on the set $\mathbb{R}^d \times \{0\}$. Since $\mu(\cdot \times Q) = \nu(x)$ by Proposition 5.1, we find that, because of the periodicity, $\mu(\mathbb{R}^d \times \cdot) = \nu(\mathbb{R}^d) \sum_{k\in\mathbb{Z}^d} \delta_{2\pi k}(\xi)$, and this restricts μ to being equal to $\nu \otimes \sum_{k\in\mathbb{Z}^d} \delta_{2\pi k}$. □

Proof of Proposition 5.3. Since the densities $|\mathcal{F}^{\varepsilon_k} U^{h_k}|^2$ are uniformly bounded in $L^1(\mathbb{R}^d)$ (and consequently in $\mathcal{M}(\mathbb{R}^d)$) it suffices to prove the result for test functions $\psi \in C_c^\infty(\mathbb{R}^d)$. Let χ and χ_R be defined as in the proof of Proposition 5.1. Because of Lemma 5.4 the following holds for every $\psi \in \mathcal{S}(\mathbb{R}^d)$:

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} |\chi_R(x)|^2 \psi(\xi) d\mu(x,\xi) = \lim_{k\to\infty} \int_{\mathbb{R}^d} \psi(\xi) |\mathcal{F}^{\varepsilon_k} \chi_R U^{h_k}(\xi)|^2 d\xi.$$

Since $|\chi_R(x)|^2 \to 1$ as $R \to \infty$ for every $x \in \mathbb{R}^d$, we have to show only that

$$\lim_{R\to\infty} \int_{\mathbb{R}^d\times\mathbb{R}^d} |\chi_R(x)|^2 \psi(\xi) d\mu(x,\xi) = \lim_{k\to\infty} \int_{\mathbb{R}^d} \psi(\xi) |\mathcal{F}^{\varepsilon_k} U^{h_k}(\xi)|^2 d\xi.$$

This appears as a consequence of the identity

$$\begin{aligned}
&\int_{\mathbb{R}^d} \psi(\xi)(|\mathcal{F}^{\varepsilon_k} U^{h_k}(\xi)|^2 - |\mathcal{F}^{\varepsilon_k} \chi_R U^{h_k}(\xi)|^2) d\xi \\
&\quad = \int_{\mathbb{R}^d} \psi(\xi)[\mathcal{F}^{\varepsilon_k}(U^{h_k} - \chi_R U^{h_k})(\xi)]\overline{\mathcal{F}^{\varepsilon_k} U^{h_k}(\xi)} d\xi \\
&\qquad + \int_{\mathbb{R}^d} \psi(\xi) \mathcal{F}^{\varepsilon_k} \chi_R U^{h_k}(\xi) \overline{[\mathcal{F}^{\varepsilon_k}(U^{h_k} - \chi_R U^{h_k})(\xi)]} d\xi,
\end{aligned}$$

which implies

$$\begin{aligned}
&\limsup_{k\to\infty} \left| \int_{\mathbb{R}^d} \psi(\xi)(|\mathcal{F}^{\varepsilon_k} U^{h_k}(\xi)|^2 - |\mathcal{F}^{\varepsilon_k} \chi_R U^{h_k}(\xi)|^2) d\xi \right| \\
&\qquad \le C_\psi \limsup_{k\to\infty} \|U^{h_k} - \chi_R U^{h_k}\|^2_{L^2(h\mathbb{Z}^d)}.
\end{aligned}$$

Since the U^{h_k} are compact at infinity, the second term in the above estimate tends to zero as R tends to infinity, and thus

$$\left| \limsup_{k\to\infty} \int_{\mathbb{R}^d} \psi(\xi) |\mathcal{F}^{\varepsilon_k} U^{h_k}(\xi)|^2 - \int_{\mathbb{R}^d\times\mathbb{R}^d} |\chi_R(x)|^2 \psi(\xi) d\mu(x,\xi) \right| \to 0 \quad \text{as } R \to \infty.$$

One easily deduces from this that the measures $|\mathcal{F}^{\varepsilon_k} U^{h_k}(\xi)|^2 d\xi$ converge in $\mathcal{M}_+(\mathbb{R}^d)$ to the measure $\int_{\mathbb{R}^d} \mu(dx, \cdot)$, as claimed. □

5.2. Defect measures of reconstructed sequences. Let (U^{h_k}) be h_k-bounded and $\varphi \in H^s(\mathbb{R}^d)$ some profile satisfying (BP_s). As a consequence of Lemma 3.1, the sequence of densities

$$|\langle h_k D_x\rangle^s T_\varphi^{h_k} U^{h_k}|^2$$

is uniformly bounded in $L^1(\mathbb{R}^d)$. Hence, Helly's compactness theorem ensures that, extracting a subsequence if necessary, there exists a measure $\nu_\varphi \in \mathcal{M}_+(\mathbb{R}^d)$ such that

$$\lim_{k\to\infty} \int_{\mathbb{R}^d} \phi(x) |\langle h_k D_x\rangle^s T_\varphi^{h_k} U^{h_k}(x)|^2 dx = \int_{\mathbb{R}^d\times Q} \phi(x) d\nu_\varphi(x).$$

The main issue addressed in this section is that of clarifying how ν_φ depends on the sequence (U^{h_k}) and the profile φ. We shall see that a formula relating ν_φ and the limit of $E^{h_k}[U^{h_k}]$ does not exist in general. However, such a formula may be established in terms of the Wigner measure of (U^{h_k}).

Suppose that $M^{h_k}[U^{h_k}]$ converges to μ. Then Theorem 4.6 may be applied to obtain that, provided $\mu(\mathbb{R}^d \times \overline{D_{\widehat{\varphi}}}) = 0$, one has

$$m^{h_k}[T_\varphi^{h_k} U^{h_k}] \rightharpoonup |\widehat{\varphi}(\xi)|^2 \mu(x,\xi).$$

In general, we are only able to ensure (see Proposition 1.7 in [9])

$$\nu_\varphi(x) \geq \int_{\mathbb{R}^d} |\langle \xi\rangle^s \widehat{\varphi}(\xi)|^2 \mu(x, d\xi),$$

and equality holds whenever $(\langle h_k D_x\rangle^s T_\varphi^{h_k} U^{h_k})$ is h_k-oscillatory. Note, however, that this is not always the case. At the end of this section we provide an example of profile φ and a sequence (U^{h_k}) for which $(\langle h_k D_x\rangle^s T_\varphi^{h_k} U^{h_k})$ fails to be h_k-oscillatory. Nevertheless, the following simple sufficient condition for h_k-oscillation holds.

PROPOSITION 5.5. *$(\langle h_k D_x\rangle^s T_\varphi^{h_k} U^{h_k})$ is h_k-oscillatory whenever* (D) *holds.*

This immediately follows from the following lemma.

LEMMA 5.6. *$(\langle h_k D_x\rangle^s T_\varphi^{h_k} U^{h_k})$ is h_k-oscillatory if and only if*

$$\limsup_{k\to\infty} (h_k)^d \int_Q \sigma_\varphi^R(\xi) |\widehat{U^{h_k}}(\xi)|^2 d\xi \to 0 \quad \text{as } R \to \infty,$$

where

$$\sigma_\varphi^R(\xi) := \sum_{|n|\geq R} |\langle \xi + 2\pi n\rangle^s \widehat{\varphi}(\xi + 2\pi n)|^2.$$

Proof. Start by noticing that

$$\int_{|\xi|\geq R/h_k} |\widehat{\langle h_k D_x\rangle^s T_\varphi^{h_k}} U^{h_k}(\xi)|^2 d\xi = \int_{|\xi|\geq R} (h_k)^d |\langle \xi\rangle^s \widehat{\varphi}(\xi) \widehat{U^{h_k}}(\xi)|^2 d\xi.$$

Periodizing in ξ, we get

$$\begin{aligned}\int_Q \sigma_\varphi^{R+\sqrt{d}\pi}(\xi) |\widehat{U^{h_k}}(\xi)|^2 d\xi &\leq \int_{|\xi|\geq R} |\langle \xi\rangle^s \widehat{\varphi}(\xi) \widehat{U^{h_k}}(\xi)|^2 d\xi \\ &\leq \int_Q \sigma_\varphi^{R-\sqrt{d}\pi}(\xi) |\widehat{U^{h_k}}(\xi)|^2 d\xi,\end{aligned}$$

and the claim follows. □

Hence, h_k-oscillation is obtained if φ decays at infinity at a uniform rate. For more general φ, it is still possible to obtain sufficient conditions; however, these depend on the particular sequence of discrete functions to be reconstructed.

PROPOSITION 5.7. *Suppose*

$$\text{(5.6)}\qquad \begin{aligned} &\text{(i)}\quad \mu(\mathbb{R}^d \times \overline{D_{\tau_{\langle D_x\rangle^s\varphi}}}) = 0;\\ &\text{(ii)}\quad (U^{h_k}) \text{ is compact at infinity.}\end{aligned}$$

Then $(\langle h_k D_x\rangle^s T_\varphi^{h_k} U^{h_k})$ *is* h_k*-oscillatory.*

Proof. For the sake of simplicity, we prove the result for $s = 0$, the proof in the general case being identical. Taking into account the periodicity of the densities involved, Proposition 5.3 ensures that

$$\text{(5.7)}\qquad \lim_{k\to\infty} (h_k)^d \int_Q \psi(\xi)|\widehat{U^{h_k}}(\xi)|^2 d\xi = \int_{\mathbb{R}^d\times Q} \psi(\xi) d\mu(x,\xi)$$

for every $\psi \in C_c(\mathbb{R}^d)$, and the claim follows. Since $\mu(\mathbb{R}^d \times \overline{D_{\tau_\varphi}}) = 0$, necessarily $\mu(\mathbb{R}^d \times \overline{D_{\sigma_\varphi^R}})$ is null for every $R > 0$. From classical results on weak convergence of measures, one deduces that relation (5.7) also holds for $\psi = \sigma_\varphi^R$. Hence, by the dominated convergence theorem,

$$\lim_{R\to\infty}\lim_{k\to\infty} (h_k)^d \int_Q \sigma_\varphi^R(\xi)|\widehat{U^{h_k}}(\xi)|^2 d\xi = \lim_{R\to\infty}\int_{\mathbb{R}^d\times Q} \sigma_\varphi^R(\xi) d\mu(x,\xi) = 0,$$

and the result follows. □

Combining Propositions 5.5, 5.7, and 1.2, we obtain the following result.

PROPOSITION 5.8. *Suppose that at least one of either* (D) *or* (5.6) *is satisfied and that* $\mu(\mathbb{R}^d \times \overline{D_{\widehat{\varphi}}}) = 0$. *Then*

$$\text{(5.8)}\qquad \lim_{k\to\infty} \int_{\mathbb{R}^d} \phi(x)|\langle h_k D_x\rangle^s T_\varphi^{h_k} U^{h_k}|^2 dx = \int_{\mathbb{R}^d\times\mathbb{R}^d} \phi(x)|\langle\xi\rangle^s \widehat{\varphi}(\xi)|^2 d\mu(x,\xi)$$

for every $\phi \in C_c(\mathbb{R}^d)$.

As anticipated above, (5.8) shows that the knowledge of the weak limit of the measures $E^{h_k}[U^{h_k}]$ and the profile φ are not enough, in general, to reconstruct the weak limit of the densities $|\langle h_k D_x\rangle^s T_\varphi^{h_k} U^{h_k}|^2 dx$. However, when the sequence of discrete functions under consideration is ε_k-oscillatory for some scale coarser than the reconstruction step h_k, there does exist a formula that relates both limits, as follows.

COROLLARY 5.9. *Let* (U^{h_k}) *be an* h_k*-bounded,* ε_k*-oscillatory sequence such that* $(E^{h_k}[U^{h_k}])$ *weakly converges to a measure* ν. *Suppose, moreover, that* $\widehat{\varphi}$ *is continuous at* Γ *and that any of* (D) *or* (5.6) *is satisfied. Then the densities* $|\langle h_k D_x\rangle^s T_\varphi^{h_k} U^{h_k}|^2$ *weakly converge to the measure*

$$\nu_\varphi(x) = \left(\sum_{n\in\mathbb{Z}^d} |\langle 2\pi n\rangle^s \widehat{\varphi}(2\pi n)|^2\right)\nu(x).$$

Proof. Using Proposition 5.2, we find that any Wigner measure at scale h_k of (U^{h_k}) equals

$$\mu(x,\xi) = \nu(x) \otimes \sum_{n\in\mathbb{Z}^d} \delta_{2\pi n}(\xi).$$

Since $\mu(\mathbb{R}^d \times \overline{D_{\widehat{\varphi}}}) = 0$, Proposition 5.8 is applicable and gives

$$|\langle h_k D_x\rangle^s T_\varphi^{h_k} U^{h_k}|^2 dx \rightharpoonup \tau_{\langle D_x\rangle^s \varphi}(0)\nu(x) \quad \text{as } k \to \infty,$$

as claimed. □

Note that condition (5.6.i) reduces in this setting to the requirement that $\tau_{\langle D_x\rangle^s\varphi}$ is continuous at $\xi = 0$.

5.3. A counterexample to *h*-oscillation. Here we construct a function $\varphi \in L^2(\mathbb{R})$ satisfying (BP_s) such that $\widehat{\varphi}$ is continuous but

$$\|\sigma_\varphi^R\|_{L^\infty(Q)} = 1 \quad \text{for every } R > 0.$$

With such a profile, we show that theres exist a sequence of discrete functions (U^h) such that $(T_\varphi^h U^h)$ is not h-oscillatory.

To construct φ, define $t_n := e^{-n}$ for $n = 0, 1, 2\ldots$ and let ψ_n be the piecewise linear function given for $n \geq 1$ by

$$\psi_n(t) := \begin{cases} \dfrac{t - t_{n+1}}{t_n - t_{n+1}} & \text{if } t \in (t_{n+1}, t_n), \\ \dfrac{t - t_{n-1}}{t_n - t_{n-1}} & \text{if } t \in (t_n, t_{n-1}), \\ 0 & \text{otherwise.} \end{cases}$$

Clearly $\sum_{n=1}^\infty \psi_n(t) = 1$ for $t \in (0, t_1)$, and the sum vanishes for $t \leq 0$. Defining

$$\widehat{\varphi}(\xi) := \sqrt{\sum_{n=1}^\infty \psi_n(\xi - 2\pi n)},$$

we obtain $\varphi \in L^2(\mathbb{R})$, $\widehat{\varphi} \in C(\mathbb{R})$, and $\tau_\varphi(\xi) = \sum_{n=1}^\infty \psi_n(\xi)$ for every $\xi \in Q$.

Moreover

$$\sigma_\varphi^n(\xi) = \begin{cases} 1 & \text{if } \xi \in (0, t_{n+1}), \\ 0 & \text{if } \xi \leq 0. \end{cases}$$

Thus $\|\sigma_\varphi^R\|_{L^\infty(Q)} = 1$ for every $R > 0$.

If we choose discrete functions $U^h \in L^2(h\mathbb{Z})$ such that

$$\widehat{U^h}(\xi) = h^{-1} \sum_{n\in\mathbb{Z}} \mathbf{1}_{(0,h)}(\xi + 2\pi n),$$

then for the φ constructed above we obtain

$$\lim_{h\to 0} \int_Q \sigma_\varphi^R(\xi) h |\widehat{U^h}(\xi)|^2 d\xi = 1 \quad \text{for every } R > 0.$$

This proves that $(T_\varphi^h U^h)$ is not h-oscillatory.

6. High-frequency analysis: $h \ll \varepsilon$. Here we shall investigate the structure of Wigner measures at scales (ε_k) asymptotically coarser than the sampling/reconstruction rate (h_k).

In the next two theorems, we suppose that φ satisfies (BP_s) and $\widehat{\varphi}$ *is continuous in a neighborhood of* $\xi = 0$. Moreover, (h_k) and (ε_k) will be scales such that $h_k \ll \varepsilon_k$.

THEOREM 6.1. *Suppose that* (U^{h_k}) *is* h_k*-bounded and* $M^{\varepsilon_k}[U^{h_k}]$ *converges to the Wigner measure* μ*. Then* $m^{\varepsilon_k}[T_\varphi^{h_k} U^{h_k}]$ *converges to a measure* μ_φ *given by*

$$\mu_\varphi(x,\xi) = |\widehat{\varphi}(0)|^2 \mu(x,\xi). \tag{6.1}$$

The proof of this result is completely analogous to that of Theorem 4.6.

Concerning the sampling operators, the situation is much similar.

THEOREM 6.2. *Let* (u_k) *be a sequence in* $H^{-s}(\mathbb{R}^d)$ *such that* $(\langle h_k D_x\rangle^{-s} u_k)$ *is bounded in* $L^2(\mathbb{R}^d)$ *and* ε_k*-oscillatory.*

(i) *Then* $(S_\varphi^{h_k} u_k)$ *is* ε_k*-oscillatory.*

(ii) *Suppose moreover that* $m^{\varepsilon_k}[u_k]$ *converges to a Wigner measure* μ*. Then* $M^{\varepsilon_k}[S_\varphi^{h_k} u_k]$ *converges to the Wigner measure* μ^φ *given by*

$$\mu^\varphi = |\widehat{\varphi}(0)|^2 \mu. \tag{6.2}$$

Proof. To prove the first part of the theorem, begin by noticing that, by the Cauchy–Schwarz inequality and Lemma 4.11, for almost every $\xi \in \mathbb{R}^d$,

$$\begin{aligned}
|\widehat{S_\varphi^{h_k} u_k}(\xi)|^2 &= \left| \frac{1}{(h_k)^d} \sum_{n\in\mathbb{Z}^d} \overline{\widehat{\varphi}(\xi + 2\pi n)} \widehat{u_k}\left(\frac{\xi + 2\pi n}{h_k}\right)\right|^2 \\
&\le \frac{\|\tau_{\langle D_x\rangle^s \varphi}\|_{L^\infty(Q)}}{(h_k)^{2d}} \sum_{n\in\mathbb{Z}^d} \left| \langle \xi + 2\pi n\rangle^{-s} \widehat{u_k}\left(\frac{\xi + 2\pi n}{h_k}\right)\right|^2.
\end{aligned}$$

Thus

$$\begin{aligned}
&\int_{Q\setminus B(0;h_k/\varepsilon_k R)} (h_k)^d |\widehat{S_\varphi^{h_k} u_k}(\xi)|^2 d\xi \\
&\quad \le \|\tau_{\langle D_x\rangle^s \varphi}\|_{L^\infty(Q)} \int_{Q\setminus B(0;R/\varepsilon_k)} \sum_{n\in\mathbb{Z}^d} |\langle h_k\xi + 2\pi n\rangle^{-s}\, \widehat{u_k}(\xi + 2\pi n)|^2 d\xi \\
&\quad \le \|\tau_{\langle D_x\rangle^s \varphi}\|_{L^\infty(Q)} \int_{\mathbb{R}^d\setminus B(0;R/\varepsilon_k)} |\langle h_k\xi\rangle^{-s}\, \widehat{u_k}(\xi)|^2 d\xi,
\end{aligned}$$

and this clearly proves that $(S_\varphi^{h_k} u_k)$ is ε_k-oscillating as soon as $(\langle h_k D_x\rangle^{-s} u_k)$ is.

The proof of identity (6.2) is essentially identical to that of Theorem 4.2. A completely analogous argument to that used in Step 2 of that proof allows us to consider only sequences such that $\widehat{\varphi}(h_k/\varepsilon_k\cdot)\widehat{u_k}(\cdot/\varepsilon_k)$ is supported in a ball $B(0;R)$. This hypothesis, together with Lemma 4.11, implies that, for h_k/ε_k small enough,

$$(h_k)^d \widehat{S_\varphi^{h_k} u_k}\left(\frac{h_k}{\varepsilon_k}\xi\right) = \overline{\widehat{\varphi}\left(\frac{h_k}{\varepsilon_k}\xi\right)} \widehat{u_k}(\xi/\varepsilon_k);$$

that is, only one summand is involved. Then the result follows from Proposition 8.3 exactly as in the proof of Theorem 4.2. ∎

We conclude with a simple remark.

COROLLARY 6.3. *Under the assumptions and notation of Theorems* 6.1 *and* 6.2*, the following hold:*

(i) *If φ has zero mean (i.e., $\widehat{\varphi}(0) = 0$), then the Wigner measure at scale (ε_k) of any sequence $(T_\varphi^{h_k} U^{h_k})$ or $(S_\varphi^{h_k} u_k)$ vanishes identically. In particular, this is the case if φ is a wavelet.*[9]

(ii) *$\mu_\varphi = \mu^\varphi = \mu$ always holds for profiles such that $|\widehat{\varphi}(0)| = 1$.*

7. Wigner measures of sampled/reconstructed sequences. Now we are able to describe Wigner measures of sequences of the form $T_\psi^h S_\varphi^h u$. In its full generality, our result requires several compatibility hypotheses, which we describe below. First of all,

$$\text{(7.1)}\qquad \begin{aligned} &\text{(i)}\quad \psi \text{ and } \varphi \text{ satisfy } (\mathrm{BP}_s) \text{ with exponents } s' \text{ and } s \text{ respectively;}\\ &\text{(ii)}\quad \varphi \text{ satisfies (D).}\end{aligned}$$

The admissible sequences will be assumed to be such that

$$\text{(7.2)}\qquad u_k \in H^{-s}(\mathbb{R}^d), \quad \text{and} \quad (\langle h_k D_x\rangle^{-s} u_k) \text{ is bounded in } L^2(\mathbb{R}^d),$$

and their Wigner measures must satisfy the following compatibility conditions for some precise representatives of $\widehat{\psi}$ and $\widehat{\varphi}$:

$$\text{(7.3)}\qquad \begin{aligned} &\text{(i)}\quad \mu \text{ fulfills (ND).}\\ &\text{(ii)}\quad \int_{\mathbb{R}^d\times\mathbb{R}^d} \mathbf{1}_{\overline{D_{\widehat{\psi}}}}(\xi + 2\pi n)|\widehat{\varphi}(\xi)|^2 d\mu(x,\xi) = 0, \quad n \in \mathbb{Z}^d.\\ &\text{(iii)}\quad \mu \text{ satisfies (MS).}\end{aligned}$$

Combining Theorems 4.6 and 4.2 we obtain the following theorem.

THEOREM 7.1. *Let ψ and φ be functions satisfying* (7.1)*; let (h_k) be a scale and let (u_k) be a sequence satisfying* (7.2)*. Suppose, moreover, that $m^{h_k}[u_k]$ converges to a Wigner measure μ that satisfies* (7.3)*.*

Then $m^{h_k}[T_\psi^{h_k} S_\varphi^{h_k} u_k]$ converges to the measure $\mu_{\varphi,\psi}$ given by

$$\text{(7.4)}\qquad \int_{\mathbb{R}^d\times\mathbb{R}^d} a(x,\xi)\, d\mu_{\varphi,\psi}(x,\xi) = \int_{\mathbb{R}^d\times\mathbb{R}^d} \sum_{n\in\mathbb{Z}^d} a(x,\xi+2\pi n)|\widehat{\psi}(\xi+2\pi n)|^2 |\widehat{\varphi}(\xi)|^2 d\mu(x,\xi)$$

for every $a \in C_c(\mathbb{R}^d \times \mathbb{R}^d)$.

Proof. Hypothesis (7.3.ii) expresses that the closure of the set of discontinuity points of $\widehat{\psi}$ is a null set for the Wigner measure of $S_\varphi^{h_k} u_k$,

$$\sum_{k\in\mathbb{Z}^d} |\widehat{\varphi}(\xi+2\pi n)|^2 \mu(x,\xi+2\pi n).$$

Hence, Theorem 4.6 is applicable, and we conclude that the distributions $m^{h_k}[T_\psi^{h_k} S_\varphi^{h_k} u_k]$ converge to the measure

$$|\widehat{\psi}(\xi)|^2 \sum_{n\in\mathbb{Z}^d} |\widehat{\varphi}(\xi+2\pi n)|^2 \mu(x,\xi+2\pi n).$$

Since $|\widehat{\psi}(\xi)|^2$ is integrable with respect to the finite measure $|\widehat{\varphi}|^2\mu$ (this is again due to (7.3.ii)), its periodization is integrable as well, and formula (7.4) follows. ∎

[9] See, for instance, [10, Proposition 2.1].

Remark 7.2. (i) When $\widehat{\psi}$ and $\widehat{\varphi}$ verify (3.6), then hypotheses (7.1), (7.3.i), and (7.3.ii) are immediately satisfied.

(ii) (7.1.ii) may be replaced by the requirement that $(\langle h_k D_x\rangle^{-s} u_k)$ be h_k-oscillatory.

From formula (7.4) one sees at once that, taking $\psi = \varphi = \delta_0$, one has that $\mu_{\varphi,\psi}$ is the periodization in ξ of the Wigner measure μ. Hence, $\mu_{\varphi,\psi}$ coincides with the limit of the Wigner series corresponding to (u_k).

When $\widehat{\psi}$ and $|\widehat{\varphi}|^2\mu$ vanish off Q it is easy to check that formula (7.4) takes the simple form

$$\mu_{\varphi,\psi}(x,\xi) = |\widehat{\psi}(\xi)|^2 |\widehat{\varphi}(\xi)|^2 \mu(x,\xi).$$

It is also clear that as soon as $|\widehat{\varphi}(\xi)|^2\mu(x,\xi)$ is not null outside Q, the measures $\mu_{\varphi,\psi}$ and μ will in general differ.

Concerning defect measures, combining Proposition 5.8 and the previous theorem, we obtain the following.

THEOREM 7.3. *Under the notation of Theorem* 7.1 *the following holds: if*

$$|\langle h_k D_x\rangle^{s'} T_\psi^{h_k} S_\varphi^{h_k} u_k|^2 dx \text{ weakly converges to a measure } \nu_{\varphi,\psi}$$

and ψ *verifies* (D)*, then*

$$\nu_{\varphi,\psi}(x) = \int_{\mathbb{R}^d_\xi} \sum_{n\in\mathbb{Z}^d} |\langle \xi + 2\pi n\rangle^{s'} \widehat{\psi}(\xi + 2\pi n)|^2 |\widehat{\varphi}(\xi)|^2 \mu(x, d\xi). \tag{7.5}$$

Remark 7.4. (i) The conclusion of the theorem still holds if condition "ψ satisfies (D)" is replaced by (5.6).

(ii) Theorems 1.3 and 1.5 follow immediately from Theorems 7.1 and 7.3.

With formula (7.5) at our disposal, we are now able to answer, in a quite general way, questions A–D addressed in the introduction. Of course, the answer to A is negative, since, in general, μ is not trivial in its ξ-component; concerning the problem of filtering, we immediately get the necessary and sufficient condition

$$c_{\varphi,\psi}(\xi) := |\widehat{\varphi}(\xi)|^2 \sum_{n\in\mathbb{Z}^d} |\langle \xi + 2\pi n\rangle^{s'} \widehat{\psi}(\xi + 2\pi n)|^2 = 0 \quad \text{for } \mu\text{-a.e. } \xi \in \mathbb{R}^d.$$

Analogously, $c_{\varphi,\psi}(\xi) = 1$ for μ-almost every $\xi \in \mathbb{R}^d$ characterizes the profiles that give $\nu_{\varphi,\psi} = \nu$. To answer D, we must, of course, assume that $\widehat{\varphi}$ and $\tau_{\langle D_x\rangle^{s'}\widehat{\psi}}$ are continuous (which, as we know, is the case if (3.6) holds). In that case, we have the equality $\nu_{\varphi,\psi} = \nu$ for every admissible sequence if and only if

$$|\widehat{\varphi}(\xi)|^2 = \frac{1}{\tau_{\langle D_x\rangle^{s'}\psi}(\xi)} \quad \text{for every } \xi \in \mathbb{R}^d \quad \text{with } \tau_{\langle D_x\rangle^{s'}\psi}(\xi) \neq 0.$$

The sampling profile φ cannot be an $L^2(\mathbb{R}^d)$ function, since $|\widehat{\varphi}|^2$ is necessarily periodic. When $\varphi = \delta_0$ and ψ generates an orthonormal basis in the sense of Lemma 3.5 we always have $\nu_{\varphi,\psi} = \nu$. If ψ merely generates a Riesz basis, $A\nu \leq \nu_{\varphi,\psi} \leq B\nu$ holds instead.

The above results may be used to compute Wigner measures of the *orthogonal projections* $P_\psi^{h_k} u_k$ of a given sequence (u_k) on the shift-invariant space defined by the range of $T_\psi^{h_k}$. As we have seen in Lemma 3.6, P_ψ^h may be written as the composition

of T_ψ^h with $S_\varphi^h\langle hD_x\rangle^s$ for a sampling profile $\varphi := \widetilde{\langle D_x\rangle^s\psi}$. Hence, Theorem 7.1 gives the following.

COROLLARY 7.5. *For ψ satisfying* (3.6) *and (u_k) such that* (7.2) *and* (MS) *hold, the defect measures of the sequence $(P_\psi^{h_k}u_k)$ is given by*

$$\nu_{P_\psi}(x) = \int_{\mathbb{R}^d} \frac{\mathbf{1}_\psi(\xi)}{\tau_{\langle D_x\rangle^s\psi}(\xi)} |\langle\xi\rangle^s\widehat{\psi}(\xi)|^2\mu(x,d\xi),$$

where $\mathbf{1}_\psi(\xi)$ denotes the characteristic function of the set of $\xi\in\mathbb{R}^d$ such that $\tau_{\langle D_x\rangle^s\psi}(\xi)\neq 0$.

In particular, when ψ gives rise to an orthonormal family, we obtain the simple formula (cf. Lemma 3.5)

$$\nu_{P_\psi}(x) = \int_{\mathbb{R}^d} |\langle\xi\rangle^s\widehat{\psi}(\xi)|^2\mu(x,d\xi).$$

To conclude, we shall see how the above results may be refined when the sequence $(\langle h_kD_x\rangle^{-s}u_k)$ is assumed to be ε_k-oscillatory at some scale $h_k\ll\varepsilon_k$. The assumptions of φ and ψ are weaker:

(7.6)
- (i) ψ and φ satisfy (BP_s) with exponents s' and s, respectively.
- (ii) $\widehat{\psi}$, $\widehat{\varphi}$ are continuous in a neighborhood of $\xi=0$.
- (iii) $\tau_{\langle D_x\rangle^{s'}\psi}$ is continuous at $\xi=0$.

Theorems 6.1 and 6.2 and Corollary 5.9 then give the following.

THEOREM 7.6. *Let ψ and φ be functions satisfying* (7.6)*, let (h_k), (ε_k) be scales with $h_k\ll\varepsilon_k$, and let (u_k) be a sequence such that* (7.2) *holds and $(\langle h_kD_x\rangle^{-s}u_k)$ is ε_k-oscillatory. Suppose, moreover, that $m^{\varepsilon_k}[u_k]$ converges to a Wigner measure μ.*

Then $m^{\varepsilon_k}[T_\psi^{h_k}S_\varphi^{h_k}u_k]$ converges to the measure $\mu_{\varphi,\psi}$ given by

$$\mu_{\varphi,\psi}(x,\xi) = |\widehat{\psi}(0)|^2|\widehat{\varphi}(0)|^2\mu(x,\xi).$$

Moreover, if $|\langle h_kD_x\rangle^{s'}T_\psi^{h_k}S_\varphi^{h_k}u_k|^2dx$ weakly converges to a measure $\nu_{\varphi,\psi}$, then

$$\nu_{\varphi,\psi}(x) = \sum_{n\in\mathbb{Z}^d} |\langle 2\pi n\rangle^{s'}\widehat{\psi}(2\pi n)|^2|\widehat{\varphi}(0)|^2\nu(x),$$

where ν is the weak limit of the densities $|\langle h_kD_x\rangle^{-s}u_k|^2dx$.

Hence, when a sequence possesses a characteristic oscillation scale (ε_k) (that is the meaning of the ε_k-oscillation condition), choosing a sampling/reconstruction rate (h_k) asymptotically finer than (ε_k) allows us to completely capture its oscillation/concentration behavior (modulo a constant that depends only on ψ and φ).

Filtering in that case can be achieved only by means of a sampling profile φ with zero mean ($\widehat{\varphi}(0)=0$) or a reconstruction profile such that $\widehat{\psi}$ vanishes at Γ.

8. Tools from the theory of Wigner measures. The main tools from the theory of Wigner measures used in this article are Propositions 8.1 and 8.3 below. The first of these is an extension of Theorem 1.1 to bounded sequences in Sobolev spaces.

PROPOSITION 8.1. *Let (ε_k) be a scale and let (u_k) be a sequence of functions in $H^{-s}(\mathbb{R}^d)$ for some $s\geq 0$ satisfying*

(8.1) $\qquad \|\langle\varepsilon_kD_x\rangle^{-s}u_k\|_{L^2(\mathbb{R}^d)}$ *are uniformly bounded in k.*

Then the sequence of distributions $(m^{\varepsilon_k}[u_k])$ is uniformly bounded in $\mathcal{S}'$. Moreover, any of its weakly converging subsequences tends to a positive measure.

As we have done so far, a measure $\mu \in \mathcal{M}_+(\mathbb{R}^d \times \mathbb{R}^d)$ will be called the *Wigner measure at scale* (ε_k) of a sequence (u_k) (satisfying the hypotheses of Proposition 8.1) provided $m^{\varepsilon_k}[u_k] \rightharpoonup \mu$ in $\mathcal{S}'$ as $k \to \infty$.

Remark 8.2. (i) When $s > 0$, condition (8.1) is stronger than just requiring that (u_k) is bounded in $H^{-s}(\mathbb{R}^d)$.

(ii) Let (h_k) be a scale such that $h_k \ll \varepsilon_k$. If $\|\langle h_k D_x\rangle^{-s} u_k\|_{L^2(\mathbb{R}^d)} \leq C$ for every $k \in \mathbb{N}$, then $\|\langle \varepsilon_k D_x\rangle^{-s} u_k\|_{L^2(\mathbb{R}^d)}$ is uniformly bounded as well.

(iii) The same result holds if $m^\varepsilon[\cdot]$ is replaced by the Wigner transform (1.4).

The second main result of this section is a localization formula for Wigner measures which was used several times in this article.

PROPOSITION 8.3. *Let (ε_k), (h_k) be scales and let (u_k) be a sequence in $H^{-s}(\mathbb{R}^d)$, $s \geq 0$, satisfying* (8.1). *Suppose that ϕ is a Borel function such that $\phi \in L^\infty(\mathbb{R}^d; \langle\xi\rangle^r)$, $r \in \mathbb{R}^d$. If $m^{\varepsilon_k}[u_k]$ converges to μ, then $m^{\varepsilon_k}[\phi(h_k D_x)u_k]$ converges to a Wigner measure μ_ϕ which has the following properties:*

(i) *If $h_k = \varepsilon_k$ and $\mu(\mathbb{R}^d \times \overline{D_\phi}) = 0$, D_ϕ being the set of points where ϕ is not continuous, then*

$$\mu_\phi(x,\xi) = |\phi(\xi)|^2 \mu(x,\xi).$$

(ii) *If $h_k \ll \varepsilon_k$ and ϕ is continuous in a neighborhood of $\xi = 0$, then*

$$\mu_\phi = |\phi(0)|^2 \mu.$$

When applied to $\phi(\xi) := \langle\xi\rangle^s$, this result gives the following.

Remark 8.4. Let (ε_k), (h_k), and (u_k) be as in Proposition 8.3. Suppose $m^{\varepsilon_k}[u_k]$ converges to μ. Then $m^{\varepsilon_k}[\langle h_k D_x\rangle^{-s} u_k]$ converges to the measure μ_s given by

$$\begin{array}{ll} \mu_s(x,\xi) = \langle\xi\rangle^{-2s}\mu(x,\xi) & \text{if } h_k = \varepsilon_k, \\ \mu_s = \mu & \text{if } h_k \ll \varepsilon_k. \end{array}$$

In particular (cf. Theorem 1.1), $\langle\xi\rangle^{-2s}\mu$ (resp., μ) is a finite measure when $h_k = \varepsilon_k$ (resp., $h_k \ll \varepsilon_k$).

For the convenience of the reader, we give detailed proofs of both results; they follow the ideas present in the existing literature on the subject (see [6, 13, 8, 9]). Proposition 8.1 will be proved in section 8.2. We shall essentially show that truncation of the high frequencies of a sequence satisfying (8.1) implies ξ-variable localization of the corresponding $m^\varepsilon[\cdot]$. Then we conclude by applying Theorem 1.1 to the localized sequence.

Proposition 8.3 is proved in section 8.3; in section 8.4, we describe two results useful for the computation of Wigner measures (Lemmas 8.12 and 8.13).

8.1. First properties of $m^\varepsilon[u]$. We begin by discussing three alternative ways of computing $m^\varepsilon[u]$ that may be used when u is merely a tempered distribution. First note that, given a $u \in \mathcal{S}'(\mathbb{R}^d)$, it makes sense to consider the distribution $m^\varepsilon[u]$ given by (1.1), since the Fourier transform of u is well-defined. Actually $m^\varepsilon[u] \in \mathcal{S}'$.

1. The action of $m^\varepsilon[u]$ on a test function $a \in \mathcal{S}$ is given by any of the following formulas (see [7]):

$$\langle m^\varepsilon[u], a\rangle_{\mathcal{S}'\times\mathcal{S}} = \begin{cases} \langle \overline{u}, a(x, \varepsilon D_x)u\rangle_{\mathcal{S}'(\mathbb{R}^d)\times\mathcal{S}(\mathbb{R}^d)}, & \text{(i)} \\ \displaystyle\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \frac{1}{\varepsilon^d} k_a\left(x, \frac{x-p}{\varepsilon}\right) u(p)\overline{u(x)}\,dp\,dx, & \text{(ii)} \end{cases} \tag{8.2}$$

where $a(x, \varepsilon D_x)$ is the *semiclassical pseudodifferential operator* of symbol a,

$$a(x, \varepsilon D_x)u(x) = \int_{\mathbb{R}^d} a(x, \varepsilon\xi)\widehat{u}(\xi)e^{ix\cdot\xi}\frac{d\xi}{(2\pi)^d}, \tag{8.3}$$

and the kernel $k_a(x, p)$ is the inverse Fourier transform of a with respect to ξ,

$$k_a(x, p) := \int_{\mathbb{R}^d} a(x, \xi)e^{ip\cdot\xi}\frac{d\xi}{(2\pi)^d}.$$

Formula (8.2.i) makes sense because the operator $a(x, \varepsilon D_x)$ continuously maps $\mathcal{S}'(\mathbb{R}^d)$ into $\mathcal{S}(\mathbb{R}^d)$ whenever $a \in \mathcal{S}$ (see, for instance, [15]). The integral in (8.2.ii) must, of course, be understood in distributional sense.

2. The distribution $m^\varepsilon[u]$ may be computed through the rescaled Fourier transform

$$\mathcal{F}^\varepsilon u(\xi) := \frac{1}{(2\pi\varepsilon)^{d/2}}\widehat{u}\left(\frac{\xi}{\varepsilon}\right), \tag{8.4}$$

using the identity

$$m^\varepsilon[u](x, \xi) = \overline{m^\varepsilon[\mathcal{F}^\varepsilon u](\xi, -x)}. \tag{8.5}$$

This follows from a direct computation from the definition (1.1).

3. Now we present two localization formulas.

LEMMA 8.5. *Let $u \in \mathcal{S}'(\mathbb{R}^d)$, $\phi \in C^\infty(\mathbb{R}^d; \langle x\rangle^r)$ for some $r \in \mathbb{R}$ and $a \in \mathcal{S}$. Then there exists $r_1^\sigma, r_2^\sigma \in \mathcal{S}$ such that*

$$\langle m^\varepsilon[\phi u], a\rangle_{\mathcal{S}'\times\mathcal{S}} = \langle|\phi(x)|^2 m^\varepsilon[u], a\rangle_{\mathcal{S}'\times\mathcal{S}} + \varepsilon\langle m^\varepsilon[u], r_1^\varepsilon\rangle_{\mathcal{S}'\times\mathcal{S}},$$
$$\langle m^\varepsilon[\phi(hD_x)u], a\rangle_{\mathcal{S}'\times\mathcal{S}} = \left\langle\left|\phi\left(\frac{h}{\varepsilon}\xi\right)\right|^2 m^\varepsilon[u], a\right\rangle_{\mathcal{S}'\times\mathcal{S}} + \varepsilon\langle m^\varepsilon[u], r_2^{h/\varepsilon}\rangle_{\mathcal{S}'\times\mathcal{S}}.$$

Moreover, the test functions r_1^σ, r_2^σ are uniformly bounded in $\mathcal{S}$ for $0 < \sigma \le 1$.

This holds as a consequence of standard results on symbolic calculus for semiclassical pseudodifferential operators; see, for instance, [15]. Note that Proposition 8.3 is not a consequence of this result, since the multiplier $\phi(hD_x)$ there may have a nonsmooth symbol.

8.2. Boundedness of the transforms $m^\varepsilon[u]$. The following lemmas are used to establish the boundedness in $\mathcal{S}'$ of the sequence $(m^{\varepsilon_k}[u_k])$, provided (u_k) satisfies the hypotheses of Proposition 8.1.

LEMMA 8.6. *For every $u \in L^2(\mathbb{R}^d; \langle x\rangle^r)$ and $a \in \mathcal{S}$ the following estimate holds:*

$$|\langle m^\varepsilon[u], a\rangle_{\mathcal{S}'\times\mathcal{S}}| \le \|u\|^2_{L^2(\mathbb{R}^d;\langle x\rangle^r)} \int_{\mathbb{R}^d} \sup_{x\in\mathbb{R}^d} |k_a(x, p)\langle x - \varepsilon p\rangle^{-r/2}\langle x\rangle^{-r/2}|dp.$$

Proof. Use formula (8.2.ii) to write

$$\langle m^\varepsilon[u], a\rangle_{\mathcal{S}'\times\mathcal{S}} = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} k_a(x, p)u(x - \varepsilon p)\overline{u(x)}dpdx,$$

noticing that this integral makes sense as $k_a \in \mathcal{S}$. Multiply and divide the integrand above by $\langle x - \varepsilon p\rangle^{r/2}\langle x\rangle^{r/2}$ to obtain, by Hölder's inequality,

$$|\langle m^\varepsilon[u], a\rangle_{\mathcal{S}'\times\mathcal{S}}| \le \int_{\mathbb{R}^d} \sup_{x\in\mathbb{R}^d} |k_a(x, p)\langle x - \varepsilon p\rangle^{-r/2}\langle x\rangle^{-r/2}| \int_{\mathbb{R}^d} |u_r(x - \varepsilon p)\overline{u_r(x)}|dxdp,$$

where we have set $u_r(x) := \langle x\rangle^{r/2}u(x)$. The conclusion follows from another application of Hölder's inequality. □

If $u \in H^{-s}(\mathbb{R}^d)$, then $\mathcal{F}^\varepsilon u \in L^2(\mathbb{R}^d; \langle\xi\rangle^{-2s})$. Clearly,

$$\|\langle \varepsilon D_x\rangle^{-s}u\|^2_{L^2(\mathbb{R}^d)} = \|\mathcal{F}^\varepsilon u\|^2_{L^2(\mathbb{R}^d;\langle\xi\rangle^{-2s})}. \tag{8.6}$$

Thus, taking identity (8.5) into account, we obtain, using the preceding lemma,

$$|\langle m^\varepsilon[u], a\rangle_{\mathcal{S}'\times\mathcal{S}}| \le \|\langle \varepsilon D_x\rangle^{-s}u\|^2_{L^2(\mathbb{R}^d)} \int_{\mathbb{R}^d} \sup_{\xi\in\mathbb{R}^d} |\widehat{a}(q,\xi)\langle\xi+\varepsilon q\rangle^s\langle\xi\rangle^s| \frac{dq}{(2\pi)^d}, \tag{8.7}$$

where $\widehat{a}(q,\xi)$ denotes the Fourier transform in x of the function $a(x,\xi)$.

LEMMA 8.7. *For every $s \ge 0$ there exists a constant $C_{s,d} > 0$ such that*

$$|\langle m^\varepsilon[u], a\rangle_{\mathcal{S}'\times\mathcal{S}}| \le C_{s,d}\|\langle \varepsilon D_x\rangle^{-s}u\|^2_{L^2(\mathbb{R}^d)} \int_{\mathbb{R}^d} \sup_{\xi\in\mathbb{R}^d} |\widehat{a}(q,\xi)\langle\xi\rangle^{2s}|\langle\varepsilon q\rangle^s dq \tag{8.8}$$

holds for every $u \in H^{-s}(\mathbb{R}^d)$ and every $a \in \mathcal{S}$.

Proof. This is obtained through the simple inequality $\langle\xi+q\rangle^s \le C_{s,d}\langle\xi\rangle^s\langle q\rangle^s$, which holds when $s \ge 0$. □

Notice that whenever $a \in \mathcal{S}$, the integrals $\int_{\mathbb{R}^d} \sup_{\xi\in\mathbb{R}^d} |\widehat{a}(q,\xi)\langle\xi\rangle^{2s}|\langle\varepsilon q\rangle^s dq$ are uniformly bounded for $0 < \varepsilon \le 1$. Consequently, we get the following corollary.

COROLLARY 8.8. *Let (ε_k) and (u_k) satisfy the hypotheses of Proposition* 8.1. *Then the sequence $(m^{\varepsilon_k}[u_k])$ is bounded in $\mathcal{S}'$.*

Estimate (8.8) immediately gives the following.

Remark 8.9. Lemma 8.7 shows that $m^\varepsilon[u]$ acts continuously on test functions a in the closure of $\mathcal{S}$ for the norm

$$[a]_s := \int_{\mathbb{R}^d} \sup_{\xi\in\mathbb{R}^d} |\widehat{a}(q,\xi)\langle\xi\rangle^{2s}|\langle q\rangle^s dq < \infty. \tag{8.9}$$

This closure contains the space

$$\Sigma^s := \{\langle D_x\rangle^s\langle\xi\rangle^{2s}a \in C_0(\mathbb{R}^d\times\mathbb{R}^d) : [a]_s < \infty\}. \tag{8.10}$$

Remark 8.10. Consequently, if (u_k) is as in Proposition 8.1 and $(m^{\varepsilon_k}[u_k])$ converges weakly in $\mathcal{S}'$, then $\langle m^{\varepsilon_k}[u], a\rangle$ converges as well for every $a \in \Sigma^s$.

Proof of Proposition 8.1. The boundedness of the sequence $(m^{\varepsilon_k}[u_k])$ was proved in Corollary 8.8. Suppose now that the distributions $m^{\varepsilon_k}[u_k]$ weakly converge to some $\mu \in \mathcal{S}'$. We next show by means of a localization argument that μ is a positive distribution and thus, due to Schwartz's theorem, a positive Radon measure.

Take $\phi \in \mathcal{S}(\mathbb{R}^d_\xi)$; Lemma 8.5 gives

$$\lim_{k\to\infty} \langle m^{\varepsilon_k}[\phi(\varepsilon_k D_x)u_k], a\rangle_{\mathcal{S}'\times\mathcal{S}} = \int_{\mathbb{R}^d\times\mathbb{R}^d} a(x,\xi)|\phi(\xi)|^2 d\mu(x,\xi)$$

for every $a \in \mathcal{S}$. Since $(\phi(\varepsilon_k D_x)u_k)$ is a bounded sequence in $L^2(\mathbb{R}^d)$, Theorem 1.1 ensures that $|\phi(\xi)|^2\mu$ is a positive Radon measure (and hence a positive distribution). But $\phi \in \mathcal{S}(\mathbb{R}^d_\xi)$ is arbitrary, so μ itself is positive and we obtain the desired result. □

Notice that a very similar proof would give a version of Proposition 8.1 in the context of weighted spaces $L^2(\mathbb{R}^d; \langle x\rangle^r)$.

8.3. Proof of Proposition 8.3. The key ingredient in the proof of the proposition is the following auxiliary result.

LEMMA 8.11. *Under the assumptions of Proposition* 8.3*,*

$$\lim_{k\to\infty}\left|\left\langle m^{\varepsilon_k}[\phi(h_kD_x)u_k]-\left|\phi\left(\frac{h_k}{\varepsilon_k}\xi\right)\right|^2 m^{\varepsilon_k}[u_k],a\right\rangle\right|=0 \tag{8.11}$$

holds for every $a\in\mathcal{S}$ if any of the following conditions hold:

(i) *$h_k=\varepsilon_k$ and a vanishes on the set of discontinuity points of ϕ.*

(ii) *$h_k\ll\varepsilon_k$ and ϕ is continuous at $\xi=0$.*

Proof. Take $a\in\mathcal{S}$ and set $\Phi_k(\xi):=\phi(h_k/\varepsilon_k\xi)$. From relations (8.5), (8.2.i), and (8.7) we obtain

$$\left|\left\langle m^{\varepsilon_k}[\phi(h_kD_x)u_k]-\left|\phi\left(\frac{h_k}{\varepsilon_k}\xi\right)\right|^2 m^{\varepsilon_k}[u_k],a\right\rangle\right|\le M_k(a)\|\langle\varepsilon_kD_x\rangle^{-s}u_k\|^2_{L^2(\mathbb{R}^d)},$$

where

$$M_k(a):=\int_{\mathbb{R}^d}\sup_{\xi\in\mathbb{R}^d}|\widehat{a}(q,\xi)\Phi_k(\xi)[\Phi_k(\xi+\varepsilon_kq)-\Phi_k(\xi)]\langle\xi+\varepsilon_kq\rangle^s\langle\xi\rangle^s|\frac{dq}{(2\pi)^d}; \tag{8.12}$$

recall that $\widehat{a}(q,\xi)$ stands for the Fourier transform of $a(x,\xi)$ in x.

We now must prove that $M_k(a)\to0$ as $k\to\infty$. First we check this for test functions a belonging to the smaller class:

$$\widehat{\mathcal{D}}:=\{a\in\mathcal{S}:\widehat{a}\in C_c^\infty(\mathbb{R}^d\times\mathbb{R}^d)\}.$$

Take $R>0$ such that $\operatorname{supp}a$ is contained in $B(0;R)\times B(0;R)$.

When $k\in\mathbb{N}$ is sufficiently large, $\varepsilon_k\le1$ and

$$\frac{h_k}{\varepsilon_k}(\xi+\varepsilon_kq)\in B(0;2R\sup h_k/\varepsilon_k)\quad\text{for every } q,\xi\in B(0;R). \tag{8.13}$$

Suppose now that (i) holds. If C_ϕ denotes the set of points where ϕ is continuous, then $\Phi_k=\phi$ is uniformly continuous over $C_\phi\cap B(0;R)$ and, consequently,

$$\sup_{q,\xi\in B(0;R)}\mathbf{1}_{C_\phi}(\xi)|\phi(\xi+\varepsilon_kq)-\phi(\xi)|\to0\quad\text{as } k\to\infty$$

because of (8.13).

On the other hand, when $h_k/\varepsilon_k\to0$ and ϕ is continuous at $\xi=0$, again as a consequence of (8.13),

$$\sup_{\xi,q\in B(0;R)}\left|\phi\left(\frac{h_k}{\varepsilon_k}(\xi+\varepsilon_kq)\right)-\phi\left(\frac{h_k}{\varepsilon_k}\xi x\right)\right|\le2\sup_{\xi\in B(0;2h_k/\varepsilon_kR)}|\phi(\xi)-\phi(0)|\to0$$

as $k\to\infty$.

Thus, in either case,

$$\sup_{\xi\in\mathbb{R}^d}|\widehat{a}(q,\xi)\Phi_k(\xi)[\Phi_k(\xi+\varepsilon_kq)-\Phi_k(\xi)]\langle\xi+\varepsilon_kq\rangle^s\langle\xi\rangle^s|\to0\quad\text{as } k\to\infty$$

for every $q\in\mathbb{R}^d$. Lebesgue's dominated convergence theorem gives the convergence to zero of the integrals (8.12). The density of $\widehat{\mathcal{D}}$ in $\mathcal{S}$ concludes the proof of the lemma. ∎

Proof of Proposition 8.3. To prove (i) and (ii) it only needs to be checked that, for any $a \in C_c^\infty(\mathbb{R}^d \times \mathbb{R}^d)$ (if $\varepsilon_k = h_k$, we further require that $a|_{\mathbb{R}^d \times \overline{D_\phi}} \equiv 0$), the functions $|\phi(h_k/\varepsilon_k \xi)|^2 a(x,\xi)$ belong to the class Σ^s. If so, then

$$\lim_{k\to\infty} \left\langle \left| \phi\left(\frac{h_k}{\varepsilon_k}\xi\right)\right|^2 m^{\varepsilon_k}[u_k], a \right\rangle_{\mathcal{S}'\times\mathcal{S}} = \int_{\mathbb{R}^d\times\mathbb{R}^d} |\phi(c\,\xi)|^2 a(x,\xi) d\mu$$

holds with $c := \lim h_k/\varepsilon_k$ because of Remark 8.10. The conclusion would then follow from identity (8.11).

First, notice that $|\phi(h_k/\varepsilon_k \cdot)|^2 a$ are compactly supported and infinitely differentiable in x. When $\varepsilon_k = h_k$ we must verify that $|\phi|^2 a \in \Sigma^s$, which is clearly the case if $a|_{\mathbb{R}^d\times\overline{D_\phi}} \equiv 0$, for then $|\phi|^2 a$ is continuous in ξ.

On the other hand, if $h_k \ll \varepsilon_k$ and ϕ is merely continuous in a ball $B(0;\delta)$, then, for k large enough, $\operatorname{supp} a \subset B(0; h_k/\varepsilon_k \delta)$, and consequently $\phi(h_k/\varepsilon_k \cdot)$ is continuous on $\operatorname{supp} a$. ∎

8.4. Additional properties. The next approximation result is sometimes useful in the computation of Wigner measures.

LEMMA 8.12. *Let (u_k) and (u_k^N) be sequences in $H^{-s}(\mathbb{R}^d)$, $s \geq 0$, satisfying* (8.1) *with the same bound and*

$$\limsup_{k\to\infty} \|\langle \varepsilon_k D_x\rangle^{-s}(u_k - u_k^N)\|_{L^2(\mathbb{R}^d)} \to 0 \quad \text{as } N \to \infty.$$

Suppose that $m^{\varepsilon_k}[u_k]$ and $m^{\varepsilon_k}[u_k^N]$ converge, respectively, to μ and μ_N. Then

$$\mu_N \rightharpoonup \mu \quad \text{in } \mathcal{M}_+(\mathbb{R}^d\times\mathbb{R}^d) \quad \text{as } N \to \infty.$$

Proof. This is a simple consequence of the identity

$$\begin{aligned}\langle m^{\varepsilon_k}[u_k] - m^{\varepsilon_k}[u_k^N], a\rangle_{\mathcal{S}'\times\mathcal{S}} = {} & \langle \overline{u_k^N}, a(x,\varepsilon_k D_x)(u_k - u_k^N)\rangle_{\mathcal{S}'\times\mathcal{S}} \\ & + \langle \overline{(u_k - u_k^N)}, a(x,\varepsilon_k D_x) u_k\rangle_{\mathcal{S}'\times\mathcal{S}}.\end{aligned}$$

This gives an estimate:

$$|\langle m^{\varepsilon_k}[u_k] - m^{\varepsilon_k}[u_k^N], a\rangle_{\mathcal{S}'\times\mathcal{S}}| \leq C \|\langle \varepsilon_k D_x\rangle^{-s}(u_k - u_k^N)\|_{L^2(\mathbb{R}^d)};$$

taking limits as $k \to \infty$, we obtain

$$\left| \int_{\mathbb{R}^d\times\mathbb{R}^d} a(x,\xi)(d\mu - d\mu_N)\right| \leq C \limsup_{k\to\infty} \|\langle \varepsilon_k D_x\rangle^{s}(u_k - u_k^N)\|_{L^2(\mathbb{R}^d)},$$

and the result follows, since the measures μ_N and μ are equibounded. ∎

We conclude this section with an almost orthogonality result.

LEMMA 8.13. *Let (u_k) and (v_k) be sequences in $H^{-s}(\mathbb{R}^d)$, $s \geq 0$, satisfying* (8.1) *for some scale (ε_k). Suppose that μ and ν, their Wigner measures at scale (ε_k), are mutually singular. Then $m^{\varepsilon_k}[u_k + v_k]$ converges to $\mu + \nu$.*

Proof. A proof of this result for $s = 0$ may be found in [6] or [13]. For the general case, it suffices to take into account Remark 8.4 to conclude that the Wigner measures of $\langle \varepsilon_k D_x\rangle^{-s} u_k$ and $\langle \varepsilon_k D_x\rangle^{-s} v_k$ are $\langle \xi\rangle^{-2s}\mu$ and $\langle \xi\rangle^{-2s}\nu$. These are clearly mutually singular, and thus the aforementioned L^2-version of the present result gives

$$m^{\varepsilon_k}[\langle \varepsilon_k D_x\rangle^{-s}(u_k + v_k)] \rightharpoonup \langle \xi\rangle^{-2s}\mu + \langle \xi\rangle^{-2s}\nu$$

and finally

$$m^{\varepsilon_k}[u_k + v_k] \rightharpoonup \langle\xi\rangle^{2s}(\langle\xi\rangle^{-2s}\mu + \langle\xi\rangle^{-2s}\nu) = \mu + \nu,$$

as claimed. □

Acknowledgments. This article extends and improves some of the results of the author's Ph.D. thesis. He is grateful for the guidance of his Ph.D. advisor Enrique Zuazua. Much of this work was done at the Département de Mathématiques et Applications of the École Normale Supérieure, Paris, where the author spent the 2002–2003 academic year as a post-doctorate fellow. He wishes to acknowledge the hospitality of this institution. Finally, he would like to thank Patrick Gérard for many helpful discussions and suggestions.

REFERENCES

[1] M. Brassart, *The Semi-classical Limit in a Crystal Subject to Exterior Forces*, preprint, 2003.
[2] C. de Boor, *A Practical Guide to Splines*, revised ed., Appl. Math. Sci. 27, Springer-Verlag, New York, 2001.
[3] C. de Boor, R. A. DeVore, and A. Ron, *Approximation from shift-invariant subspaces of* $L^2(\mathbf{R}^d)$, Trans. Amer. Math. Soc., 341 (1994), pp. 787–806.
[4] L. C. Evans, *Weak Convergence Methods for Nonlinear Partial Differential Equations*, CBMS Reg. Conf. Ser. Math. 74, AMS, Providence, RI, 1990.
[5] P. Gérard, *Microlocal defect measures*, Comm. Partial Differential Equations, 16 (1991), pp. 1761–1794.
[6] P. Gérard, *Mesures semi-classiques et ondes de Bloch*, Séminaire sur les Équations aux Dérivées Partielles, 1990–1991, Exp. No. XVI, École Polytech., Palaiseau, 1991.
[7] P. Gérard, *Oscillations and concentration effects in semilinear dispersive wave equations*, J. Funct. Anal., 141 (1996), pp. 60–98.
[8] P. Gérard and E. Leichtnam, *Ergodic properties of eigenfunctions for the Dirichlet problem*, Duke Math. J., 71 (1993), pp. 559–607.
[9] P. Gérard, P. A. Markowich, N. J. Mauser, and F. Poupaud, *Homogenization limits and Wigner transforms*, Comm. Pure Appl. Math., 50 (1997), pp. 323–379.
[10] E. Hernández and G. Weiss, *A First Course on Wavelets*, Stud. Adv. Math., CRC Press, Boca Raton, FL, 1996.
[11] P.-L. Lions, *The concentration-compactness principle in the calculus of variations. The limit case.* I, Rev. Mat. Iberoamericana, 1 (1) (1985), pp. 145–201.
[12] P.-L. Lions, *The concentration-compactness principle in the calculus of variations. The limit case.* II, Rev. Mat. Iberoamericana, 1 (2) (1985), pp. 45–121.
[13] P.-L. Lions and T. Paul, *Sur les mesures de Wigner*, Rev. Mat. Iberoamericana, 9 (1993), pp. 553–618.
[14] P. A. Markowich, N. J. Mauser, and F. A. Poupaud, *Wigner-function approach to (semi) classical limits: Electrons in a periodic potential*, J. Math. Phys., 35 (1994), pp. 1066–1094.
[15] A. Martinez, *An Introduction to Semiclassical and Microlocal Analysis*, Universitext, Springer-Verlag, New York, 2002.
[16] Y. Meyer, *Ondelettes et opérateurs.* I. *Ondelettes*, Actualités Mathématiques, Hermann, Paris, 1990.
[17] A. Ron, *Introduction to shift-invariant spaces. Linear independence*, in Multivariate Approximation and Applications, Cambridge University Press, Cambridge, UK, 2001, pp. 112–151.
[18] L. Tartar, *H-measures, a new approach for studying homogenisation, oscillations and concentration effects in partial differential equations*, Proc. Roy. Soc. Edinburgh Sect. A, 115 (1990), pp. 193–230.
[19] M. Unser, *Sampling—50 years after Shannon*, Proc. IEEE, 88 (2000), pp. 569–587.
[20] E. P. Wigner, *On the quantum correction for thermodynamic equilibrium*, Phys. Rev., 40 (1932), pp. 749–759.

SIAM J. MATH. ANAL.
Vol. 36, No. 2, pp. 384–404

EXISTENCE OF MINIMIZERS IN INCREMENTAL ELASTO-PLASTICITY WITH FINITE STRAINS*

ALEXANDER MIELKE†

Abstract. We consider elasto-plastic deformations of a body which is subjected to a time-dependent loading. The model includes fully nonlinear elasticity as well as the multiplicative split of the deformation gradient into an elastic part and a plastic part. Using the energetic formulation for this rate-independent process we derive a time-incremental problem, which is a minimization problem with respect to the deformation and the plastic variables. We provide assumptions on the constitutive laws of the material which guarantee that the incremental problem can be solved for as many time steps as desired. The methods relies on the polyconvexity of the so-called condensed energy functional and on a priori estimates for the plastic variables using the dissipation distance.

Key words. nonlinear elasticity, plasticity, polyconvexity, time-incremental minimization problems, energetic formulation

AMS subject classifications. Primary, 74C15, 49J40; Secondary, 74A20, 49J52

DOI. 10.1137/S0036141003429906

1. Introduction.

The mathematical theory of linearized elasto-plasticity was developed in the 1970s by Moreau [Mor74, Mor76] and subsequently developed further up to efficient numerical implementations; see, e.g., [Joh76, HaR95]. This theory relies on the additive decomposition

$$\varepsilon = \frac{1}{2}(\mathrm{D}u + \mathrm{D}u^{\mathsf{T}}) = \varepsilon_{\mathrm{elast}} + \varepsilon_{\mathrm{plast}}$$

of the linearized strain tensor ε, where $u : \Omega \to \mathbb{R}^d$ denotes the displacement. Moreover, the energy is assumed to be a quadratic functional such that the problem takes the form of a quasi-variational inequality. More general approaches with nonlinear hardening laws and viscoplastic effects can be found in [BeF96, Alb98, ACZ99, Che01a, Che01b, Nef02].

With this work we want to start a mathematical investigation of elasto-plasticity which allows for large strains and which is based on the multiplicative decomposition

$$F = \mathrm{D}\varphi = F_{\mathrm{elast}}F_{\mathrm{plast}}. \tag{1.1}$$

Here, $\varphi : \Omega \to \mathbb{R}^d$ is the deformation of the body $\Omega \subset \mathbb{R}^d$. The energy $\mathcal{E}$ stored in a deformed body depends only on the elastic part F_{elast} of the deformation tensor and suitable hardening parameters $p \in \mathbb{R}^m$, but not on the plastic part F_{plast}, which is contained in $\mathrm{SL}(\mathbb{R}^d)$ or another Lie group $\mathfrak{G}$ contained in $\mathrm{GL}_+(\mathbb{R}^d) = \{\, P \in \mathbb{R}^{d\times d} \mid \det P > 0 \,\}$. The energy functional takes the form

$$\mathcal{E}(t, \varphi, (F_{\mathrm{plast}}, p)) = \int_\Omega W(x, \mathrm{D}\varphi(x)F_{\mathrm{plast}}(x)^{-1}, p(x))\,\mathrm{d}x - \langle \ell(t), \varphi\rangle,$$

*Received by the editors June 19, 2003; accepted for publication (in revised form) February 13, 2004; published electronically July 29, 2004. This work was partially supported by DFG through SFB 404 *Multifield Problems in Continuum Mechanics* under the subproject C11.
http://www.siam.org/journals/sima/36-2/42990.html
†Institut für Analysis, Dynamik und Modellierung, Universität Stuttgart, Pfaffenwaldring 57, 70569 Stuttgart, Germany (mielke@mathematik.uni-stuttgart.de).

where the external loading $\ell(t)$ is given via

$$\langle \ell(t), \varphi \rangle = \int_\Omega f_{\text{ext}}(t,x) \cdot \varphi(x)\,\mathrm{d}x + \int_\Gamma g_{\text{ext}}(t,x) \cdot \varphi(x)\,\mathrm{d}a.$$

To model the plastic effects, one prescribes either a plastic flow law or, equivalently, a dissipation potential $\Delta : \Omega \times \mathrm{T}(\mathfrak{G} \times \mathbb{R}^m) \to [0,\infty]$. We consider $\Delta(x,\cdot,\cdot)$ as an infinitesimal metric which defines the global dissipation distance $D(x,\cdot,\cdot)$ on $\mathfrak{G} \times \mathbb{R}^m$. Thus, the second ingredient to our material model is the dissipation distance between two internal states $z_j = (F^{(j)}_{\text{plast}}, p_j) : \Omega \to \mathrm{SL}(\mathbb{R}^d) \times \mathbb{R}^m$:

$$\mathcal{D}(z_1,z_2) = \int_\Omega D(x, (F^{(1)}_{\text{plast}}(x), p_1(x)),\, (F^{(2)}_{\text{plast}}(x), p_2(x)))\,\mathrm{d}x.$$

Allowing for finite strains, we are forced to abolish convexity assumptions on the stored-energy density W, since it has to be frame indifferent (i.e., $W(x,RF,z) = W(x,F,z)$ for $R \in \mathrm{SO}(\mathbb{R}^d)$) and enforce local invertibility (i.e., $W(F) = \infty$ for $F \notin \mathrm{GL}_+(\mathbb{R}^d)$). It was a major breakthrough in [Bal77] when it was discovered that these conditions are compatible with quasi-convexity and polyconvexity. The aim of this work is to show that it is possible to find constitutive functions W (being polyconvex) and Δ which, on the one hand, satisfy all the above-mentioned natural, physical conditions of finite-strain elasticity as well as the multiplicative plastic decomposition (1.1) (giving rise to the Lie group structure for $P = F_{\text{plast}}$) and, on the other hand, allow for a mathematical existence theory.

We follow the work found in [MiT99, MTL02, Mie02, Mie03a, MiR03], which shows that rate-independent evolution for elastic materials with internal variables ("standard generalized materials") can be formulated by energy principles as follows. A pair $(\varphi, z) : [0,T] \times \Omega \to \mathbb{R}^d \times \mathrm{SL}(\mathbb{R}^d) \times \mathbb{R}^m$ is called a solution of the elasto-plastic process associated with $\mathcal{E}(t,\cdot,\cdot)$ and $\mathcal{D}$ if *stability* (S) and the *energy inequality* (E) hold:

(S) For all $t \in [0,T]$ we have
$\mathcal{E}(t,\varphi(t),z(t)) \le \mathcal{E}(t,\widetilde{\varphi},\widetilde{z}) + \mathcal{D}(z(t),\widetilde{z})$ for all admissible states $(\widetilde{\varphi},\widetilde{z})$.

(E) For all $s,t \in [0,T]$ with $s < t$ we have
$$\mathcal{E}(s,\varphi(s),z(s)) + \mathrm{Diss}(z,[s,t]) \le \mathcal{E}(t,\varphi(t),z(t)) - \int_s^t \langle \dot{\ell}(\tau,\varphi(\tau)) \rangle\,\mathrm{d}\tau.$$

So far, we are not able to provide existence results for (S)–(E) in the present elasto-plastic setting. However, analogous models in phase transformations [MTL02, MiR03], in delamination [KMR03], in micromagnetism [Kru02, RoK04], and in fracture [FrM93, FrM98, DMT02] have been treated with mathematical success. In these works two major restrictions had to be made: (i) $\mathcal{E}$ has to be convex in the strains (leading to infinitesimal strains), and (ii) the internal variable z has to lie in a closed convex subset of a Banach space. In finite-strain elasto-plasticity these two assumptions are clearly violated. For a more general nonlinear version we refer to [MaM03], where severe compactness assumptions are used to construct solutions. So far it is not clear how this compactness can be established in elasto-plasticity; however, in [MiM04] the first steps are being taken by introducing a suitable regularization.

Since most of the above-mentioned existence results are based on time-incremental approximations we devote this work to an existence theory for the following incremental problem (IP). The hope is that after having developed a suitable existence theory

for (IP) that the methods in [MaM03] could be adjusted to pass to the limit for step size tending to 0 and thus find solutions for (S)–(E).

(IP) For given $t_0 = 0 < t_1 < \cdots < t_N = T$ and z_0, find incrementally, for $k = 1, \dots, N$,

$$(\varphi_k, z_k) \in \operatorname*{Arg\,min}_{(\varphi,z)} [\mathcal{E}(t_k, \varphi, z) + \mathcal{D}(z_{k-1}, z)].$$

Here "Arg min" denotes the set of all global minimizers. Hence, (IP) consists of k minimization problems which are coupled via the dissipation distance. The problem in solving (IP) is that the minimization at the kth step involves the solution z_{k-1} from the previous step. For solving the N minimization problems in (IP) it needs a careful bookkeeping of the properties of the solutions; in particular we have to control the integrability conditions of P_k and P_k^{-1} independently of k. This will be done by the help of the dissipation distance $\mathcal{D}$, whereas the elastic energy $\mathcal{E}$ is used to control the Sobolev norm of φ_k.

Such incremental minimization problems are heavily used in the engineering community (cf. [OrR99, OrS99, MSS99, ORS00, MiL03, MSL02, HaH03]), which justifies studying (IP) in its own right. In fact, existence and nonexistence for (IP) relates to questions of formation of microstructure, localization, or failure; see the discussions in [Mie03a, Mie04]. The failure mechanisms in elasto-plasticity are currently an active research area. However, the aim of our work is to provide examples and to isolate general conditions which exclude these failures. In fact, there are many commercial codes for the numerical simulation of plastic processes (like deep drawing) which are expected to describe nice solutions in regions where no failure arises. We want to contribute to the challenging task of providing a mathematical understanding of these models and hopefully improve the numerical simulation techniques.

The plan of the paper is as follows. In section 2 we introduce the notions of finite-strain elasto-plasticity in detail and establish the relation between the classical flow rules of elasto-plasticity with our energetic formulation (S)–(E). For a more extensive and mechanical treatment we refer to [Mie03a]. In section 3 we start the mathematical analysis by studying the incremental problem (IP) in specific function spaces $\mathcal{F} \times \mathcal{Z}$. To start with, we establish a rather general result which says that any solution $(\varphi_k, z_k)_{k=1,\dots,N}$ of (IP) is stable in the sense of (S) and satisfies a two-sided discretized energy inequality replacing (E).

The key to the analysis of (IP) is realizing that the internal variables $z = (F_{\text{plast}}, p)$ occur under the integral over the body Ω only in a local fashion. Hence, it is possible to minimize in (IP) with respect to z pointwise in $x \in \Omega$. This leads to the condensed energy density

$$W^{\text{cond}}(z_{\text{old}}; F) = \min\{W(FP, p) + D(z_{\text{old}}, (P, p)) \mid (P, p) \in \mathrm{SL}(\mathbb{R}^d) \times \mathbb{R}^m\}.$$

In [CHM02, Mie03a] it is shown that W^{cond} has also mechanical significance, as it contains the effective information of the interplay between energy storage through W and the dissipation mechanism through D. The first major assumption for our existence theory is that $W^{\text{cond}}((\mathbf{1}, p_*); \cdot) : \mathbb{R}^{d\times d} \to \mathbb{R}_\infty$ is polyconvex. The second major assumption is that the condensed energy density W^{cond} and the dissipation distance D are coercive:

$$W^{\text{cond}}((\mathbf{1}, p_*); F) \geq c|F|^{q_F} - C \quad \text{and} \quad D((\mathbf{1}, p_*), (P, p)) \geq c|P|^{q_P} - C.$$

If the growth exponents satisfy $\frac{1}{q_P} + \frac{1}{q_F} \le \frac{1}{q} < \frac{1}{d}$, then the existence of solutions $(\varphi_k, F^{(k)}_{\text{plast}}, p_k)$ for (IP) is obtained with $\varphi_k \in \mathrm{W}^{1,q}(\Omega, \mathbb{R}^d)$ and $F^{(k)}_{\text{plast}} \in \mathrm{L}^{q_P}(\Omega, \mathbb{R}^{d\times d})$.

In section 4 we supply a specific two-dimensional example in which all assumptions can be checked explicitly and are fulfilled for suitable parameter values. Thus, we provide a first existence theory for a multidimensional elasto-plastic incremental problem in the geometric nonlinear case.

In section 5 we treat a one-dimensional example where again the existence theory for (IP) can be carried out explicitly. Using this example we discuss the difficulties in proving the existence of solutions for the time-continuous problem (S)–(E) by letting the step size of the time discretizations go to 0. In section 6, using the very specific properties of the one-dimensional case (like $\operatorname{div}\sigma = 0 \implies \sigma = \text{const.}$), we finally prove a convergence result for the incremental solution which implies that the time-continuous problem (S)–(E) has a solution as well.

2. Elasto-plasticity at finite strain. We consider an elastic body $\Omega \subset \mathbb{R}^d$ which is bounded and has a Lipschitz boundary $\partial\Omega$. A deformation is a mapping $\varphi : \Omega \to \mathbb{R}^d$ such that the deformation gradient $F(x) = \mathrm{D}\varphi(x)$ exists for a.e. $x \in \Omega$ and satisfies

$$F(x) \in \mathrm{GL}_+(\mathbb{R}^d) = \{\, F \in \mathbb{R}^{d\times d} \mid \det F > 0 \,\}.$$

The internal plastic state at a material point $x \in \Omega$ is described by the plastic tensor $P = F_{\text{plast}} \in \mathrm{GL}_+(\mathbb{R}^d)$ and a possibly vector-valued hardening variable $p \in \mathbb{R}^m$. We write the shorthand $z = (P, p)$ to denote the set of all plastic variables. The major assumption in finite-strain elasto-plasticity is the multiplicative decomposition of the deformation gradient F into an elastic and a plastic part,

$$F = F_{\text{elast}} F_{\text{plast}} = F_{\text{elast}} P. \tag{2.1}$$

The point of this decomposition is that the elastic properties will depend only on F_{elast}, whereas previous plastic transformations through P are completely forgotten. However, the hardening variable p will record changes in P and may influence the elastic properties.

The deformation process is governed by two principles. First, we have energy storage which gives rise to the equilibrium equations, and, second, we have dissipation due to plastic transformations which give rise to the plastic flow rule. Energy storage is described by the Gibbs energy

$$\mathcal{E}(t, \varphi, z) = \int_\Omega W(x, \mathrm{D}\varphi(x), z(x))\,\mathrm{d}x - \langle \ell(t), \varphi\rangle, \tag{2.2}$$

where $\langle \ell(t), \varphi\rangle = \int_\Omega f_{\text{ext}}(t,x)\cdot\varphi(x)\,\mathrm{d}x + \int_{\Gamma_{\text{Neu}}} g_{\text{ext}}(t,x)\cdot\varphi(x)\,\mathrm{d}a(x)$ denotes the loading depending on the process time $t \in [0, T]$. The major constitutive assumption is the multiplicative decomposition

$$W(x, F, (P, p)) = \widehat{W}(x, FP^{-1}, p). \tag{2.3}$$

From now on we drop the variable x for notational convenience. However, the whole theory and analysis works in the inhomogeneous case as well.

The dissipation effects are usually modeled by prescribing yield surfaces. For our purpose it is more convenient and mathematically clearer to start on the other side, namely, the dissipation metric. In mechanics this metric is called dissipation

potential, since the dissipational friction forces are obtained from it via differentiation with respect to the plastic rates. We emphasize that the natural setup for the plastic transformation $P \in \mathrm{GL}_+(\mathbb{R}^d)$ is that of an element of a Lie group $\mathfrak{G} \subset \mathrm{GL}_+(\mathbb{R}^d)$. A usual assumption is incompressibility, which gives $\mathfrak{G} = \mathrm{SL}(\mathbb{R}^d) = \{ P \mid \det P = 1 \}$. However, $\mathfrak{G} = \mathrm{GL}_+(\mathbb{R}^d)$ or a single-slip system $\mathfrak{G} = \{ \mathbf{1} + \gamma e_1 \otimes e_2 \mid \gamma \in \mathbb{R} \}$ may also be possible. A dissipation potential is a mapping

$$\Delta : \Omega \times \mathrm{T}(\mathfrak{G} \times \mathbb{R}^m) \to [0, \infty], \tag{2.4}$$

which is called a dissipation metric if it is continuous and $\Delta(x, (P,p), \cdot)$ is convex and positively homogeneous of degree 1:

$$\Delta(x, (P,p), \alpha(\dot P, \dot p)) = \alpha \Delta(x, (P,p), (\dot P, \dot p)) \text{ for } \alpha \geq 0. \tag{2.5}$$

(Again we will drop the variable x for notational convenience.) This condition leads to rate-independent material behavior. One assumes, together with the multiplicative decomposition (2.1), *plastic indifference*:

$$\Delta((P\widehat P, p), (\dot P \widehat P, \dot p)) = \Delta((P,p), (\dot P, \dot p)) \text{ for all } \widehat P \in \mathfrak{G}. \tag{2.6}$$

This amounts in the existence of a function $\widehat\Delta : \mathbb{R}^m \times \mathbb{R}^m \times \mathfrak{g} \to [0, \infty]$ such that

$$\Delta((P,p), (\dot P, \dot p)) = \widehat\Delta(p, \dot p, \dot P P^{-1}). \tag{2.7}$$

Here $\mathfrak{g} = \mathrm{T}_{\mathbf{1}}\mathfrak{G}$ is the Lie algebra associated with the Lie group $\mathfrak{G}$, and $\dot P P^{-1}$ is strictly speaking the right translation of $\dot P(t) \in \mathrm{T}_{P(t)}\mathfrak{G}$ to $\mathfrak{g} = \mathrm{T}_{\mathbf{1}}\mathfrak{G}$.

An important feature of our theory is the induced dissipation distance D on $\mathfrak{G} \times \mathbb{R}^m$ defined via (recall $z = (P,p)$)

$$D(z_0, z_1) = \inf \left\{ \int_0^1 \Delta(z(s), \dot z(s)) \, \mathrm{d}s \mid z \in \mathrm{C}^1([0,1], \mathfrak{G} \times \mathbb{R}^m), z(0) = z_0, z(1) = z_1 \right\}. \tag{2.8}$$

It is important to note that we didn't assume symmetry (i.e., $\Delta(z, -\dot z) \neq \Delta(z, \dot z)$ is allowed), which would contradict hardening. Thus, $D(\cdot, \cdot)$ will not be symmetric either. However, we will often use the triangle inequality

$$D(z_1, z_3) \leq D(z_1, z_2) + D(z_2, z_3), \tag{2.9}$$

which is immediate from the definition. Plastic difference implies that the dissipation distance satisfies

$$D((P_1, p_1), (P_2, p_2)) = D((\mathbf{1}, p_1), (P_2 P_1^{-1}, p_2)). \tag{2.10}$$

Integration over the body Ω gives the total dissipation between two internal states $z_j : \Omega \to \mathfrak{G} \times \mathbb{R}^m$ via

$$\mathcal{D}(z_0, z_1) = \int_\Omega D(z_0(x), z_1(x)) \, \mathrm{d}x. \tag{2.11}$$

To make the energetic formulation mathematically rigorous we define the set of kinematically admissible deformations via

$$\mathcal{F} = \{ \varphi \in \mathrm{W}^{1,q}(\Omega; \mathbb{R}^d) \mid \varphi|_{\Gamma_{\mathrm{Dir}}} = \varphi_{\mathrm{Dir}} \}, \tag{2.12}$$

where $\Gamma_{\rm Dir} = \partial\Omega/\Gamma_{\rm Neu}$ is a part of the boundary with positive surface measure. Moreover, $\varphi_{\rm Dir} = \widehat{\varphi}|_{\Gamma_{\rm Dir}}$, where $\widehat{\varphi} \in {\rm C}^1(\overline{\Omega};\mathbb{R}^d)$ with ${\rm D}\widehat{\varphi}(x) \in {\rm GL}_+(\mathbb{R}^d)$ for all $x \in \overline{\Omega}$. The integrability power q in ${\rm W}^{1,q}$ will be chosen larger than the space dimension d in order to apply the theory of polyconvexity. The loading can then be considered as a function $\ell : [0,T] \to {\rm W}^{1,q}(\Omega,\mathbb{R}^d)^*$, where * denotes the dual space (space of all continuous linear forms).

The set of admissible internal states is simply

$$\mathcal{Z} = \{\, z : \Omega \to \mathfrak{G} \times \mathbb{R}^m \mid z \text{ measurable} \,\}. \tag{2.13}$$

Because of the image space, which is a manifold, it is not clear whether it is reasonable to consider $\mathcal{Z}$ as a subset of a Banach space like $L^1(\Omega,\mathbb{R}^{d\times d}\times\mathbb{R}^m)$. It rather seems natural to equip $\mathcal{Z}$ with the metric $\mathcal{D}$ and use arguments of general metric spaces. Nevertheless, our analysis will be based on states $z = (P,p) \in \mathcal{Z}$ with $P \in {\rm L}^{q_P}(\Omega,\mathbb{R}^{d\times d})$ for a suitable $q_P > 1$. However, the topology on the set $\mathcal{Z}$ will not be important.

DEFINITION 2.1. *A process $(\varphi,z) : [0,T] \to \mathcal{F}\times\mathcal{Z}$ is called a solution of the elasto-plastic problem defined via $\mathcal{E}(t,\cdot,\cdot)$ and $\mathcal{D}$ if the stability condition* (S) *and the energy inequality* (E) *hold:*

(S) *For all $t \in [0,T]$ we have*
$\mathcal{E}(t,\varphi(t),z(t)) \le \mathcal{E}(t,\widetilde{\varphi},\widetilde{z}) + \mathcal{D}(z(t),\widetilde{z})$ *for all $(\widetilde{\varphi},\widetilde{z}) \in \mathcal{F}\times\mathcal{Z}$.*

(E) *For all $s,t \in [0,T]$ with $s < t$ we have*

$$\mathcal{E}(t,\varphi(t),z(t)) + {\rm Diss}(z,[s,t]) \le \mathcal{E}(s,\varphi(s),z(s)) - \int_s^t \langle \dot{\ell}(r),\varphi(r)\rangle \,{\rm d}r. \tag{2.14}$$

Here $-\int_s^t \langle \dot{\ell},\varphi\rangle \,{\rm d}r = \int_s^t \langle \ell,\dot{\varphi}\rangle \,{\rm d}r - \langle \ell,\varphi\rangle|_s^t$ is called the reduced work of the external forces, since $\mathcal{E}$ denotes the Gibbs energy instead of the Helmholtz energy. The dissipation is defined as

$${\rm Diss}(z,[s,t]) = \sup\left\{ \sum_{j=1}^N \mathcal{D}(z(t_{j-1}),z(t_j)) \mid N \in \mathbb{N}, s \le t_0 < \cdots < t_N \le t \right\}$$

for general processes, which equals ${\rm Diss}(z,[s,t]) = \int_s^t \int_\Omega \Delta(z(r,x),\dot{z}(r,x)) \,{\rm d}x \,{\rm d}t$ for differentiable processes.

The major advantage of the energetic formulation via (S) and (E) is that derivatives of neither the constitutive functions W and Δ nor the solution $({\rm D}\varphi,z)$ are needed. Nevertheless, (S) and (E) are strong enough to determine the physically relevant solutions. We refer to [MiT03] for uniqueness results under additional convexity assumptions. Moreover, it is shown in [Mie03a] that sufficiently smooth solutions (φ,z) of (S) and (E) satisfy the classical equations of elasto-plasticity, namely, the *equilibrium equation*

$$\begin{cases} -\operatorname{div} T(t,x) = f_{\rm ext}(t,x) & \text{in } \Omega, \\ \varphi(t,x) = Y_{\rm Dir}(x) & \text{on } \Gamma_{\rm Dir}, \\ T(t,x)\nu(x) = g_{\rm ext}(t,x) & \text{on } \Gamma_{\rm Neu}, \end{cases} \tag{2.15}$$

with $T(t,x) = \frac{\partial}{\partial F} W({\rm D}\varphi(t,x),z(t,x)) = \frac{\partial}{\partial F_{\rm elast}} \widehat{W}({\rm D}\varphi(t,x)P(t,x)^{-1},p(t,x))P(t,x)^{-\mathsf{T}}$, and the *flow rule*

$$0 \in \partial_{\dot{z}}^{\rm sub} \Delta(z(t,x),\dot{z}(t,x)) - Q(t,x), \tag{2.16}$$

where $\partial^{\rm sub}_{\dot z}\Delta(z,\dot z)$ denotes the subgradient of the convex function $\Delta(z,\cdot):\mathrm{T}_z(\mathfrak{G}\times\mathbb{R}^m)\to[0,\infty]$ and Q is the driving force thermodynamically conjugated to z, i.e.,

$$Q=-\frac{\partial}{\partial(P,p)}W(F,(P,p))=\left(P^{-\mathsf{T}}F^{\mathsf{T}}\frac{\partial}{\partial F_{\rm elast}}\widehat{W}(FP^{-1},p)P^{-\mathsf{T}},-\frac{\partial}{\partial p}\widehat{W}(FP{-}1,p)\right).$$

Defining the elastic domain as $\mathbb{Q}(z)=\partial^{\rm sub}_{\dot z}\Delta(z,0)\subset\mathrm{T}^*_z(\mathfrak{G}\times\mathbb{R}^m)$, the Legendre–Fenchel transform shows that (2.16) is equivalent to

$$\dot z\in\partial\mathcal{X}_{\mathbb{Q}(z)}(Q)=\mathrm{N}_Q\mathbb{Q}(z). \tag{2.17}$$

If $\mathbb{Q}(z)$ is given by a yield function Φ in the form

$$\mathbb{Q}(z)=\{\,Q\mid\Phi(z,Q)\le 0\,\}$$

and $\frac{\partial}{\partial Q}\Phi(z,Q)\neq 0$ at $\Phi(z,Q)=0$, then (2.17) can be reformulated via the Karush–Kuhn–Tucker conditions

$$\dot z=\lambda\frac{\partial}{\partial Q}\Phi(z,Q),\quad \lambda\ge 0,\quad \Phi(z,Q)\le 0,\quad \lambda\Phi(z,Q)=0.$$

3. Incremental problems. Until now no existence theory for the time-continuous problem (S)–(E) was available, except for the case $d=1$ given in section 5 below. Following the abstract developments in [MiT03] and the applications of the same energetic approach to models for shape-memory alloys [MTL02, MiR03], it is clear that for proving existence results for the highly nonlinear problem (S)–(E) it is essential to provide an existence theory for suitable associated time-discretized problems. Moreover, such incremental problems are the basis of all engineering simulations and, hence, provide a first step to the mathematical understanding of elasto-plasticity.

It was realized in [OrR99, ORS00, CHM02, Mie03a, Mie04] that existence of solutions for the incremental problem is not to be expected in general situations. In fact, nonexistence can be connected either with failure of the material due to localization (e.g., in shear bands) or fracture or with formation of microstructure in material domains of positive measure. Here we present constitutive assumptions which allow us to prove existence of solutions for each incremental step.

We now start with the mathematical analysis and recall that $\mathcal{F}$ and $\mathcal{Z}$ are defined in (2.12) and (2.13), respectively. Consider a time discretization $0=t_0<t_1<\cdots<t_{N-1}<t_N=T$ of the interval $[0,T]$. Moreover, assume that an initial state $(\varphi_0,z_0)\in\mathcal{F}\times\mathcal{Z}$ is given which is stable according to (S) at $t=0$.

(IP) Incremental Problem:
For $k=1,\dots,N$ find $(\varphi_k,z_k)\in\mathcal{F}\times\mathcal{Z}$ such that
$(\varphi_k,z_k)\in\operatorname{Arg\,min}\{\,\mathcal{E}(t_k,\varphi,z)+\mathcal{D}(z_{k-1},z)\mid(\varphi,z)\in\mathcal{F}\times\mathcal{Z}\,\}$. (3.1)

Here "Arg min" denotes the set of global minimizers. The main point is to show that this set is nonempty, i.e., there exists $(\varphi_k,z_k)\in\mathcal{F}\times\mathcal{Z}$ such that

$$\mathcal{E}(t_k,\varphi_k,z_k)+\mathcal{D}(z_{k-1},z_k)=\inf\{\,\mathcal{E}(t_k,\varphi,z)+\mathcal{D}(z_{k-1},z)\mid(\varphi,z)\in\mathcal{F}\times\mathcal{Z}\,\}.$$

We say that the minimum of $\mathcal{E}(t_k,\cdot,\cdot)+\mathcal{D}(z_{k-1},\cdot)$ is attained at the minimizer (φ_k,z_k).

Before we start the analysis of (IP) we first establish a result which emphasizes the fact that the given incremental problem is the most natural one. In particular, it

illuminates the positive role of the dissipation distance $\mathcal{D}$, which is difficult to characterize, as it is defined only implicitly via Δ in (2.8). However, replacing $\mathcal{D}(z_{k-1}, z)$ in (IP) by some approximation (e.g., $\Delta(z_{k-1}, z_k - z)$) would destroy at least one of the three estimates provided in (i) and (ii) below.

THEOREM 3.1. *Let $(\varphi_k, z_k)_{k=0,\dots,N}$ be any solution of* (IP). *Then the following discrete versions of* (S) *and* (E) *hold:*

(i) *For $k = 0, \dots, N$ the state (φ_k, z_k) is stable at t_k, i.e.,*

$$\mathcal{E}(t_k, \varphi_k, z_k) \le \mathcal{E}(t_k, \widetilde{\varphi}, \widetilde{z}) + \mathcal{D}(z_k, \widetilde{z}) \text{ for all } (\widetilde{\varphi}, \widetilde{z}) \in \mathcal{F} \times \mathcal{Z}.$$

(ii) *For all $s, t \in \{ t_j \mid j = 0, 1, \dots, N \}$ with $s < t$ we have*

$$\begin{aligned} -\int_s^t \langle \dot{\ell}(r), \varphi^{\mathrm{cl}}(r) \rangle \, \mathrm{d}r &\le \mathcal{E}(t, \varphi^{\mathrm{cr}}(t), z^{\mathrm{cr}}(t)) + \mathrm{Diss}(z^{\mathrm{cr}}, [s,t]) - \mathcal{E}(s, \varphi^{\mathrm{cr}}(s), z^{\mathrm{cr}}(s)) \\ &\le -\int_s^t \langle \dot{\ell}(r), \varphi^{\mathrm{cr}}(r) \rangle \, \mathrm{d}r. \end{aligned}$$

Here, φ^{cr} and φ^{cl} are the piecewise constant interpolants which are continuous from the right "cr" and from the left "cl", i.e., $\varphi^{\mathrm{cr}}(t) = \varphi_{k-1}$ for $t \in [t_{k-1}, t_k)$ and $\varphi^{\mathrm{cl}}(t) = \varphi_k$ for $t \in (t_{k-1}, t_k]$ with $\varphi^{\mathrm{cr}}(t_N) = \varphi_N$ and $\varphi^{\mathrm{cl}}(t_0) = \varphi_0$. Hence,

$$\int_{t_j}^{t_k} \langle \dot{\ell}(r), \varphi^{\mathrm{cr}}(r) \rangle \, \mathrm{d}r = \sum_{i=j+1}^{k} \langle \ell(t_i) - \ell(t_{i-1}), \varphi_{i-1} \rangle,$$

and with the same notation for z^{cr} we have $\mathrm{Diss}(z^{\mathrm{cr}}, [t_j, t_k]) = \sum_{i=j+1}^{k} \mathcal{D}(z_{i-1}, z_i)$.

The proof does not need any specific assumptions on the function space $\mathcal{F} \times \mathcal{Z}$ or on the functionals $\mathcal{E}$ and $\mathcal{D}$, since it assumes the existence of a solution. Essential to the proof are the minimization property and the triangle inequality (2.9) for $\mathcal{D}$.

Proof. To simplify the proof we write $y_k = (\varphi_k, z_k)$ and $\widetilde{y} = (\widetilde{\varphi}, \widetilde{z})$.

(i) For arbitrary $\widetilde{y} \in \mathcal{F} \times \mathcal{Z}$ and $k \in \{1, \dots, N\}$ we have

$$\begin{aligned} &\mathcal{E}(t_k, \widetilde{y}) + \mathcal{D}(z_k, \widetilde{z}) = \mathcal{E}(t_k, \widetilde{y}) + \mathcal{D}(z_{k-1}, \widetilde{z}) + \mathcal{D}(z_k, \widetilde{z}) - \mathcal{D}(z_{k-1}, \widetilde{z}) \\ &\ge \mathcal{E}(t_k, y_k) + \mathcal{D}(z_{k-1}, z_k) + \mathcal{D}(z_k, \widetilde{z}) - \mathcal{D}(z_{k-1}, \widetilde{z}) \ge \mathcal{E}(t_k, y_k), \end{aligned}$$

where the first estimate follows since y_k is a minimizer and the second estimate follows from the triangle inequality for $\mathcal{D}$.

(ii) The lower estimate follows since y_{i-1} is stable at t_{i-1}:

$$\begin{aligned} &-\int_{t_{i-1}}^{t_i} \langle \dot{\ell}(r), \varphi^{\mathrm{cl}}(r) \rangle \, \mathrm{d}r = -\langle \ell(t_i), \varphi_i \rangle + \langle \ell(t_{i-1}), \varphi_i \rangle \\ &= \mathcal{E}(t_i, y_i) - \mathcal{E}(t_{i-1}, y_i) = \mathcal{E}(t_i, y_i) - \mathcal{E}(t_{i-1}, y_{i-1}) + \mathcal{E}(t_{i-1}, y_{i-1}) - \mathcal{E}(t_{i-1}, y_i) \\ &\le \mathcal{E}(t_i, y_i) - \mathcal{E}(t_{i-1}, y_{i-1}) + \mathcal{D}(z_{i-1}, z_i). \end{aligned}$$

Summing over i from $j+1$ to k gives the lower estimate. The upper estimate follows similarly since y_i is a minimizer at t_i:

$$\mathcal{E}(t_i, y_i) - \mathcal{E}(t_{i-1}, y_{i-1}) + \mathcal{D}(z_{i-1}, z_i) \le \mathcal{E}(t_i, y_{i-1}) - \mathcal{E}(t_{i-1}, y_{i-1}) = -\int_{t_{i-1}}^{t_i} \langle \dot{\ell}, \varphi^{\mathrm{cr}} \rangle \, \mathrm{d}r.$$

Thus, the result is proved. ∎

We now study the existence of solutions to (IP). For this we need specific properties of the space $\mathcal{F} \times \mathcal{Z}$ and strong conditions on the functionals $\mathcal{E}$ and $\mathcal{D}$. In each time step we have to solve the global minimization problem for the functional $\mathcal{I}_k : \mathcal{F} \times \mathcal{Z} \to \mathbb{R}_\infty$, given as

$$\mathcal{I}_k(\varphi, z) := \int_\Omega [W(\mathrm{D}\varphi(x), z(x)) + D(z_{k-1}(x), z(x))]\,\mathrm{d}x - \langle \ell(t_k), \varphi \rangle. \tag{3.2}$$

The special structure here is that $z \in \mathcal{Z}$ occurs under the integral only with its point values and that no derivatives of z appear. We note that $\mathcal{I}_k : \mathcal{F} \times \mathcal{Z} \to \mathbb{R}_\infty$ is not lower semicontinuous because of the geometric nonlinearity coming from the multiplicative decomposition, i.e., $W(F, (P, p)) = \widehat{W}(FP^{-1}, p)$. It is shown in [FKP94, LDR00] that lower semicontinuity of $\mathcal{I}_k$ implies cross-quasi-convexity of

$$(F, P, p) \mapsto W(F, (P, p)) + D(z_{k-1}(x), (P, p)),$$

which in turn implies convexity in $z = (P, p)$. However, this can only be achieved if $F_{\mathrm{elast}} \mapsto \widehat{W}(F_{\mathrm{elast}})$ is convex, but this contradicts the standard axioms of finite-strain elasto-plasticity; see [CHM02] and below.

Of course, lower semicontinuity of $\mathcal{I}_k$ is not necessary, and we may obtain minimizers without it. The idea is that we can minimize with respect to z for each point $x \in \Omega$ separately. To prepare the following result we define the *condensed energy density*

$$W^{\mathrm{cond}}(z_{\mathrm{old}}; F) = \min\{\, W(F, z) + D(z_{\mathrm{old}}, z) \mid z \in \mathfrak{G} \times \mathbb{R}^m \,\}$$

and the condensed functional

$$\mathcal{I}_k^{\mathrm{cond}}(\varphi) = \int_\Omega W^{\mathrm{cond}}(z_{k-1}(x); \mathrm{D}\varphi(x))\,\mathrm{d}x - \langle \ell(t_k), \varphi \rangle.$$

According to [EkT76, Chap. VIII, sect. 1.6] we can choose a measurable *update function*

$$\begin{aligned} &z^{\mathrm{upd}} : (\mathfrak{G} \times \mathbb{R}^m) \times \mathbb{R}^{d\times d} \to \mathfrak{G} \times \mathbb{R}^m \text{ with} \\ &z^{\mathrm{upd}}(z_{\mathrm{old}}; F) \in Z(z_{\mathrm{old}}; F) := \operatorname{Arg\,min}\{\, W(F, z) + D(z_{\mathrm{old}}, z) \mid z \in \mathfrak{G} \times \mathbb{R}^m \,\}, \end{aligned}$$

i.e., $W^{\mathrm{cond}}(z_{\mathrm{old}}; F) = (W(F, z) + D(z_{\mathrm{old}}, z))|_{z = z^{\mathrm{upd}}(z_{\mathrm{old}}; F)}$.

LEMMA 3.2. *Let W and D be nonnegative, measurable functions, such that for each $(z_{\mathrm{old}}; F)$ the function $z \mapsto W(F, z) + D(z_{\mathrm{old}}, z)$ is coercive. Then W^{cond} and z^{upd} as above are well defined. Moreover, we have the following:*

(a) *For all $(\varphi, z) \in \mathcal{F} \times \mathcal{Z}$ we have $\mathcal{I}_k^{\mathrm{cond}}(\varphi) \le \mathcal{I}_k(\varphi, z)$ with equality if and only if $z(x) \in Z(z_{k-1}(x); \mathrm{D}\varphi(x))$ for a.a. $x \in \Omega$.*

(b) *A pair $(\varphi, z) \in \mathcal{F} \times \mathcal{Z}$ minimizes $\mathcal{I}_k$ in (3.2) if and only if φ is a minimizer of $\mathcal{I}_k^{\mathrm{cond}} : \mathcal{F} \to \mathbb{R}_\infty$ and $z(x) \in Z(z_{k-1}(x); \mathrm{D}\varphi(x))$ for a.a. $x \in \Omega$.*

(c) *If $\widetilde{\varphi} \in \mathcal{F}$ minimizes $\mathcal{I}_k^{\mathrm{cond}}$ and $\widetilde{z} \in \mathcal{Z}$ satisfies $\widetilde{z}(x) = z^{\mathrm{upd}}(z_{k-1}(x); \mathrm{D}\widetilde{\varphi}(x))$, then $(\widetilde{\varphi}, \widetilde{z})$ minimizes $\mathcal{I}_k$.*

Proof. Part (a) is obvious, as $W^{\mathrm{cond}}(z_{k-1}; F) \le W(F, z) + D(z_{k-1}, z)$.

For part (b) first assume that $(\varphi, z) \in \mathcal{F} \times \mathcal{Z}$ minimizes $\mathcal{I}_k$ and let $A = \{\, x \in \Omega \mid z(x) \in Z(z_{k-1}(x); \mathrm{D}\varphi(x)) \,\}$. Outside of A we can change z, while keeping φ fixed, such that the integrand $W + D$ becomes strictly smaller. However, decreasing an integrand strictly on a set of positive measure decreases the integral $\mathcal{I}_k$. Hence, A must have measure 0.

Assume that φ minimizes $\mathcal{I}^{\mathrm{cond}}$ and that $z \in \mathcal{Z}$ is given such that A has full measure in Ω. Then $W^{\mathrm{cond}} = W + D$ on A implies $\mathcal{I}_k^{\mathrm{cond}}(\varphi) = \mathcal{I}_k(\varphi, z)$. With part (a) we conclude that (φ, z) minimizes $\mathcal{I}_k$.

Part (c) is obtained exactly the same way, as now $A = \Omega$. ∎

This simple lemma shows that each step in the incremental problem (IP) reduces to a classical variational problem of nonlinear elasticity. Using the multiplicative decomposition (2.3) and the plastic indifference of the dissipation (2.10) we immediately see that W^{cond} satisfies

$$W^{\mathrm{cond}}((P_{\mathrm{old}}, p_{\mathrm{old}}); F) = W^{\mathrm{cond}}((\mathbf{1}, p_{\mathrm{old}}); FP_{\mathrm{old}}^{-1}), \tag{3.3}$$

and thus it is uniquely determined by $W^{\mathrm{cond}}((\mathbf{1}, \cdot); \cdot) : \mathbb{R}^m \times \mathbb{R}^{d\times d} \to \mathbb{R}_\infty$. Similarly, we may choose z^{upd} such that it satisfies

$$z^{\mathrm{upd}}((P_{\mathrm{old}}, p_{\mathrm{old}}); F) = z^{\mathrm{upd}}((\mathbf{1}, p_{\mathrm{old}}); FP_{\mathrm{old}}^{-1}) \begin{pmatrix} P_{\mathrm{old}} & 0 \\ 0 & 1 \end{pmatrix}. \tag{3.4}$$

We now list all assumptions which are stated in terms of W^{cond} and D. Thus, the assumptions are quite implicit, since in practice the stored-energy density W and the dissipation potential Δ are given. From Δ one has to calculate the dissipation distance $D(\cdot, \cdot)$ and then the condensed energy density W^{cond}. However, currently there are no conditions on W and Δ which are known to be sufficient for our conditions. In the next section we provide an example where all these conditions are satisfied.

$$\begin{cases} \text{(i)} & W^{\mathrm{cond}}((\mathbf{1}, \cdot); \cdot) : \mathbb{R}^m \times \mathbb{R}^{d\times d} \to [0, \infty] \text{ and } D(\cdot, \cdot) : (\mathfrak{G} \times \mathbb{R}^m)^2 \to [0, \infty] \\ & \text{are lower semicontinuous.} \\ \text{(ii)} & \text{For each } p \in \mathbb{R}^m \text{ the function } W^{\mathrm{cond}}((\mathbf{1}, p), \cdot) : \mathbb{R}^{d\times d} \to [0, \infty] \\ & \text{is polyconvex.} \\ \text{(iii)} & \text{There exist } C, c > 0, p_* \in \mathbb{R}^m \text{ and exponents } q_F, q_P \geq 1 \text{ such that} \\ & \qquad D((\mathbf{1}, p_*), (P, p)) \geq c|P|^{q_P} - C \\ & \text{for all } (P, p), \text{and} \\ & \qquad W^{\mathrm{cond}}((\mathbf{1}, p); F) \geq c|F|^{q_F} - C \\ & \text{for all } (F, P, p) \text{ with } D((\mathbf{1}, p_*), (P, p)) < \infty. \\ \text{(iv)} & z^{\mathrm{upd}}((\mathbf{1}, \cdot); \cdot) : \mathbb{R}^m \times \mathbb{R}_+^{d\times d} \to \mathfrak{G} \times \mathbb{R}^m \text{ is Borel measurable.} \end{cases} \tag{3.5}$$

Note that we do not need any additional assumptions on W or Δ.

THEOREM 3.3. *Let the assumptions* (3.5) *be satisfied such that additionally*

$$\frac{1}{q_F} + \frac{1}{q_P} \leq \frac{1}{q} < \frac{1}{d}$$

holds, where q occurs in the definition of $\mathcal{F}$ in (2.12).

Then, for each $z_0 \in \mathcal{Z}$ with $\mathcal{D}((\mathbf{1}, p_*), z_0) = \int_\Omega D((\mathbf{1}, p_*), (P_0(x), p_0(x)))\,\mathrm{d}x < \infty$ and each $\ell \in \mathrm{C}^0([0, T], \mathrm{W}^{1,q}(\Omega, \mathbb{R}^d)^*)$ the incremental problem (IP) *(see* (3.1)*) has a solution $((\varphi_k, z_k))_{k=1,\ldots,N}$ with*

$$\varphi_k \in \mathcal{F} \subset \mathrm{W}^{1,q}(\Omega, \mathbb{R}^d) \quad \textit{and} \quad z_k = z^{\mathrm{upd}}(z_{k-1}; \mathrm{D}\varphi_k(\cdot)) \in \mathcal{Z} \cap \mathrm{L}^{q_P}(\Omega, \mathbb{R}^{d\times d}).$$

Proof. Obviously, the result is proved by induction over $k = 1, 2, \ldots, N$.

For the kth step we assume that $z_{k-1} \in \mathcal{Z}$ is known to satisfy $\mathcal{D}((\mathbf{1}, p_*), z_{k-1}) < \infty$, which certainly holds for $k = 1$. With (3.5)(iii) we conclude $P_{k-1} \in \mathrm{L}^{q_P}(\Omega, \mathbb{R}^{d\times d})$. By Lemma 3.2, the kth minimization problem for $\mathcal{I}_k$ (cf. (3.2)) reduces to minimization of $\mathcal{I}_k^{\mathrm{cond}} : \mathcal{F} \to \mathbb{R}_\infty$, where $\mathcal{I}_k^{\mathrm{cond}}(\varphi) = \int_\Omega W_k(x, \mathrm{D}\varphi(x))\,\mathrm{d}x - \langle \ell(t_k), \varphi\rangle$ with

$$W_k(x, F) = W^{\mathrm{cond}}(z_{k-1}(x); F) = W^{\mathrm{cond}}((\mathbf{1}, p_{k-1}(x)); FP_{k-1}(x)^{-1}).$$

Clearly, $W_k : \Omega \times \mathbb{R}^{d\times d} \to [0, \infty]$ is measurable in x and lower semicontinuous in F. Moreover, by (3.5)(iii) we have the lower bound

$$\begin{aligned} W_k(x, F) &\geq c|FP_{k-1}(x)^{-1}|^{q_F} - C \\ &\geq \frac{cq_F}{q}|F|^q - c\left(\frac{q_F}{q} - 1\right)|P_{k-1}(x)|^{q_F q/(q_F - q)} - c, \end{aligned}$$

where we have used $|FP^{-1}| \geq |F|/|P|$ and $|a/b|^{q_F} \geq ra^{q_F/r} - (r-1)b^{q_F/(r-1)}$ with $r = q_F/q > 1$. Using the assumption $\frac{1}{q_P} \leq \frac{1}{q} - \frac{1}{q_F}$ we conclude $W_k(x, F) \geq \widetilde{c}|F|^q - h(x)$ for $\widetilde{c} > 0$ and $h \in \mathrm{L}^1(\Omega)$. Hence, W_k is coercive.

Moreover, the minors (of order s) of the product FP_{k-1}^{-1} are in fact linear combinations of products of the minors (of order s) of F and P_{k-1}^{-1}. Since by (3.5)(ii) W^{cond} is polyconvex we conclude that $F \mapsto W_k(x, F)$ is polyconvex as well.

The existence theory of Ball [Bal76, Bal77] provides $\varphi_k \in \mathcal{F} \subset \mathrm{W}^{1,q}(\Omega, \mathbb{R}^d)$ such that $\mathcal{I}_k^{\mathrm{cond}}(\varphi_k) = \inf\{\, \mathcal{I}_k^{\mathrm{cond}}(\varphi) \mid \varphi \in \mathcal{F}\,\}$. By Lemma 3.2 we see that (φ_k, z_k) with $z_k = z^{\mathrm{upd}}(z_{k-1}; \mathrm{D}\varphi_k) \in \mathcal{Z}$ minimizes $\mathcal{I}_k : \mathcal{F} \times \mathcal{Z} \to \mathbb{R}_\infty$.

To finish the induction we have to show $\mathcal{D}((\mathbf{1}, p_*), z_k) < \infty$. To see this we use the triangle inequality for $\mathcal{D}$ and the minimization property of (φ_k, z_k) in the form of the energy estimate as in part (ii) of Theorem 3.1. We have

$$\begin{aligned} &\mathcal{D}((\mathbf{1}, p_*), z_k) \leq \mathcal{D}((\mathbf{1}, p_*), z_{k-1}) + \mathcal{D}(z_{k-1}, z_k) \\ &\leq \mathcal{D}((\mathbf{1}, p_*), z_{k-1}) + \mathcal{I}_{k-1}^{\mathrm{cond}}(\varphi_{k-1}) - \mathcal{I}_k^{\mathrm{cond}}(\varphi_k) + \langle \ell(t_{k-1}) - \ell(t_k), \varphi_{k-1}\rangle < \infty. \end{aligned}$$

This concludes the induction step, and hence the whole proof. ∎

4. A two-dimensional example. The purpose of this section is to supply a multidimensional example with $\mathfrak{G} = \mathrm{SL}(\mathbb{R}^d)$ where all assumptions of the previous section can be fulfilled. Unfortunately, our example only works in $d = 2$, since it depends on the fact that everything can be calculated explicitly.

We consider the isotropic elastic energy density

$$W : \left\{ \begin{array}{rcl} \mathbb{R}^{2\times 2} & \to & \mathbb{R}_\infty, \\ F & \mapsto & \frac{1}{\alpha}(\nu_1^\alpha + \nu_2^\alpha) + V(\det F), \end{array} \right. \tag{4.1}$$

where $\nu_1, \nu_2 \geq 0$ are the two singular values of F (i.e., the eigenvalues of $(F^{\mathsf{T}}F)^{1/2}$) and $V : \mathbb{R} \to [0, \infty]$ is convex and continuous and satisfies

$$V(\delta) = \infty \text{ for } \delta \leq 0, \quad V(\delta) \nearrow \infty \text{ for } \delta \searrow 0.$$

For the plastic variables we take $z = (P, p) \in \mathrm{SL}(2) \times \mathbb{R}$ with the dissipation metric

$$\Delta(P, p, \dot{P}, \dot{p}) = \left\{ \begin{array}{ll} A'(p)\|\dot{P}P^{-1}\| & \text{for } \dot{p} \geq \|\dot{P}P^{-1}\|, \\ \infty & \text{else.} \end{array} \right. \tag{4.2}$$

Here, $\|\cdot\|$ denotes the classical Euclidean norm on $\mathfrak{g} \subset \mathbb{R}^{2\times 2}$, i.e., $\|\xi\|^2 = \sum_{i,j=1}^{2} \xi_{ij}^2$, and $A(p) = \mathrm{e}^{\beta p}$ for $\beta > 0$. The associated dissipation distance D is plastically invariant and isotropic, i.e.,

$$D((RP_0\widehat{P}, p_0), (RP_1\widehat{P}, p_1)) = D((P_0, p_0), (P_1, p_1))$$

for all arguments. From the analysis in [Mie02, HMM03, Mie03a] we know that

$$D((\mathbf{1}, p_0), (E(s), p_1)) = \begin{cases} \mathrm{e}^{\beta(p_0+\sqrt{2}|s|)} - \mathrm{e}^{\beta p_0} & \text{for } p_1 \geq p_0+\sqrt{2}|s|, \\ \infty & \text{else,} \end{cases} \tag{4.3}$$

where $E(s) = \mathrm{diag}(\mathrm{e}^s, \mathrm{e}^{-s})$, and, for all $R, \widehat{R} \in \mathrm{SO}(2)$,

$$D((\mathbf{1}, p_0), (RE(s)\widehat{R}, p_1)) \geq D((\mathbf{1}, p_0), (E(s), p_1)). \tag{4.4}$$

With this information, it is shown in [Mie03a] that the condensed stored-energy density takes the form

$$W^{\mathrm{cond}}((\mathbf{1}, p); F) = \min_{s\in\mathbb{R}} \frac{1}{\alpha}((\mathrm{e}^{-s}\nu_1)^\alpha + (\mathrm{e}^s\nu_2)^\alpha) + V(\nu_1\nu_2) + \mathrm{e}^{p\beta}(\mathrm{e}^{\sqrt{2}\beta|s|} - 1).$$

To see this, one uses the isotropy of W and D together with (4.4) to deduce that the minimum in W^{cond} with $F = \mathrm{diag}(\nu_1, \nu_2)$ is attained for $P = E(s) = \mathrm{diag}(\mathrm{e}^s, \mathrm{e}^{-s})$ for some $s \in \mathbb{R}$.

The minimum over $s \in \mathbb{R}$ can be evaluated explicitly if we choose $\beta = \alpha/\sqrt{2}$. This gives the final form

$$W^{\mathrm{cond}}((\mathbf{1}, p); F) = V(\nu_1\nu_2) - \mathrm{e}^{\alpha p/\sqrt{2}} + \begin{cases} \frac{2}{\alpha}\sqrt{\nu_1^\alpha(\nu_2^\alpha + b_p)} & \text{for } \nu_1^\alpha \geq \nu_2^\alpha + b_p, \\ \frac{1}{\alpha}(\nu_1^\alpha + \nu_2^\alpha + b_p) & \text{for } |\nu_1^\alpha - \nu_2^\alpha| \leq b_p, \\ \frac{2}{\alpha}\sqrt{\nu_2^\alpha(\nu_1^\alpha + b_p)} & \text{for } \nu_2^\alpha \geq \nu_1^\alpha + b_p, \end{cases}$$

where $b_p = \alpha\mathrm{e}^{\alpha p/\sqrt{2}}$. Moreover, the update functions can be given explicitly as well. With the auxiliary function

$$S(\nu, p) = \begin{cases} -\frac{1}{2\alpha}\log\frac{\nu_1^\alpha}{\nu_2^\alpha + b_p} & \text{for } \nu_1^\alpha \geq \nu_2^\alpha + b_p, \\ 0 & \text{for } |\nu_1^\alpha - \nu_2^\alpha| \leq b_p, \\ \frac{1}{2\alpha}\log\frac{\nu_2^\alpha}{\nu_1^\alpha + b_p} & \text{for } \nu_2^\alpha \geq \nu_1^\alpha + b_p, \end{cases}$$

we find the update functions (for $\det F = \nu_1\nu_2 > 0$)

$$P^{\mathrm{upd}}((\mathbf{1}, p_0); F) = R_F^{-1}E(S(\nu, p_0))R_F \quad \text{and} \quad p^{\mathrm{upd}}((\mathbf{1}, p_0); F) = p_0 + \sqrt{2}|S(\nu, p_0)|,$$

where $\nu_1, \nu_2 > 0$ and R_F are defined via $F = \widehat{R}\,\mathrm{diag}(\nu_1, \nu_2)R_F$ with $\widehat{R}, R_F \in \mathrm{SO}(2)$. Both update functions are locally Lipschitz continuous since R_F is uniquely defined where $S(\nu, p) \neq 0$.

We summarize the properties of W^{cond} and D in the following proposition, which establishes the conditions (3.5).

PROPOSITION 4.1. *Let W and Δ be defined as above with $\beta = \alpha/\sqrt{2}$. Then the following hold:*

(i) $W^{\mathrm{cond}}((\mathbf{1}, \cdot); \cdot) : \mathbb{R} \times \mathbb{R}^{2\times 2} \to \mathbb{R}_\infty$ *is continuous and* $D(\cdot, \cdot) : (\mathrm{SL}(2) \times \mathbb{R})^2 \to [0, \infty]$ *is lower semicontinuous.*

(ii) *For $\alpha \geq 2$ and $p \in \mathbb{R}$ the function $W^{\mathrm{cond}}((\mathbf{1},p);\cdot) : \mathbb{R}^{2\times d} \to \mathbb{R}_\infty$ is polyconvex.*

(iii) *For all $F \in \mathbb{R}^{2\times 2}, p_*, p \in \mathbb{R}$, and $P \in \mathrm{SL}(2)$ with $D((\mathbf{1},p_*),(P,p)) < \infty$ we have*

$$D((\mathbf{1},p_*),(P,p)) \geq \frac{\mathrm{e}^{\alpha p_*/\sqrt{2}}}{2}\Big(\|P\|^\alpha - 1\Big),$$
$$W^{\mathrm{cond}}((\mathbf{1},p);F) \geq \frac{1}{\alpha}\Big(\sqrt{b_p}\, 2^{1-\alpha/2}\, \|F\|^{\alpha/2} - b_p\Big).$$

(iv) *The update function $z^{\mathrm{upd}} = (P^{\mathrm{upd}}, p^{\mathrm{upd}})$ is continuous.*

Proof. Parts (i) and (iv) are immediate from the definitions and formulas. Part (ii) is the most difficult part; its proof is given in [Mie03b].

To prove the lower estimates in (iii) we first note that $P \in \mathrm{SL}(2)$ has the form $P = R_1 \operatorname{diag}(g, 1/g) R_2 = R_1 E(\log g) R_2$. With (4.3) and (4.4) we obtain

$$D((\mathbf{1},p_*),(P,p)) \geq \mathrm{e}^{\alpha p_*/\sqrt{2}}(\mathrm{e}^{\alpha|\log g|} - 1).$$

Using $\|P\| = \sqrt{g^2 + 1/g^2} \leq \sqrt{2}\max\{g, 1/g\} = \sqrt{2}\mathrm{e}^{|\log g|}$ gives the first estimate. For the second estimate we use the explicit form of $W^{\mathrm{cond}}((\mathbf{1},p);F)$ and $V \geq 0$ to find the lower estimate $\frac{2}{\alpha}\sqrt{b_p}(\max\{\nu_1,\nu_2\})^{\alpha/2}$. With $\|F\| = \sqrt{\nu_1^2+\nu_2^2} \leq \sqrt{2}\max\{\nu_1,\nu_2\}$ the desired estimate follows. ∎

Thus, we have shown that this example satisfies the assumptions (3.5) for $\alpha \geq 2$ with $q_F = \alpha/2$ and $q_P = \alpha$. Hence, Theorem 3.3 is applicable if

$$\frac{1}{2} = \frac{1}{d} > \frac{1}{q} \geq \frac{1}{q_F} + \frac{1}{q_P} = \frac{3}{\alpha}$$

holds. We summarize the existence result for this example in the following statement.

THEOREM 4.2. *Let $d = 2$ and $\mathfrak{G} = \mathrm{SL}(2)$. With $\alpha > 6$ and $\beta = \alpha/\sqrt{2}$ let $W : \mathbb{R}^{2\times 2} \to [0,\infty]$ and $\Delta : \mathrm{T}(\mathfrak{G}\times\mathbb{R}) \to [0,\infty]$ be defined via (4.1) and (4.2), respectively. Assume that there exists a $p_* \in \mathbb{R}$ such that the initial condition $z_0 \in \mathcal{Z}$ satisfies $\mathcal{D}((\mathbf{1},p_*),z_0) < \infty$ and let $q = \alpha/3$.*

Then for each $\ell : [0,T] \to (\mathrm{W}^{1,\alpha/3}(\Omega,\mathbb{R}^2))^*$ the incremental problem (IP) (see (3.1)) has a solution $((\varphi_k,z_k))_{k=1,\dots,N} \in (\mathcal{F}\times\mathcal{Z})^N$. Moreover, there exists a constant C which depends only on α, ℓ, and z_0, but neither on the partition $t_1,\dots,t_N$ nor on the solution, such that

$$\|\varphi_k\|_{\mathrm{W}^{1,\alpha/3}} + \|P_k\|_{\mathrm{L}^\alpha} + \|\mathrm{e}^{\alpha p_k/\sqrt{2}}\|_{\mathrm{L}^1} \leq C \text{ for } k = 1,\dots,N.$$

5. A one-dimensional example. The one-dimensional case is quite special and much simpler for two reasons. First, polyconvexity is equivalent to convexity, and, second, the equilibrium equation is an ordinary differential equation which can be solved easily. Nevertheless this case is interesting, since we will be able to discuss the problems with convergence for step size going to 0 of the incremental solutions towards a solution of the time-continuous problem (S)–(E); see (2.14). We will see that general arguments, which are available in higher space dimensions as well, are not sufficient. In section 6, using the special one-dimensional structure, we then prove convergence (of a subsequence) and obtain finally an existence result for (S)–(E).

Again we treat a special case, but far more general constitutive laws W and Δ could be considered. We let

$$W(F) = \begin{cases} \frac{1}{\alpha}(F^\alpha + F^{-\alpha}) & \text{for } F > 0, \\ \infty & \text{else,} \end{cases}$$

$\mathfrak{G} = \mathrm{GL}_+(1) = (0,\infty)$, $z = (P,p) \in \mathfrak{G} \times \mathbb{R}$, and

$$\Delta((P,p),(\dot P,\dot p)) = \begin{cases} \alpha \mathrm{e}^{\alpha p}\dot p & \text{for } \dot p \geq |\dot P/P|, \\ \infty & \text{else.} \end{cases}$$

As in the previous section (see also [Mie03a]), we obtain the dissipation distance

$$\mathcal{D}((P_0,p_0),(P_1,p_1)) = \begin{cases} \mathrm{e}^{\alpha p_1} - \mathrm{e}^{\alpha p_0} & \text{for } p_1 \geq p_0 + |\log P_1/P_0|\,, \\ \infty & \text{else.} \end{cases}$$

From this we find the condensed stored-energy density

$$W^{\mathrm{cond}}((1,p);F) = \frac{1}{\alpha}\begin{cases} 2\sqrt{1+b_pF^\alpha} - b_p & \text{for } F^\alpha \geq b_p + F^{-\alpha}, \\ F^\alpha + F^{-\alpha} & \text{for } |F^\alpha - F^{-\alpha}| \leq b_p, \\ 2\sqrt{1+b_pF^{-\alpha}} - b_p & \text{for } F^{-\alpha} \geq b_p + F^\alpha, \\ \infty & \text{for } F \leq 0, \end{cases} \tag{5.1}$$

where $b_p = \alpha \mathrm{e}^{\alpha p}$. For $F > 0$ the update functions read

$$P^{\mathrm{upd}}((1,p);F) = \begin{cases} F/(1+b_pF^\alpha)^{1/(2\alpha)} & \text{for } F^\alpha \geq b_p + F^{-\alpha}, \\ 1 & \text{for } |F^\alpha - F^{-\alpha}| \leq b_p, \\ F\,(1+b_pF^{-\alpha})^{1/(2\alpha)} & \text{for } F^{-\alpha} \geq b_p + F^\alpha; \end{cases}$$

$$z^{\mathrm{upd}}((1,p);F) = p + \left|\log P^{\mathrm{upd}}((1,p);F)\right|.$$

As in section 4 we see that the abstract theory of section 3 applies for $\alpha > 3$ since $q_F = \alpha/2$ and $q_P = \alpha$ in condition (3.5).

We consider the one-dimensional domain $\Omega = (0,1) \subset \mathbb{R}^1$. The space $\mathcal{F}^q$ of admissible deformation may be either $\mathcal{F}^q_{\mathrm{displ}} = \mathrm{W}^{1,q}_0(\Omega) = \{\,\varphi \in \mathrm{W}^{1,q}(\Omega) \mid \varphi(0) = \varphi(1) = 0\,\}$ or $\mathcal{F}^q_{\mathrm{tract}} = \{\,\varphi \in \mathrm{W}^{1,q}(\Omega) \mid \varphi(0) = 0\,\}$. The loading takes the form

$$\langle \ell(t), \varphi\rangle = \int_0^1 h_{\mathrm{ext}}(t,x)\varphi(x)\,\mathrm{d}x + \sigma_1(t)\varphi(1) = \int_0^1 H_{\mathrm{ext}}(t,x)\varphi'(x)\,\mathrm{d}x,$$

where $H_{\mathrm{ext}}(t,x) = \sigma_1(t) + \int_x^1 h_{\mathrm{ext}}(t,\widetilde{x})\,\mathrm{d}\widetilde{x}$ and $\varphi'(x) = \mathrm{D}\varphi(x) \in \mathbb{R}^{1\times 1}$. At this point it suffices to assume $H_{\mathrm{ext}} \in \mathrm{C}^0([0,T]\times\overline{\Omega})$.

PROPOSITION 5.1. *Fix $\alpha > 3$ and $p_* \in \mathbb{R}$. Then the above one-dimensional model generates an incremental problem* (IP) *as above, and* (IP) *has, for each $z_0 \in \mathcal{Z}$ with $\mathcal{D}((1,p_*),z_0) < \infty$, a unique solution $(\varphi_k,z_k)_{k=1,\dots,N}$.*

Moreover, there exists $C > 0$, which depends only on α, ℓ, and z_0, such that

$$\|\varphi_k\|_{\mathrm{W}^{1,\alpha/3}} + \|P_k\|_{\mathrm{L}^\alpha} + \|P_k^{-1}\|_{\mathrm{L}^\alpha} + \|\mathrm{e}^{\alpha p_k}\|_{\mathrm{L}^1} \leq C \text{ for } k = 1,\dots,N. \tag{5.2}$$

Proof. Using Lemma 3.2 φ_k is a minimizer of the condensed functional $\mathcal{I}^{\mathrm{cond}}_k$ which is based on W^{cond}; see (5.1). Because of $\alpha > 3$, this density, and hence the functional $\mathcal{I}^{\mathrm{cond}}_k$, is strictly convex. Hence, φ_k is uniquely defined for given z_{k-1} and t_k.

For given F and z_{k-1}, the set $\operatorname{Arg\,min}\{\,W(FP)+D(z_{k-1},(P,p))\,|\,(P,p) \in (0,\infty)\times\mathbb{R}\,\}$ contains just one point. Hence, z_k is also uniquely defined. By induction we conclude uniqueness of the whole solution to (IP).

Estimate (5.2) follows the standard energy estimates as given in section 3. ∎

Finally we want to discuss the problem of establishing convergence for the step size $\max\{\, t_k - t_{k-1} \mid k = 1, \dots, N \,\}$ going to 0. In [MiT99, MTL02, MiT03, MaM03] conditions are given which guarantee that from the sequence of the piecewise constant interpolants

$$(\varphi_{\mathrm{cr}}^N, z_{\mathrm{cr}}^N) : \left\{ \begin{array}{ccc} [0,T) & \to & \mathcal{F} \times \mathcal{Z}, \\ t & \mapsto & \sum_{k=0}^{N-1} \chi_{[t_k, t_{k+1})}(t)(\varphi^k, z^k) \end{array} \right. \tag{5.3}$$

a subsequence can be extracted which converges to a solution $(\varphi, z) : [0,T] \to \mathcal{F} \times \mathcal{Z}$ of the time-continuous problem (S)–(E); see (2.14). The dissipation $\mathcal{D}$ can be used to bound possible oscillations in time yielding temporal compactness. The problem is to control possible spatial oscillation, i.e., in $x \in \Omega$.

A crucial tool developed there (see also [Efe03, MaM03, MiR03]) is the set of stable states

$$\mathcal{S}_{[0,T]} = \{\, (t, \varphi, z) \in [0,T] \times \mathcal{F} \times \mathcal{Z} \mid \text{ for all } \widetilde{\varphi}, \widetilde{z} : \mathcal{E}(t, \varphi, z) \le \mathcal{E}(t, \widetilde{\varphi}, \widetilde{z}) + \mathcal{D}(z, \widetilde{z}) \,\}.$$

The important condition in the abstract theory developed in the above-mentioned papers is that any limit $(\varphi, z) : [0,T] \to \mathcal{F} \times \mathcal{Z}$ of the subsequence $(\varphi^{N_m}(t), z^{N_m}(t)) \to (\varphi(t), z(t))$ occurs in a topology in which the stable set $\mathcal{S}_{[0,T]}$ is closed. We want to study this question in our explicit one-dimensional example now.

For simplicity, we restrict ourselves to the traction case $\mathcal{F} = \mathcal{F}_{\mathrm{tract}}^{\alpha/3}$, which allows us to characterize $\mathcal{S}_{[0,T]}$ explicitly. A similar result was obtained already in [Mie03a].

LEMMA 5.2. *In the above one-dimensional example $(t, \varphi, P, p) \in \mathcal{S}_{[0,T]}$ if and only if for a.a. $x \in \Omega$ we have*

$$|(\varphi'/P)^\alpha - (\varphi'/P)^{-\alpha}| \le \alpha \mathrm{e}^{\alpha p} \text{ and } ((\varphi'/P)^{\alpha-1} - (\varphi'/P)^{-\alpha-1})/P = H_{\mathrm{ext}}(t, \cdot). \tag{5.4}$$

Proof. Stability of (t, φ, z) is equivalent to the fact that (φ, z) is a global minimizer of $J : (\widetilde{\varphi}, \widetilde{z}) \mapsto \mathcal{E}(t, \widetilde{\varphi}, \widetilde{z}) + \mathcal{D}(z, \widetilde{z})$. Minimizing with respect to $\widetilde{z} \in \mathcal{Z}$ leads to the condensed functional

$$J^{\mathrm{cond}} : \widetilde{\varphi} \mapsto \int_\Omega W^{\mathrm{cond}}(z(x); \widetilde{\varphi}'(x)) \, \mathrm{d}x - \langle \ell(t), \widetilde{\varphi} \rangle.$$

For $\widetilde{\varphi} = \varphi$ we know that this minimum is attained for $\widetilde{z} = z$, and hence we know

$$W^{\mathrm{cond}}(z(x); \varphi'(x)) = W(\varphi'(x)/P(x)) \text{ for a.a. } x \in \Omega. \tag{5.5}$$

This gives the first condition in (5.4).

Since φ minimizes J^{cond} we have $\mathrm{D}J^{\mathrm{cond}}(\varphi) = 0$, which implies the second condition in (5.4), after using (5.5) once again. Thus, we conclude that (5.4) is necessary. The sufficiency follows from the convexity. ∎

Defining the two-dimensional subsets $M(t,x)$ of $\mathbb{R}^3$ via

$$M(t,x) = \Big\{ (F,P,p) \in (0,\infty)^2 \times \mathbb{R} \;\Big|\; \begin{array}{l} \big|(\frac{F}{P})^\alpha - (\frac{F}{P})^{-\alpha}\big| \le \alpha \mathrm{e}^{\alpha p}, \\ (\frac{F}{P})^{\alpha-1} - (\frac{F}{P})^{-\alpha-1} = P\, H_{\mathrm{ext}}(t,x) \Big\} \subset \mathbb{R}^3, \end{array}$$

the stability condition (5.4) can be reformulated as

$$(\varphi'(x), P(x), p(x)) \in M(t,x) \text{ for a.a. } x \in \Omega.$$

We note that the sets $M(t,x)$ are closed but not convex in $\mathbb{R}^3$. Hence, $\mathcal{S}_{[0,T]}$ is closed in the strong topology of $[0,T]\times\mathcal{F}\times\mathcal{Z}\subset\mathbb{R}\times \mathrm{W}^{1,\alpha/3}(\Omega)\times \mathrm{L}^\alpha(\Omega)\times \mathrm{L}^\alpha(\Omega)$.

However, $\mathcal{S}_{[0,T]}$ is not closed in the weak topology of this Banach space. Yet, so far the a priori estimate (5.2) is the only one available, and from it we obtain just weak convergence (at fixed times $t\in[0,T]$):

$$
\begin{aligned}
\varphi^{N_m}(t) &\rightharpoonup \varphi(t) && \text{in } \mathrm{W}^{1,\alpha/3}(\Omega),\\
\frac{\partial}{\partial x}(\varphi^{N_m}(t))P^{N_m}(t)^{-1} &\rightharpoonup F_{\mathrm{elast}}(t) && \text{in } \mathrm{L}^\alpha(\Omega),\\
P^{N_m}(t) &\rightharpoonup P(t) && \text{in } \mathrm{L}^\alpha(\Omega),\\
P^{N_m}(t)^{-1} &\rightharpoonup K(t) && \text{in } \mathrm{L}^\alpha(\Omega),\\
z^{N_m}(t) &\rightharpoonup p(t) && \text{in } \mathrm{L}^\alpha(\Omega).
\end{aligned}
\tag{5.6}
$$

However, this does not imply $\varphi'(t,x)/P(t,x)=F_{\mathrm{elast}}(t,x)$ or $P(t,x)^{-1}=K(t,x)$ for a.a. $x\in\Omega$, which would be needed to conclude from $((\varphi^{N_m})',P^{N_m},p^{N_m})\in M(t,x)$ the desirable condition $(\varphi',P,p)\in M(t,x)$.

Thus, the convergence of the incremental solutions can be shown only by establishing convergence in stronger topologies. Below we will show that the solutions $(\frac{\mathrm{d}}{\mathrm{d}x}\varphi^k,P_k,p_k)$ converge pointwise in $[0,T]\times\Omega$.

Before providing this result, we want to mention another abstract approach to obtain strong convergence, which is implemented in section 7 of [MiT03]. It relies on the reduced problem where only the internal variable z is kept, whereas the deformation φ is minimized out. We define

$$
\mathcal{I}^{\mathrm{red}}(t,z)=\min\{\,\mathcal{E}(t,\varphi,z)\mid\varphi\in\mathcal{F}\,\}.
$$

In the case of $\mathcal{F}=\mathcal{F}^{\alpha/3}_{\mathrm{tract}}$ this minimization can be made explicit, since $\mathcal{E}$ contains φ only via φ'. We denote by W^* the Legendre–Fenchel transform of W, i.e.,

$$
W^*(\sigma)=\sup\{\,\sigma F-W(F)\mid F\in\mathbb{R}\,\}.
\tag{5.7}
$$

Then $W^*:\mathbb{R}\to\mathbb{R}$ is convex and satisfies $W^*(\sigma)\sim\frac{1}{\alpha_+}\sigma^{\alpha_+}$ for $\sigma\to+\infty$ and $W^*(\sigma)\sim-\frac{1}{\alpha_-}(-\sigma)^{\alpha_-}$ for $\sigma\to-\infty$, where $\alpha_\pm=\frac{\alpha}{\alpha\mp1}$. Moreover, a simple calculation gives

$$
\mathcal{I}^{\mathrm{red}}(t,z)=-\int_0^1 W^*(H_{\mathrm{ext}}(t,x)P(x))\,\mathrm{d}x.
$$

Unfortunately, this functional is concave in P. Hence, the strong convergence theory in the uniformly convex case is not applicable.

6. Convergence in the one-dimensional case. To derive a convergence result we use the very specific structure of the one-dimensional traction problem with $\mathcal{F}=\mathcal{F}^{\alpha/3}_{\mathrm{tract}}$. As already used in Lemma 5.2 the incremental problem has the special property that it can be solved independently for each point $x\in\Omega$ to obtain $(F_k,P_k,p_k)=(\frac{\mathrm{d}}{\mathrm{d}x}\varphi_k(x),P_k(x),p_k(x))$ as the solution of the finite-dimensional, x-dependent minimization problem

$$
\begin{aligned}
&(F_k(x),P_k(x),p_k(x))\\
&\in \operatorname*{Arg\,min}_{(F,P,p)\in\mathbb{R}^3} W(F/P)-H_{\mathrm{ext}}(t_k,x)F+D((P_{k-1}(x),p_{k-1}(x)),(P,p)),
\end{aligned}
$$

which has a unique solution.

We now additionally assume $z_0 = (P_0, p_0) \in \mathrm{C}^0(\overline{\Omega}, \mathbb{R}^2)$ with $P_0(x) > 0$ for all $x \in \overline{\Omega}$. Moreover, the loading should satisfy $H_{\mathrm{ext}} \in \mathrm{C}^1([0,T] \times \overline{\Omega})$. Using energy estimates as for Proposition 5.1, we find a constant $C > 0$, which is independent of $x \in \overline{\Omega}$ and the time discretization, such that all incremental solutions satisfy

$$|F_k(X)| + |P_k(x)| + |1/P_k(x)| + |p_k(x)| \le C \tag{6.1}$$

for all $x \in \overline{\Omega}$ and $k = 0, 1, \dots, N$.

From now on we omit the x-dependence in most cases and use the shorthand $H_k = H_{\mathrm{ext}}(t_k, x)$. Introducing the logarithm $\gamma = \log P$ and eliminating F, we are left with the following incremental problem in $\mathbb{R}^2$:

$$(\gamma_k, p_k) \in \operatorname{Arg\,min}\{\, D((\mathrm{e}^{\gamma_{k-1}}, p_{k-1}), (\mathrm{e}^{\gamma}, p)) - W^*(\mathrm{e}^{\gamma} H_k) \mid \gamma, p \in \mathbb{R} \,\}.$$

Because of the special form of D, this reduces to a scalar problem

$$\begin{aligned} &\gamma_k \in \operatorname{Arg\,min}\{\, \mathrm{e}^{\alpha(p_{k-1} + |\gamma - \gamma_{k-1}|)} - W^*(\mathrm{e}^{\gamma} H_k) \mid \gamma \in \mathbb{R} \,\}, \\ &p_k = p_{k-1} + |\gamma_k - \gamma_{k-1}|. \end{aligned} \tag{6.2}$$

This problem can be solved almost explicitly by using monotonicity arguments relying on the total ordering of the real line.

The essential scalar variable is $\zeta_{k-1}^{\pm} = \gamma_{k-1} \mp p_{k-1} + \log(\pm H_k)$, which allows us to write the iteration (6.2) in the form

$$\begin{pmatrix} \gamma_k \\ p_k \end{pmatrix} = \begin{cases} \begin{pmatrix} \Gamma_+(\zeta_{k-1}^+) - \log H_k \\ \Gamma_+(\zeta_{k-1}^+) - \zeta_{k-1}^+ \end{pmatrix} & \text{if } \Gamma_+(\zeta_{k-1}^+) > \gamma_{k-1} + \log H_k, \\ \begin{pmatrix} \gamma_{k-1} \\ p_{k-1} \end{pmatrix} & \text{if } \Gamma_+(\zeta_{k-1}^+) \le \gamma_{k-1} + \log |H_k| \le \Gamma_-(\zeta_{k-1}^-), \\ \begin{pmatrix} \Gamma_-(\zeta_{k-1}^-) - \log(-H_k) \\ \zeta_{k-1}^- - \Gamma_-(\zeta_{k-1}^-) \end{pmatrix} & \text{if } \Gamma_-(\zeta_{k-1}^-) < \gamma_{k-1} + \log(-H_k), \end{cases} \tag{6.3}$$

where $\Gamma_{\pm}(\zeta) = \operatorname{Arg\,min}\{\, \mathrm{e}^{\pm\alpha(\gamma - \zeta)} - W^*(\pm \mathrm{e}^{\gamma}) \mid \gamma \in \mathbb{R} \,\}$.

We call the first case, where $\gamma_k > \gamma_{k-1}$, plastic loading and the third case, where $\gamma_k < \gamma_{k-1}$, plastic unloading. In the second case the plastic variables do not change. The major observation is that if in a time interval the solution stays either always in cases one and two or always in cases two and three, then the solution can be calculated directly from the initial data when entering this time interval and the loading history, but one does not need to know the solution in between. In particular, the number of steps done in between is irrelevant. We now make this precise.

With $\Gamma_{\pm}(\zeta) \sim \alpha_{\pm}\zeta$ for $\zeta \to -\infty$, $\alpha_- < 1 < \alpha_+$, and the a priori estimate (6.1) we find a constant $H^* > 0$ such that $|H_k| \le H^*$ implies that the second case (no change in the plastic variables) occurs. We now decompose the time interval $[0,T]$ into a finite number of subintervals $J_m = [\tau_{m-1}, \tau_m]$ with $0 = \tau_0 \le \tau_1 < \tau_2 < \cdots < \tau_M = T$ such that $H^* + (-1)^m H(t) \ge 0$ for all $t \in J_m$. For the given time discretization $0 = t_0 < t_1 < \cdots < t_N = T$ we define, for $m = 1, \dots, M$, the exit times $t_{j_m} \in J_m$ of the subintervals J_m via

$$j_0 = 0 \quad \text{and} \quad j_m = \max\{\, k \mid t_k \le \tau_m \,\}.$$

On the subintervals J_m we change the loading H_k into a monotone version $\widetilde{H}_k$, which is defined for $t_k \in J_m$ via

$$\widetilde{H}_k = (-1)^m \max\{\, (-1)^m H(t_n) \mid n \in \{j_{m-1}, \dots, k\} \,\}. \tag{6.4}$$

Hence $(-1)^m \widetilde{H}_k$ is nondecreasing for $k = j_{m-1}, \dots, j_m$.

By induction over the subintervals and by induction over the number of steps inside each subinterval, we obtain the following representation formula.

PROPOSITION 6.1. *Let m be even and $t_k \in J_m$. Then the solution takes the form*

$$\begin{pmatrix} \gamma_k \\ p_k \end{pmatrix} = \begin{pmatrix} \Gamma_+(\gamma_{j_{m-1}} - p_{j_{m-1}} + \log \widetilde{H}_k) - \log \widetilde{H}_k \\ \Gamma_+(\gamma_{j_{m-1}} - p_{j_{m-1}} + \log \widetilde{H}_k) - \gamma_{j_{m-1}} + p_{j_{m-1}} - \log \widetilde{H}_k \end{pmatrix}. \tag{6.5}$$

A similar formula using Γ_- holds for m odd; cf. (6.3).

Finally, we obtain the desired convergence result, which is formulated in terms of functions over $x \in \Omega = (0,1) \subset \mathbb{R}^1$.

THEOREM 6.2. *Consider the one-dimensional traction problem of section* 5 *with $\alpha > 2$ and $H_{\mathrm{ext}} \in \mathrm{C}^1([0,T] \times \overline{\Omega})$. Then there exists a function $(\varphi, P, p) \in \mathrm{C}^0([0,T], \mathrm{W}^{1,\infty}(\Omega) \times \mathrm{L}^\infty(\Omega)^2)$, which is a solution of* (S)–(E) *(cf.* (2.14)*). Moreover, there exists a constant $C > 0$ such that for each time discretization $0 = t_0 < t_1 < \cdots < t_N = T$ the unique solution $(\varphi_k, P_k, p_k)_{k=0,\ldots,N}$ of the incremental problem* (3.1) *satisfies, for $k = 1, \ldots, N$,*

$$\|\varphi(t_k,\cdot) - \varphi_k\|_{\mathrm{W}^{1,\infty}} + \|P(t_k,\cdot) - P_k\|_{\mathrm{L}^\infty} + \|p(t_k,\cdot) - p_k\|_{\mathrm{L}^\infty} \\ \leq C \max\{\, t_n - t_{n-1} \mid n = 1, \ldots, k \,\}.$$

Proof. We use the fact that Proposition 6.1 can be applied in a uniform manner for $x \in \overline{\Omega}$.

First, consider the division into subintervals J_m. Since H_{ext} is continuous, the sets Σ_+ and Σ_- with

$$\Sigma_\pm = \{\, (t,x) \in [0,T] \times \overline{\Omega} \mid \pm H_{\mathrm{ext}}(t,x) \geq H^* \,\}$$

are strictly separated. Because of this, the only restrictions to the subintervals are $J_m(x) \supset \Sigma_+ \cap ([0,T] \times \{x\})$ for even m and $J_m(x) \supset \Sigma_- \cap ([0,T] \times \{x\})$ for odd m. Hence, it is possible to choose the intervals piecewise constant on a finite number of subintervals $\Omega_l = (x_{l-1}, x_l)$. In particular, the number of time intervals $J_m(x)$, $m = 1, \ldots, M_l$, is bounded from above.

Second, we apply the formula (6.5). To show convergence we define the function $\widetilde{H}_{\mathrm{ext}}$ as in (6.4):

$$\widetilde{H}_{\mathrm{ext}}(t,x) = (-1)^m \max\{\, (-1)^m H_{\mathrm{ext}}(s,x) \mid s \in J_m(x) \cap [0,t] \,\}.$$

By Lipschitz continuity of $H_{\mathrm{ext}}(\cdot, x)$ we obtain

$$|\widetilde{H}_k(x) - \widetilde{H}_{\mathrm{ext}}(t_k,x)| \leq C_1 \delta_k \quad \text{with } \delta_k = \max\{\, t_n - t_{n-1} \mid n = 1, \ldots, k \,\}$$

for a constant C_1 independent of $x \in \Omega$ and the partition.

Now, we may take a sequence of partitions $0 < t_1^{N_l} < \cdots < t_{N_l}^{N_l}$ such that the fineness $\widetilde{\delta}_l := \delta_{N_l}^{N_l}$ tends to 0. Now, the exit points $t_{j_m^l(x)}^{N_l}$ have a distance to the end points $\tau_m(x)$ of the intervals $J_m(x)$ of at most $\widetilde{\delta}_l$. Moreover, by induction over m we find that $(\gamma_{j_m^l(x)}(x), p_{j_m^l(x)}(x))$ converges for $l \to \infty$. The limits, called $(\widetilde{\gamma}_m(x), \widetilde{p}_m(x))$, satisfy the recursion

$$\begin{pmatrix} \widetilde{\gamma}_m \\ \widetilde{p}_m \end{pmatrix} = \begin{pmatrix} \Gamma_+(\widetilde{\gamma}_{m-1} - \widetilde{p}_{m-1} + \log \widetilde{H}_{\mathrm{ext}}(\tau_m)) - \log \widetilde{H}_{\mathrm{ext}}(\tau_m) \\ \Gamma_+(\widetilde{\gamma}_{m-1} - \widetilde{p}_{m-1} + \log \widetilde{H}_{\mathrm{ext}}(\tau_m)) - \widetilde{\gamma}_{m-1} + \widetilde{p}_{m-1} - \log \widetilde{H}_{\mathrm{ext}}(\tau_m) \end{pmatrix}$$

for even m and similarly for odd m. The error is bounded by $C_2\widetilde{\delta}_l$, since $\Gamma_\pm$ are Lipschitz continuous.

Third, we define the function $(\gamma, p) : [0,T] \times \Omega \to \mathbb{R}^2$ via

$$\begin{pmatrix} \gamma(t,x) \\ p(t,x) \end{pmatrix} = \begin{pmatrix} \Gamma_+(\widetilde{\gamma}_{m-1}(x) - \widetilde{p}_{m-1}(x) + \log \widetilde{H}_{\rm ext}(t,x)) - \log \widetilde{H}_{\rm ext}(t,x) \\ \Gamma_+(\widetilde{\gamma}_{m-1}(x) - \widetilde{p}_{m-1}(x) + \log \widetilde{H}_{\rm ext}(t,x)) - \widetilde{\gamma}_{m-1}(x) + \widetilde{p}_{m-1}(x) - \log \widetilde{H}_{\rm ext}(t,x) \end{pmatrix}$$

for $t \in J_m(x)$. By our construction the incremental solutions $(t_k^{N_l}, \gamma_k^{N_l}(x), p_k^{N_l}(x))$ converge to $(t, \gamma(t,x), p(t,x))$ with an error bounded by $C_3\delta_l$, uniformly on $[0,T] \times \Omega$.

Finally, it remains to show that (γ, p) define a solution of (S)–(E). Let $\widehat{F}(P,H)$ be the unique minimizer of $F \mapsto W(F/P) - HF$. Then the desired function (φ, P, p) is obtained from (γ, p) via

$$P(t,x) = \mathrm{e}^{\gamma(t,x)} \quad \text{and} \quad \varphi(t,x) = \int_0^x \widehat{F}(P(t,\xi), H_{\rm ext}(t,\xi))\,\mathrm{d}\xi.$$

Since the function $\widehat{F}$ is also Lipschitz continuous, we obtain uniform convergence of the (unique) incremental solutions towards this limit function. Now we use the abstract theorem, Theorem 3.1, which guarantees that the incremental solutions are stable and satisfy the discrete version of the energy inequality. The characterization of the stable sets in Lemma 5.2 show that uniform limits (with pointwise convergence almost everywhere) are stable again, i.e., $(t, \varphi(t), P(t), p(t)) \in \mathcal{S}_{[0,T]}$ for each $t \in [0,T]$. Thus, (S) is established.

Similarly, we start from the discrete energy inequality (ii) in Theorem 3.1 for the incremental solutions (φ^{N_l}, z^{N_l}). For $l \to \infty$ the uniform convergence guarantees that all terms converge:

$$\mathcal{E}(t, \varphi(t), z(t)) + \mathrm{Diss}(z, [s,t]) = \mathcal{E}(s, \varphi(s), z(s)) - \int_s^t \int_\Omega \partial_t H_{\rm ext}(\tau,\xi) \partial_x \varphi(\tau,\xi)\,\mathrm{d}\xi\,\mathrm{d}\tau.$$

For the convergence of the dissipation, uniform convergence is not sufficient. There we use that the piecewise constant interpolants $P^{\rm cr}(\cdot, x)$ are monotone in t when restricted to the subintervals $J_m(x)$ and that $p^{\rm cr}(\cdot, x)$ is always monotone. This, together with the uniform convergence, implies convergence of the dissipation as well. This establishes (E) as an energy equality. ∎

REFERENCES

[ACZ99] J. Alberty, C. Carstensen, and D. Zarrabi, *Adaptive numerical analysis in primal elastoplasticity with hardening*, Comput. Methods Appl. Mech. Engrg., 171 (1999), pp. 175–204.

[Alb98] H.-D. Alber, *Materials with Memory*, Lecture Notes in Math. 1682, Springer-Verlag, Berlin, 1998.

[Bal76] J. M. Ball, *Convexity conditions and existence theorems in nonlinear elasticity*, Arch. Ration. Mech. Anal., 63 (1976), pp. 337–403.

[Bal77] J. Ball, *Constitutive inequalities and existence theorems in nonlinear elastostatics*, in Nonlinear Analysis and Mechanics: Heriot-Watt Symposium (Edinburgh, 1976), Vol. I, Res. Notes Math. 17, Pitman, London, 1977, pp. 187–241.

[BeF96] A. Bensoussan and J. Frehse, *Asymptotic behaviour of the time dependent Norton-Hoff law in plasticity theory and H^1 regularity*, Comment. Math. Univ. Carolin., 37 (1996), pp. 285–304.

[Che01a] K. Chełmiński, *Coercive approximation of viscoplasticity and plasticity*, Asymptot. Anal., 26 (2001), pp. 105–133.

[Che01b] K. Chełmiński, *Perfect plasticity as a zero relaxation limit of plasticity with isotropic hardening*, Math. Methods Appl. Sci., 24 (2001), pp. 117–136.

[CHM02] C. Carstensen, K. Hackl, and A. Mielke, *Non-convex potentials and microstructures in finite-strain plasticity*, Proc. Roy. Soc. London Ser. A, 458 (2002), pp. 299–317.

[DMT02] G. Dal Maso and R. Toader, *A model for quasi-static growth of brittle fractures: Existence and approximation results*, Arch. Ration. Mech. Anal., 162 (2002), pp. 101–135.

[Efe03] M. Efendiev, *On the compactness of the stable set for rate-independent processes*, Commun. Pure Appl. Anal., 2 (2003), pp. 495–509.

[EkT76] I. Ekeland and R. Témam, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1976; corrected reprint, Classics. Appl. Math. 28, SIAM, Philadelphia, 1999.

[FKP94] I. Fonseca, D. Kinderlehrer, and P. Pedregal, *Energy functionals depending on elastic strain and chemical composition*, Calc. Var. Partial Differential Equations, 2 (1994), pp. 283–313.

[FrM93] G. Francfort and J.-J. Marigo, *Stable damage evolution in a brittle continuous medium*, European J. Mech. A Solids, 12 (1993), pp. 149–189.

[FrM98] G. Francfort and J.-J. Marigo, *Revisiting brittle fracture as an energy minimization problem*, J. Mech. Phys. Solids, 46 (1998), pp. 1319–1342.

[HaH03] K. Hackl and U. Hoppe, *On the calculation of microstructures for inelastic materials using relaxed energies*, in IUTAM Symposium on Computational Mechanics of Solids at Large Strains, C. Miehe, ed., Solid Mech. Appl. 108, Kluwer Academic, Dordrecht, The Netherlands, 2003, pp. 77–86.

[HaR95] W. Han and B. D. Reddy, *Computational plasticity: The variational basis and numerical analysis*, Comput. Mech. Adv., 2 (1995), pp. 283–400.

[HMM03] K. Hackl, A. Mielke, and D. Mittenhuber, *Dissipation distances in multiplicative elastoplasticity*, in Analysis and Simulation of Multifield Problems, W. Wendland and M. Efendiev, eds., Springer-Verlag, Berlin, 2003, pp. 87–100,

[Joh76] C. Johnson, *Existence theorems for plasticity problems*, J. Math. Pures Appl. (9), 55 (1976), pp. 431–444.

[KMR03] M. Kočvara, A. Mielke, and T. Roubíček, *A rate-independent approach to the delamination problem*, Math. Mech. Solids, to appear.

[Kru02] M. Kružík, *Variational models for microstructure in shape memory alloys and in micromagnetics and their numerical treatment*, in Proceedings of the Bexbach Kolloquium on Science 2000 (Bexbach, 2000), A. Ruffing and M. Robnik, eds., Shaker-Verlag, Aachen, 2002.

[LDR00] H. Le Dret and A. Raoult, *Variational convergence for nonlinear shell models with directors and related semicontinuity and relaxation results*, Arch. Ration. Mech. Anal., 154 (2000), pp. 101–134.

[MaM03] A. Mainik and A. Mielke, *Existence results for energetic models for rate-independent systems*, Calc. Var. Partial Differential Equations, to appear.

[Mie02] A. Mielke, *Finite elastoplasticity, Lie groups and geodesics on $SL(d)$*, in Geometry, Dynamics, and Mechanics, P. Newton, A. Weinstein, and P. Holmes, eds., Springer-Verlag, New York, 2002, pp. 61–90.

[Mie03a] A. Mielke, *Energetic formulation of multiplicative elasto-plasticity using dissipation distances*, Contin. Mech. Thermodyn., 15 (2003), pp. 351–382.

[Mie03b] A. Mielke, *Necessary and sufficient conditions for polyconvexity of isotropic functions*, J. Convex Anal., to appear.

[Mie04] A. Mielke, *Deriving new evolution equations for microstructures via relaxation of variational incremental problems*, Comput. Methods Appl. Mech. Engrg., to appear.

[MiL03] C. Miehe and M. Lambrecht, *Analysis of microstructure development in shearbands by energy relaxation of incremental stress potentials: Large-strain theory for generalized standard solids*, Internat. J. Numer. Methods Engrg., 58 (2003), pp. 1–41.

[MiM04] A. Mielke and S. Müller, *Lower semi-continuity and existence of minimizers for a functional in elasto-plasticity*, in preparation, 2004.

[MiR03] A. Mielke and T. Roubíček, *A rate-independent model for inelastic behavior of shape-memory alloys*, Multiscale Model. Simul., 1 (2003), pp. 571–597.

[MiT99] A. Mielke and F. Theil, *A mathematical model for rate-independent phase transformations with hysteresis*, in Proceedings of the Workshop on "Models of Continuum Mechanics in Analysis and Engineering," H.-D. Alber, R. Balean, and R. Farwig, eds., Shaker-Verlag, Aachen, 1999, pp. 117–129.

[MiT03] A. Mielke and F. Theil, *On rate-independent hysteresis models*, NoDEA Nonlinear Differential Equations Appl., to appear.

[Mor74] J.-J. Moreau, *On unilateral constraints, friction and plasticity*, in New Variational Techniques in Mathematical Physics (Centro Internaz. Mat. Estivo (C.I.M.E.), II Ciclo, Bressanone, 1973), Edizioni Cremonese, Rome, 1974, pp. 171–322.

[Mor76] J.-J. Moreau, *Application of convex analysis to the treatment of elastoplastic systems*, in Applications of Methods of Functional Analysis to Problems in Mechanics, P. Germain and B. Nayroles, eds., Lecture Notes in Math. 503, Springer-Verlag, Berlin, New York, 1976, pp. 56–89.

[MSL02] C. Miehe, J. Schotte, and M. Lambrecht, *Homogenization of inelastic solid materials at finite strain based on incremental minimization principles. Application to texture analysis of polycrystals*, J. Mech. Physics Solids, 50 (2002), pp. 2123–2167.

[MSS99] C. Miehe, J. Schröder, and J. Schotte, *Computational homogenization analysis in finite plasticity. Simulation of texture development in polycrystalline materials*, Comput. Methods Appl. Mech. Engrg., 171 (1999), pp. 387–418.

[MTL02] A. Mielke, F. Theil, and V. Levitas, *A variational formulation of rate-independent phase transformations using an extremum principle*, Arch. Ration. Mech. Anal., 162 (2002), pp. 137–177.

[Nef02] P. Neff, *Finite multiplicative plasticity for small elastic strains with linear balance equations and grain boundary relaxation*, Contin. Mech. Thermodyn., 15 (2003), pp. 161–195.

[OrR99] M. Ortiz and E. Repetto, *Nonconvex energy minimization and dislocation structures in ductile single crystals*, J. Mech. Phys. Solids, 47 (1999), pp. 397–462.

[OrS99] M. Ortiz and L. Stainier, *The variational formulation of viscoplastic constitutive updates*, Comput. Methods Appl. Mech. Engrg., 171 (1999), pp. 419–444.

[ORS00] M. Ortiz, E. Repetto, and L. Stainier, *A theory of subgrain dislocation structures*, J. Mech. Phys. Solids, 48 (2000), pp. 2077–2114.

[RoK04] T. Roubíček and M. Kružík, *Microstructure evolution model in micromagnetics*, Z. Angew. Math. Phys., 55 (2004), pp. 159–182.

SIAM J. MATH. ANAL.
Vol. 36, No. 2, pp. 405–422

RENORMALIZED ENTROPY SOLUTIONS FOR QUASI-LINEAR ANISOTROPIC DEGENERATE PARABOLIC EQUATIONS*

MOSTAFA BENDAHMANE† AND KENNETH H. KARLSEN†‡

Abstract. We prove the well-posedness (existence and uniqueness) of renormalized entropy solutions to the Cauchy problem for quasi-linear anisotropic degenerate parabolic equations with L^1 data. This paper complements the work by Chen and Perthame [*Ann. Inst. H. Poincaré Anal. Non Linéaire*, 20 (2003), pp. 645–668], who developed a pure L^1 theory based on the notion of kinetic solutions.

Key words. degenerate parabolic equation, quasi-linear, anisotropic diffusion, entropy solution, renormalized solution, uniqueness, existence

AMS subject classifications. 35K65, 35L65

DOI. 10.1137/S0036141003428937

1. Introduction.

We consider the Cauchy problem for quasi-linear anisotropic degenerate parabolic equations with L^1 data. This convection–diffusion-type problem is of the form

$$\partial_t u + \operatorname{div} f(u) = \nabla \cdot (a(u)\nabla u) + F, \qquad u(0,x) = u_0(x), \tag{1.1}$$

where $(t,x) \in (0,T) \times \mathbf{R}^d$; $T > 0$ is fixed; div and ∇ are with respect to $x \in \mathbf{R}^d$; and $u = u(t,x)$ is the scalar unknown function that is sought. The (initial and source) data $u_0(x)$ and $F(t,x)$ satisfy

$$u_0 \in L^1(\mathbf{R}^d), \qquad F \in L^1((0,T) \times \mathbf{R}^d). \tag{1.2}$$

The diffusion function $a(u) = (a_{ij}(u))$ is a symmetric $d \times d$ matrix of the form

$$a(u) = \sigma(u)\sigma(u)^\top \geq 0, \qquad \sigma \in \left(L^\infty_{\mathrm{loc}}(\mathbf{R})\right)^{d \times K}, \qquad 1 \leq K \leq d, \tag{1.3}$$

and hence has entries

$$a_{ij}(u) = \sum_{k=1}^{K} \sigma_{ik}(u)\sigma_{jk}(u), \qquad i,j = 1,\dots,d.$$

The inequality in (1.3) means that for all $u \in \mathbf{R}$

$$\sum_{i,j=1}^{d} a_{ij}(u)\lambda_i\lambda_j \geq 0 \qquad \forall \lambda = (\lambda_1,\dots,\lambda_d) \in \mathbf{R}^d.$$

Finally, the convection flux $f(u)$ is a vector-valued function that satisfies

$$f(u) = (f_1(u),\dots,f_d(u)) \in \left(\mathrm{Lip}_{\mathrm{loc}}(\mathbf{R})\right)^d. \tag{1.4}$$

*Received by the editors June 11, 2003; accepted for publication (in revised form) March 5, 2004; published electronically July 29, 2004. This work was supported by the BeMatA program of the Research Council of Norway and the European network HYKE, funded by the EC as contract HPRN-CT-2002-00282.

http://www.siam.org/journals/sima/36-2/42893.html

†Department of Mathematics, University of Bergen, Johs. Brunsgt. 12, N–5008 Bergen, Norway (mostafab@math.uio.no).

‡Centre of Mathematics for Applications, Department of Mathematics, University of Oslo, P.O. Box 1053, Blindern, N–0316 Oslo, Norway (kennethk@mi.uib.no, http://www.mi.uib.no/~kennethk/).

It is well known that (1.1) possesses discontinuous solutions and that weak solutions are not uniquely determined by their initial data (the scalar conservation law is a special case of (1.1)). Hence (1.1) must be interpreted in the sense of entropy solutions [16, 20, 21]. In recent years the isotropic diffusion case, for example, the equation

$$\partial_t u + \operatorname{div} f(u) = \Delta A(u), \quad A(u) = \int_0^u a(\xi)\, d\xi, \quad 0 \le a \in L^\infty_{\mathrm{loc}}(\mathbf{R}), \tag{1.5}$$

has received much attention, at least when the data are regular enough (say $L^1 \cap L^\infty$) to ensure $\nabla A(u) \in L^2$. Various existence results for entropy solutions of (1.5) (and (1.1)) can be derived from the work by Vol'pert and Hudjaev [21]. Some general uniqueness results for entropy solutions have been proved in the one-dimensional context by Wu and Yin [22] and Bénilan and Touré [2]. In the multidimensional context a general uniqueness result is more recent and was proved by Carrillo [6, 5] using Kružkov's doubling-of-variables device. Various extensions of his result can be found in [4, 13, 14, 15, 17, 18, 19]; see also [7] for a different approach. Explicit "continuous dependence on the nonlinearities" estimates were proved in [10]. In the literature just cited it is essential that the solutions u possess the regularity $\nabla A(u) \in L^2$. This excludes the possibility of imposing general L^1 data, since it is well known that in this case one cannot expect that much integrability.

The general anisotropic diffusion case (1.1) is more delicate and was successfully solved only recently by Chen and Perthame [9]. Chen and Perthame introduced the notion of kinetic solutions and provided a well-posedness theory for (1.1) with L^1 data. Using their kinetic framework, explicit continuous dependence and error estimates for $L^1 \cap L^\infty$ entropy solutions were obtained in [8]. With the only assumption that the data belong to L^1, we cannot expect a solution of (1.1) to be more than L^1. Hence it is in general impossible to make distributional sense to (1.1) (or its entropy formulation). In addition, as already mentioned above, we cannot expect $\sqrt{a(u)}\nabla u$ to be square-integrable, which seems to be an essential condition for uniqueness. Both these problems were elegantly dealt with in [9] using the kinetic approach.

The purpose of the present paper is to offer an alternative "pure" L^1 well-posedness theory for (1.1) based on a notion of renormalized entropy solutions and the classical Kružkov method [16]. The notion of renormalized solutions was introduced by DiPerna and Lions in the context of Boltzmann equations [11]. This notion (and a similar one called entropy solutions) was then adapted to nonlinear elliptic and parabolic equations with L^1 (or measure) data by various authors. We refer to [3] for some recent results in this context and a list of relevant references. Bénilan, Carrillo, and Wittbold [1] introduced a notion of renormalized Kružkov entropy solutions for scalar conservation laws with L^1 data and proved the existence and uniqueness of such solutions. Their theory generalizes the Kružkov well-posedness theory for L^∞ entropy solutions [16].

Motivated by [1, 3] and [9], we introduce herein a notion of renormalized entropy solutions for (1.1) and prove its well-posedness. Let us illustrate our notion of an L^1 solution on the isotropic diffusion equation (1.5) with initial data $u|_{t=0} = u_0 \in L^1$. To this end, let $T_l : \mathbf{R} \to \mathbf{R}$ denote the truncation function at height $l > 0$ and let $\zeta(z) = \int_0^z \sqrt{a(\xi)}\, d\xi$. A renormalized entropy solution of (1.5) is a function $u \in L^\infty(0,T; L^1(\mathbf{R}^d))$ such that (i) $\nabla \zeta(T_l(u))$ is square-integrable on $(0,T) \times \mathbf{R}^d$ for any $l > 0$; (ii) for any convex C^2 entropy-entropy flux triple (η, q, r), with η' bounded and $q' = \eta' f'$, $r' = \eta' a$, there exists for any $l > 0$ a nonnegative bounded Radon

measure μ_l on $(0,T)\times\mathbf{R}^d$, whose total mass tends to zero as $l\uparrow\infty$, such that

$$\begin{aligned}(1.6)\qquad &\partial_t\eta(T_l(u))+\operatorname{div}q(T_l(u))-\Delta r(T_l(u))\\ &\qquad\le -\eta''(T_l(u))\,|\nabla\zeta(T_l(u))|^2+\mu_l(t,x)\quad\text{in }\mathcal{D}'((0,T)\times\mathbf{R}^d).\end{aligned}$$

Roughly speaking, (1.6) expresses the entropy condition satisfied by the truncated function $T_l(u)$. Of course, if u is bounded by M, choosing $l>M$ in (1.6) yields the usual entropy formulation for u, i.e., a bounded renormalized entropy solution is an entropy solution. However, in contrast to the usual entropy formulation, (1.6) makes sense also when u is merely L^1 and possibly unbounded. Intuitively the measure μ_l should be supported on $\{|u|=l\}$ and carry information about the behavior of the "energy" on the set where $|u|$ is large. The requirement is that the energy should be small for large values of $|u|$, that is, the total mass of the renormalization measure μ_l should vanish as $l\uparrow\infty$. This is essential for proving uniqueness of a renormalized entropy solution. Being explicit, the existence proof reveals that

$$\mu_l((0,T)\times\mathbf{R}^d)\le\int_{\{|u_0|>l\}}|u_0|\,dx\to 0\quad\text{as } l\uparrow\infty.$$

We prove existence of a renormalized entropy solution to (1.1) using an approximation procedure based on artificial viscosity [21] and bounded data. We derive a priori estimates and pass to the limit in the approximations.

Uniqueness of renormalized entropy solutions is proved by adapting the doubling-of-variables device due to Kružkov [16]. In the first order case, the uniqueness proof of Kružkov depends crucially on the fact that

$$\nabla_x\Phi(x-y)+\nabla_y\Phi(x-y)=0,\qquad \Phi\text{ smooth function on }\mathbf{R}^d,$$

which allows for a cancellation of certain singular terms. The proof herein for the second order case relies in addition crucially on the following identity involving the Hessian matrices of $\Phi(x-y)$:

$$\nabla_{xx}\Phi(x-y)+2\nabla_{xy}\Phi(x-y)+\nabla_{yy}\Phi(x-y)=0,$$

which, when used together with the parabolic dissipation terms (like the one found in (1.6)), allows for a cancellation of certain singular terms involving the second order operator in (1.1). Compared to [9], our uniqueness proof is new even in the case of bounded entropy solutions.

The remaining part of this paper is organized as follows: In section 2 we introduce the notion of a renormalized entropy solution for (1.1) and state our main well-posedness theorem. The proof of this theorem is given in section 3 (uniqueness) and section 4 (existence).

2. Definitions and statement of main result. We start by defining an entropy-entropy flux triple.

DEFINITION 2.1 (entropy-entropy flux triple). *For any convex C^2 entropy function $\eta:\mathbf{R}\to\mathbf{R}$, the corresponding entropy fluxes*

$$q=(q_1,\dots,q_d):\mathbf{R}\to\mathbf{R}^d\quad\textit{and}\quad r=(r_{ij}):\mathbf{R}\to\mathbf{R}^{d\times d}$$

are defined by $q'(u)=\eta'(u)f'(u)$ and $r'(u)=\eta'(u)a(u)$. We will refer to (η,q,r) as an entropy-entropy flux triple.

For $1 \le k \le K$ and $1 \le i \le d$, we let

$$\zeta_{ik}(u) = \int_0^u \sigma_{ik}(\xi)\, d\xi, \qquad \zeta_k(u) = (\zeta_{1k}(u), \dots, \zeta_{dk}(u)),$$

and for any $\psi \in C(\mathbf{R})$

$$\zeta_{ik}^{\psi}(u) = \int_0^u \psi(\xi)\sigma_{ik}(\xi)\, d\xi, \qquad \zeta_k^{\psi}(u) = \left(\zeta_{1k}^{\psi}(u), \dots, \zeta_{dk}^{\psi}(u)\right).$$

Let us introduce the following set of vector fields:

$$\begin{aligned} &L^2(0,T;L^2(\mathrm{div};\mathbf{R}^d)) \\ &\quad = \left\{ w = (w_1, \dots, w_d) \in \left(L^2((0,T)\times\mathbf{R}^d)\right)^d : \mathrm{div} w \in L^2((0,T)\times\mathbf{R}^d) \right\}. \end{aligned}$$

Following [9] we define an entropy solution as follows.

DEFINITION 2.2 (entropy solution). *An entropy solution of* (1.1) *is a measurable function* $u : (0,T)\times\mathbf{R}^d \to \mathbf{R}$ *satisfying the following conditions:*

(D.1) $u \in L^\infty(0,T;L^1(\mathbf{R}^d)) \cap L^\infty((0,T)\times\mathbf{R}^d)$.

(D.2) *For any* $k = 1, \dots, K$, $\zeta_k(u) \in L^2(0,T;L^2(\mathrm{div};\mathbf{R}^d))$.

(D.3) *(chain rule) For any* $k = 1, \dots, K$ *and* $\psi \in C(\mathbf{R})$,

$$\mathrm{div}\zeta_k^{\psi}(u) = \psi(u)\mathrm{div}\zeta_k(u)$$

a.e. in $(0,T)\times\mathbf{R}^d$ *and in* $L^2((0,T)\times\mathbf{R}^d)$.

(D.4) *Define the parabolic dissipation measure* $n_l^{u,\psi}(t,x)$ *by*

$$n^{u,\psi}(t,x) = \psi(u(t,x)) \sum_{k=1}^{K} \left(\mathrm{div}\zeta_k(u(t,x))\right)^2.$$

For any entropy-entropy flux triple (η, q, r),

$$\begin{aligned} &\partial_t \eta(u) + \sum_{i=1}^{d} \partial_{x_i} q_i(u) - \sum_{i,j=1}^{d} \partial^2_{x_i x_j} r_{ij}(u) \\ &\qquad - \eta'(u)F \le -n^{u,\eta''} \quad \text{in } \mathcal{D}'((0,T)\times\mathbf{R}^d). \end{aligned} \tag{2.1}$$

(D.5) $\operatorname{ess\,lim}_{t\downarrow 0} \|u(t,\cdot) - u_0\|_{L^1(\mathbf{R}^d)} = 0$.

An important contribution of Chen and Perthame [9] is to make explicit the point that the chain rule (D.3) should be included in the definition of an entropy solution in the anisotropic diffusion case. They also note that (D.3) is automatically fulfilled when $a(u)$ is a diagonal matrix, and can then be deleted from Definition 2.2. This applies to the isotropic case (1.5).

Uniqueness of an entropy solution in the sense of Definition 2.2 was proved in [9] using a kinetic formulation and regularization by convolution. The present paper offers an alternative proof based on the more classical Kružkov method of doubling the variables [16].

Let us mention that (D.4) implies that the following Kružkov-type entropy condition holds for all $c \in \mathbf{R}$ (here $A'_{ij}(\cdot) = a_{ij}(\cdot)$):

$$\begin{aligned}&\partial_t |u-c| + \sum_{i=1}^{d} \partial_{x_i} \Big[\operatorname{sign}(u-c)\,(f_i(u) - f_i(c)) \Big] \\ &\quad - \sum_{i,j=1}^{d} \partial^2_{x_i x_j} \Big[\operatorname{sign}(u-c)\,(A_{ij}(u) - A_{ij}(c)) \Big] - \operatorname{sign}(u-c)\,F \le 0.\end{aligned} \tag{2.2}$$

In the isotropic case (1.5), (2.2) simplifies to

$$\partial_t |u-c| + \operatorname{div} \Big[\operatorname{sign}(u-c)\,(f(u) - f(c)) \Big] - \Delta\, |A(u) - A(c)| \le 0. \tag{2.3}$$

After Carrillo's work [6, 5], it is known that (2.3) implies uniqueness in the isotropic case (1.5). In the anisotropic case (1.1), (2.2) is not sufficient for uniqueness. Indeed, it is necessary to explicitly include the parabolic dissipation measure in the entropy condition, as is done in (D.4).

As we discussed in section 1, for unbounded L^1 solutions Definition 2.2 is in general not meaningful. In [9] the authors use a notion of kinetic solutions to handle this problem. It is the purpose of this paper to use instead a notion of renormalized entropy solutions. Before we can introduce this notion, let us recall the definition of the (Lipschitz continuous) truncation function $T_l : \mathbf{R} \to \mathbf{R}$ at height $l > 0$:

$$T_l(u) = \begin{cases} -l, & u < -l, \\ u, & |u| \le l, \\ l, & u > l. \end{cases} \tag{2.4}$$

We then suggest the following notion of an L^1 solution.

DEFINITION 2.3 (renormalized entropy solution). *A renormalized entropy solution of* (1.1) *is a measurable function* $u : (0,T) \times \mathbf{R}^d \to \mathbf{R}$ *satisfying the following conditions:*

(D.1) $u \in L^\infty(0,T; L^1(\mathbf{R}^d))$.

(D.2) *For any* $k = 1, \dots, K$, $\zeta_k(T_l(u)) \in L^2(0,T; L^2(\operatorname{div}; \mathbf{R}^d))$ *for all* $l > 0$.

(D.3) *(renormalized chain rule) For any* $k = 1, \dots, K$ *and* $\psi \in C(\mathbf{R})$,

$$\operatorname{div} \zeta_k^\psi(T_l(u)) = \psi(T_l(u)) \operatorname{div} \zeta_k(T_l(u))$$

a.e. in $(0,T) \times \mathbf{R}^d$ *and in* $L^2((0,T) \times \mathbf{R}^d)$ *for all* $l > 0$.

(D.4) *For* $l > 0$, *introduce the renormalized parabolic dissipation measure*

$$n_l^{u,\psi}(t,x) = \psi(T_l(u(t,x))) \sum_{k=1}^{K} \Big(\operatorname{div} \zeta_k(T_l(u(t,x))) \Big)^2.$$

For any $l > 0$ *and any entropy-entropy flux triple* (η, q, r), *with* $|\eta'|$ *bounded by* K *(for some given* K*), there exists a nonnegative bounded Radon measure* $\mu_l^{u,K}$ *on* $(0,T) \times \mathbf{R}^d$ *such that*

$$\begin{aligned}&\partial_t \eta(T_l(u)) + \sum_{i=1}^{d} \partial_{x_i} q_i(T_l(u)) - \sum_{i,j=1}^{d} \partial^2_{x_i x_j} r_{ij}(T_l(u)) \\ &\quad - \eta'(T_l(u)) F \le -n_l^{u,\eta''} + \mu_l^{u,K} \quad \text{in } \mathcal{D}'((0,T) \times \mathbf{R}^d).\end{aligned} \tag{2.5}$$

(D.5) *The total mass of the renormalization measure* $\mu_l^{u,K}$ *vanishes as* $l \uparrow \infty$*:*

$$\lim_{l\uparrow\infty} \mu_l^{u,K}((0,T)\times\mathbf{R}^d) = 0.$$

(D.6) $\operatorname{ess\,lim}_{t\downarrow 0} \|u(t,\cdot) - u_0\|_{L^1(\mathbf{R}^d)} = 0.$

Note that since $T_l(u) \in L^\infty((0,T)\times\mathbf{R}^d)$, the terms in (2.5) are all well defined. Moreover, if a renormalized entropy solution u belongs to $L^\infty((0,T)\times\mathbf{R}^d)$, then it is also an entropy solution in the sense of Definition 2.2 (let $l \uparrow \infty$ in Definition 2.3).

Our well-posedness result is contained in the following theorem, which is proved in section 3 (uniqueness) and section 4 (existence).

THEOREM 2.1 (well-posedness). *Suppose that* (1.2)*,* (1.3)*, and* (1.4) *hold. Then there exists a unique renormalized entropy solution* u *of* (1.1)*.*

It is worthwhile mentioning that Theorem 2.1 holds under merely local regularity assumptions on $f(u)$, $a(u)$. Moreover, $a(u)$ can be discontinuous, which is of interest in some applications [4].

Remark 2.1. In the isotropic case it is not necessary to include the chain rule (D.3) as a part of Definition 2.3, since it is then automatically fulfilled. Indeed, let $0 \le \sigma \in L^\infty_{\mathrm{loc}}(\mathbf{R})$ and $\psi \in L^\infty_{\mathrm{loc}}(\mathbf{R})$. Set

$$\beta(z) = \int_0^z \sigma(\xi)\,d\xi, \qquad \beta^\psi(z) = \int_0^z \psi(\xi)\sigma(\xi)\,d\xi.$$

Then, for any measurable function $u(x)$ such that $\partial_{x_i}\beta(T_l(u)) \in L^1_{\mathrm{loc}}(\mathbf{R}^d)$, for some fixed $i = 1,\dots,d$, there holds

$$\partial_{x_i}\beta^\psi(T_l(u(x))) = \psi(T_l(u(x)))\partial_{x_i}\beta(T_l(u(x))) \tag{2.6}$$

for a.e. $x \in \mathbf{R}^d$ and in $L^2_{\mathrm{loc}}(\mathbf{R}^d)$ for all $l > 0$. To establish (2.6) we can apply the proof in [9] to the function $v := \beta(T_l(u)) \in L^\infty(\mathbf{R})$, which satisfies $\partial_{x_i} v \in L^1_{\mathrm{loc}}(\mathbf{R}^d)$.

Remark 2.2 (the initial condition). In Definition 2.3 of a renormalized entropy solution we require that the initial condition at $t = 0$ be satisfied in the strong L^1 sense. When proving convergence of approximate solution sequences without having BV estimates at our disposal, it can be difficult to verify condition (D.6) for a limit function. To have a more flexible framework, the initial condition can be included into the renormalized entropy formulation, that is, delete condition (D.6) and require instead that the renormalized entropy inequality in (2.5) hold in $\mathcal{D}'([0,T)\times\mathbf{R}^d)$. In this case, the Radon measure $\mu_l^{u,K}$ should be bounded on $[0,T)\times\mathbf{R}^d$ and satisfy $\lim_{l\uparrow\infty}\mu_l^{u,K}([0,T)\times\mathbf{R}^d) = 0$. Such a weak formulation of the initial condition is much easier to verify for limits of certain approximate solution sequences. This point was made explicit in [12]; see also [13] for degenerate parabolic equations. To prove Theorem 2.1 with a weak formulation of the initial condition we simply have to combine the proof of Theorem 3.1 below with a straightforward adaptation of the arguments in [12, 13] (we leave the details to the reader). Of course, the comments above also apply to Definition 2.3 of an entropy solution.

3. Uniqueness of renormalized entropy solution. For the uniqueness proof, we need a C^1 approximation of $\operatorname{sign}(\cdot)$ and a corresponding C^2 approximation of the Kružkov entropy flux $|\cdot - c|$, $c \in \mathbf{R}$.

For $\varepsilon > 0$, set

$$\operatorname{sign}_\varepsilon(\xi) = \begin{cases} -1, & \xi < -\varepsilon, \\ \sin\left(\frac{\pi}{2\varepsilon}\xi\right), & |\xi| \le \varepsilon, \\ 1, & \xi > \varepsilon. \end{cases} \tag{3.1}$$

For each $c \in \mathbf{R}$, the corresponding entropy function

$$u \mapsto \eta_\varepsilon(u,c) = \int_c^u \operatorname{sign}_\varepsilon(\xi - c)\, d\xi$$

is convex and belongs to $C^2(\mathbf{R})$ with $\eta_\varepsilon'' \in C_c(\mathbf{R})$ and $|\eta_\varepsilon'| \le 1$ (so that the constant K appearing in Definition 2.3 is 1). Moreover, η_ε is symmetric in the sense that $\eta_\varepsilon(u,c) = \eta_\varepsilon(c,u)$ and

$$\eta_\varepsilon(u,c) \to \eta(u,c) := |u-c| \quad \text{as } \varepsilon \downarrow 0.$$

For each $c \in \mathbf{R}$ and $1 \le i,j \le d$, we define the entropy flux functions

$$\begin{aligned} u \mapsto q_i^\varepsilon(u,c) &= \int_c^u \operatorname{sign}_\varepsilon(\xi - c)\, f_i'(\xi)\, d\xi, \\ u \mapsto r_{ij}^\varepsilon(u,c) &= \int_c^u \operatorname{sign}_\varepsilon(\xi - c)\, A_{ij}'(\xi)\, d\xi, \end{aligned} \tag{3.2}$$

where $A_{ij}'(\cdot) = a_{ij}(\cdot)$ for $1 \le i,j \le d$. Then as $\varepsilon \downarrow 0$

$$\begin{aligned} q_i^\varepsilon(u,c) &\to q_i(u,c) := \operatorname{sign}(u-c)\,(f_i(u) - f_i(c)), \\ r_{ij}^\varepsilon(u,c) &\to r_{ij}(u,c) := \operatorname{sign}(u-c)\,(A_{ij}(u) - A_{ij}(c)) \end{aligned} \tag{3.3}$$

for $1 \le i,j \le d$. Let $q^\varepsilon = (q_1^\varepsilon, \dots, q_d^\varepsilon)$, $r^\varepsilon = \left(r_{ij}^\varepsilon\right)$, and similarly for q, r.

We are now ready to prove uniqueness of renormalized entropy solutions.

THEOREM 3.1 (uniqueness). *Suppose that* (1.3) *and* (1.4) *hold. Let u and v be renormalized entropy solutions of* (1.1) *with data $F \in L^1((0,T)\times\mathbf{R}^d)$, $u_0 \in L^1(\mathbf{R}^d)$ and $G \in L^1((0,T)\times\mathbf{R}^d)$, $v_0 \in L^1(\mathbf{R}^d)$, respectively. Then for a.e. $t \in (0,T)$,*

$$\begin{aligned} &\|u(\cdot,t) - v(\cdot,t)\|_{L^1(\mathbf{R}^d)} \\ &\quad \le \|u_0 - v_0\|_{L^1(\mathbf{R}^d)} + \int_0^t \|F(s,\cdot) - G(s,\cdot)\|_{L^1(\mathbf{R}^d)}\, ds. \end{aligned} \tag{3.4}$$

In particular, (1.1) *admits at most one renormalized entropy solution.*

Proof. We shall prove (3.4) using Kružkov's doubling-of-variables method [16]. When it is notationally convenient we drop the domain of integration.

Let $(\eta_\varepsilon, q_i^\varepsilon, r_{ij}^\varepsilon)$ be the entropy flux triple defined above, and denote by μ_l^u, μ_l^v the corresponding renormalization measures.

From the definition of a renormalized entropy solution for $u = u(t,x)$,

$$\begin{aligned} &\int \left(\eta_\varepsilon(T_l(u),c)\partial_t\phi + \sum_{i=1}^d q_i^\varepsilon(T_l(u),c)\partial_{x_i}\phi + \sum_{i,j=1}^d r_{ij}^\varepsilon(T_l(u),c)\partial_{x_ix_j}^2\phi \right) dx\,dt \\ &\quad - \int \operatorname{sign}_\varepsilon(T_l(u) - c)\, F(t,x)\phi\, dx\,dt \\ &\quad \ge \int n^{u,\operatorname{sign}_\varepsilon'(\cdot - c)}(t,x)\phi\, dx\,dt - \int \phi\, d\mu_l^u(t,x) \end{aligned} \tag{3.5}$$

for all $c \in \mathbf{R}$, for all $l > 0$, and for every $0 \le \phi = \phi(t,x) \in \mathcal{D}((0,T) \times \mathbf{R}^d)$.

From the definition of a renormalized entropy solution for $u = u(s,y)$,

$$
\begin{aligned}
&\int \left(\eta_\varepsilon(T_l(v),c)\partial_s\phi + \sum_{i=1}^d q_i^\varepsilon(T_l(v),c)\partial_{y_j}\phi + \sum_{i,j=1}^d r_{ij}^\varepsilon(T_l(v),c)\partial^2_{y_iy_j}\phi \right) dy\,ds \\
&\quad - \int \operatorname{sign}_\varepsilon (T_l(v)-c)\, G(s,y)\phi\, dy\,ds \\
&\qquad \ge \int n^{v,\operatorname{sign}'_\varepsilon(c-\cdot)}(s,y)\phi\, dy\,ds - \int \phi\, d\mu_l^v(s,y)
\end{aligned} \tag{3.6}
$$

for all $c \in \mathbf{R}$, for all $l > 0$, and for every $0 \le \phi = \phi(s,y) \in \mathcal{D}((0,T) \times \mathbf{R}^d)$.

Choose $c = T_l(v(s,y))$ in (3.5) and integrate over (s,y). Choose $c = T_l(u(t,x))$ in (3.6) and integrate over (t,x). Then adding the two resulting inequalities yields

$$
\begin{aligned}
&\int \Bigg(\eta_\varepsilon(T_l(u),T_l(v))\,(\partial_t+\partial_s)\,\phi \\
&\quad + \sum_{i=1}^d \left[q_i^\varepsilon(T_l(u),T_l(v))\partial_{x_i}\phi + q_i^\varepsilon(T_l(v),T_l(u))\partial_{y_i}\phi \right] \\
&\quad + \sum_{i,j=1}^d \left[r_{ij}^\varepsilon(T_l(u),T_l(v))\partial^2_{x_ix_j}\phi + r_{ij}^\varepsilon(T_l(v),T_l(u))\partial^2_{y_iy_j}\phi \right] \Bigg) dx\,dt\,dy\,ds \\
&\qquad - \int \operatorname{sign}_\varepsilon (T_l(u)-T_l(v))\,(F(t,x)-G(s,y))\; dx\,dt\,dy\,ds \\
&\qquad\quad \ge \int \left(n^{u,\operatorname{sign}'_\varepsilon(\cdot-c)}(t,x) + n^{v,\operatorname{sign}'_\varepsilon(\cdot-c)}(s,y) \right) \phi\, dx\,dt\,dy\,ds \\
&\qquad\qquad - \int \phi(t,x,s,y)\, d\mu_l^u(t,x)\, dy\,ds \\
&\qquad\qquad\quad - \int \phi(t,x,s,y)\, d\mu_l^v(s,y)\, dx\,dt,
\end{aligned} \tag{3.7}
$$

where $\phi = \phi(t,x,s,y)$ is any nonnegative function in $\mathcal{D}(((0,T) \times \mathbf{R}^d)^2)$.

We introduce next a standard mollifier sequence $\omega_\rho : \mathbf{R} \times \mathbf{R}^d \to \mathbf{R}$, $\rho > 0$, and take our test function $\phi = \phi(t,x,s,y)$ to be of the form

$$
\phi(t,x,s,y) = \varphi\left(\frac{t+s}{2}, \frac{x+y}{2}\right)\omega_\rho\left(\frac{t-s}{2}, \frac{x-y}{2}\right), \qquad \varphi \in \mathcal{D}((0,T)\times\mathbf{R}^d), \qquad 0 \le \varphi \le 1.
$$

With this choice, we have $(\partial_t+\partial_s)\,\phi = (\partial_t+\partial_s)\,\varphi\left(\frac{t+s}{2}, \frac{x+y}{2}\right)\omega_\rho\left(\frac{t-s}{2}, \frac{x-y}{2}\right)$ and $(\nabla_x+\nabla_y)\,\phi = (\nabla_x+\nabla_y)\,\varphi\left(\frac{t+s}{2}, \frac{x+y}{2}\right)\omega_\rho\left(\frac{t-s}{2}, \frac{x-y}{2}\right)$.

Introduce the Hessian matrices

$$
\nabla_{xx}\phi = \left(\partial^2_{x_ix_j}\phi\right), \quad \nabla_{xy}\phi = \left(\partial^2_{x_iy_j}\phi\right), \quad \nabla_{yy}\phi = \left(\partial^2_{y_iy_j}\phi\right).
$$

Then one can check that the following crucial matrix equality holds:

$$
(\nabla_{xx} + 2\nabla_{xy} + \nabla_{yy})\,\phi = (\nabla_{xx} + 2\nabla_{xy} + \nabla_{yy})\,\varphi\left(\frac{t+s}{2}, \frac{x+y}{2}\right)\omega_\rho\left(\frac{t-s}{2}, \frac{x-y}{2}\right).
$$

Note that the two latter properties imply that for $1 \le i \le d$,

$$
\begin{aligned}
&q_i^\varepsilon(T_l(u),T_l(v))\partial_{x_i}\phi + q_i^\varepsilon(T_l(v),T_l(u))\partial_{y_i}\phi \\
&\quad = q_i^\varepsilon(T_l(u),T_l(v))\left(\partial_{x_i}+\partial_{y_i}\right)\varphi\left(\frac{t+s}{2},\frac{x+y}{2}\right)\omega_\rho\left(\frac{t-s}{2},\frac{x-y}{2}\right) \\
&\qquad + \left[q_i^\varepsilon(T_l(v),T_l(u)) - q_i^\varepsilon(T_l(u),T_l(v))\right]\partial_{y_i}\phi,
\end{aligned}
\tag{3.8}
$$

and for $1 \le i,j \le d$,

$$
\begin{aligned}
&r_{ij}^\varepsilon(T_l(u),T_l(v))\partial^2_{x_ix_j}\phi + r_{ij}^\varepsilon(T_l(v),T_l(u))\partial^2_{y_iy_j}\phi \\
&\quad = r_{ij}^\varepsilon(T_l(u),T_l(v))\left(\partial^2_{x_ix_j}+2\partial^2_{x_iy_j}+\partial^2_{y_iy_j}\right)\varphi\left(\frac{t+s}{2},\frac{x+y}{2}\right)\omega_\rho\left(\frac{t-s}{2},\frac{x-y}{2}\right) \\
&\qquad - 2r_{ij}^\varepsilon(T_l(u),T_l(v))\partial^2_{x_iy_j}\phi \\
&\qquad\quad + \left[r_{ij}^\varepsilon(T_l(v),T_l(u)) - r_{ij}^\varepsilon(T_l(u),T_l(v))\right]\partial^2_{y_iy_j}\phi.
\end{aligned}
\tag{3.9}
$$

We also have

$$
\begin{aligned}
&-\int \phi(t,x,s,y)\,d\mu_l^u(t,x)\,dy\,ds \\
&\qquad \ge -\int \omega_\rho\left(\frac{t-s}{2},\frac{x-y}{2}\right)\,dy\,ds\,d\mu_l^u(t,x) \ge -\mu_l^u((0,T)\times\mathbf{R}^d)
\end{aligned}
\tag{3.10}
$$

and similarly

$$
-\int \phi(t,x,s,y)\,d\mu_l^v(s,y)\,dx\,dt \ge -\mu_l^v((0,T)\times\mathbf{R}^d). \tag{3.11}
$$

Insertion of (3.8)–(3.11) into (3.7) gives

$$
\begin{aligned}
&\int\Bigg(\eta_\varepsilon(T_l(u),T_l(v))\left(\partial_t+\partial_s\right)\varphi\left(\frac{t+s}{2},\frac{x+y}{2}\right) \\
&\qquad + \sum_{i=1}^d q_i^\varepsilon(T_l(u),T_l(v))\left(\partial_{x_i}+\partial_{y_i}\right)\varphi\left(\frac{t+s}{2},\frac{x+y}{2}\right) \\
&\qquad + \sum_{i,j=1}^d r_{ij}^\varepsilon(T_l(u),T_l(v))\left(\partial^2_{x_ix_j}+2\partial^2_{x_iy_j}+\partial^2_{y_iy_j}\right)\varphi\left(\frac{t+s}{2},\frac{x+y}{2}\right)\Bigg) \\
&\qquad\qquad \times \omega_\rho\left(\frac{t-s}{2},\frac{x-y}{2}\right)\,dx\,dt\,dy\,ds \\
&\qquad - \int \operatorname{sign}_\varepsilon\left(T_l(u)-T_l(v)\right)\left(F(t,x)-G(s,y)\right)\,dx\,dt\,dy\,ds \\
&\quad \ge E_1(\varepsilon)+E_2(\varepsilon)+E_3(\varepsilon)-\mu_l^v((0,T)\times\mathbf{R}^d)-\mu_l^v((0,T)\times\mathbf{R}^d),
\end{aligned}
\tag{3.12}
$$

where $E_j(\varepsilon) = \int I_j(\varepsilon)\,dx\,dt\,dy\,ds$, $j = 1, 2, 3$, with

$$
\begin{aligned}
I_1(\varepsilon) &= \left(n^{u,\operatorname{sign}'_\varepsilon(\cdot-c)}(t,x) + n^{v,\operatorname{sign}'_\varepsilon(\cdot-c)}(s,y)\right)\phi,\\
I_2(\varepsilon) &= 2\sum_{i,j=1}^{d} r_{ij}^{\varepsilon}(T_l(u),T_l(v))\partial^2_{x_iy_j}\phi,\\
I_3(\varepsilon) &= \sum_{i=1}^{d}\left[q_i^{\varepsilon}(T_l(u),T_l(v)) - q_i^{\varepsilon}(T_l(v),T_l(u))\right]\partial_{y_i}\phi\\
&\qquad + \sum_{i,j=1}^{d}\left[r_{ij}^{\varepsilon}(T_l(u),T_l(v)) - r_{ij}^{\varepsilon}(T_l(v),T_l(u))\right]\partial^2_{y_iy_j}\phi.
\end{aligned}
$$

Clearly, we have $\lim_{\varepsilon\downarrow 0} E_3(\varepsilon) = 0$ and

$$
\lim_{\varepsilon\downarrow 0} E_2(\varepsilon) = \int 2\sum_{i,j=1}^{d} r_{ij}(T_l(u),T_l(v))\partial^2_{x_iy_j}\phi\,dx\,dt\,dy\,ds. \tag{3.13}
$$

Our goal now is to show that

$$
\lim_{\varepsilon\downarrow 0} E_1(\varepsilon) + \lim_{\varepsilon\downarrow 0} E_2(\varepsilon) \geq 0. \tag{3.14}
$$

To this end, note first that, since $\operatorname{sign}'_\varepsilon(\cdot) \geq 0$,

$$
I_1(\varepsilon) \geq 2\sum_{k=1}^{K} \operatorname{sign}'_\varepsilon\left(T_l(u) - T_l(v)\right)\operatorname{div}_x\zeta_k(T_l(u))\operatorname{div}_y\zeta_k(T_l(v))\phi,
$$

so that

$$
\begin{aligned}
E_1(\varepsilon) \geq \int 2\sum_{k=1}^{K} &\operatorname{sign}'_\varepsilon\left(T_l(u) - T_l(v)\right)\operatorname{div}_x\zeta_k(T_l(u))\\
&\times \operatorname{div}_y\zeta_k(T_l(v))\phi\,dx\,dt\,dy\,ds.
\end{aligned}
$$

Invoking the chain rule (D.3) in Definition 2.3 (we can do this since $\operatorname{sign}'_\varepsilon(\cdot)$ belongs to $C(\mathbf{R})$), we have for $1 \leq k \leq K$

$$
\operatorname{sign}'_\varepsilon\left(T_l(u) - T_l(v)\right)\operatorname{div}_y\zeta_k(T_l(v)) = \operatorname{div}_y\zeta_k^{\operatorname{sign}'_\varepsilon(T_l(u)-\cdot)}(T_l(v)). \tag{3.15}
$$

If we now use (3.15), then we have

$$
E_1(\varepsilon) \geq \int 2\sum_{k=1}^{K} \operatorname{div}_x\zeta_k(T_l(u))\operatorname{div}_y\zeta_k^{\operatorname{sign}'_\varepsilon(T_l(u)-\cdot)}(T_l(v))\phi\,dx\,dt\,dy\,ds.
$$

Integration by parts in y yields

$$\begin{aligned}
E_1(\varepsilon) \geq \int 2 \sum_{k=1}^{K} \sum_{i,j=1}^{d} & \partial_{x_i} \zeta_{ik}(T_l(u)) \\
& \times \partial_{y_j} \left(\int_{T_l(u)}^{T_l(v)} \operatorname{sign}'_\varepsilon (T_l(u) - \xi)\, \sigma_{jk}(\xi)\, d\xi \right) \phi \, dx\, dt\, dy\, ds \\
= - \int 2 \sum_{k=1}^{K} \sum_{i,j=1}^{d} & \partial_{x_i} \zeta_{ik}(T_l(u)) \\
& \times \left(\int_{T_l(u)}^{T_l(v)} \operatorname{sign}'_\varepsilon (T_l(u) - \xi)\, \sigma_{jk}(\xi)\, d\xi \right) \partial_{y_j} \phi \, dx\, dt\, dy\, ds.
\end{aligned}$$

For $1 \leq k \leq K$ and $1 \leq j \leq d$, define the function $\psi^\varepsilon_{jk} : \mathbf{R} \to \mathbf{R}$ by

$$\psi^\varepsilon_{jk}(\eta) = \int_{\eta}^{T_l(v)} \operatorname{sign}'_\varepsilon (\eta - \xi)\, \sigma_{jk}(\xi)\, d\xi.$$

Since $\operatorname{sign}'_\varepsilon(\cdot) \in C(\mathbf{R})$ and $\sigma_{jk}(\cdot) \in L^\infty_{\mathrm{loc}}(\mathbf{R})$, we have $\psi^\varepsilon_{jk}(\cdot) \in C(\mathbf{R})$ and the chain rule can therefore be used.

Using the chain rule (D.3) in Definition 2.3 and then doing integration by parts in x, we derive

$$\begin{aligned}
E_1(\varepsilon) &\geq - \int 2 \sum_{k=1}^{K} \sum_{i,j=1}^{d} \partial_{x_i} \zeta_{ik}(T_l(u)) \psi^\varepsilon_{jk}(T_l(u)) \partial_{y_j} \phi \, dx\, dt\, dy\, ds \\
&= - \int 2 \sum_{k=1}^{K} \sum_{i,j=1}^{d} \partial_{x_i} \zeta_{ik}^{\psi^\varepsilon_{jk}}(T_l(u)) \partial_{y_j} \phi \, dx\, dt\, dy\, ds \\
&= - \int 2 \sum_{k=1}^{K} \sum_{i,j=1}^{d} \partial_{x_i} \left(\int_{T_l(v)}^{T_l(u)} \psi^\varepsilon_{jk}(\xi) \sigma_{ik}(\xi)\, d\xi \right) \partial_{y_j} \phi \, dx\, dt\, dy\, ds \\
&= \int 2 \sum_{k=1}^{K} \sum_{i,j=1}^{d} \left(\int_{T_l(v)}^{T_l(u)} \psi^\varepsilon_{jk}(\xi) \sigma_{ik}(\xi)\, d\xi \right) \partial^2_{x_i y_j} \phi \, dx\, dt\, dy\, ds.
\end{aligned}$$

Observe that for a.e. $\eta \in \mathbf{R}$

$$\lim_{\varepsilon \downarrow 0} \psi^\varepsilon_{jk}(\eta) = -\operatorname{sign}(\eta - T_l(v))\, \sigma_{jk}(\eta),$$

so that by the dominated convergence theorem we have for a.e. (t, x, s, y)

$$\lim_{\varepsilon \downarrow 0} \int_{T_l(v)}^{T_l(u)} \psi^\varepsilon_{jk}(\xi) \sigma_{ik}(\xi)\, d\xi = - \int_{T_l(v)}^{T_l(u)} \operatorname{sign}(\xi - T_l(v))\, \sigma_{jk}(\xi) \sigma_{ik}(\xi)\, d\xi.$$

Hence, after another application of the dominated convergence theorem,

$$\begin{aligned}\lim_{\varepsilon\downarrow 0} E_1(\varepsilon) \geq -\int 2\sum_{k=1}^{K}\sum_{i,j=1}^{d}\left(\int_{T_l(v)}^{T_l(u)} \operatorname{sign}\left(\xi - T_l(v)\right)\sigma_{jk}(\xi)\sigma_{ik}(\xi)\,d\xi\right)\\ \times \partial^2_{x_i y_j}\phi\,dx\,dt\,dy\,ds\\ = -\int 2\sum_{i,j=1}^{d} r_{ij}(T_l(u),T_l(v))\partial^2_{x_i y_j}\phi\,dx\,dt\,dy\,ds.\end{aligned} \tag{3.16}$$

Finally, adding (3.16) to (3.13) yields (3.14).

Summing up, sending $\varepsilon \downarrow 0$ in (3.12) gives

$$\begin{aligned}&\int \Big(I_{\text{time}} + I_{\text{conv}} + I_{\text{diff}}\Big)(t,x,s,y)\omega_\rho\left(\frac{t-s}{2},\frac{x-y}{2}\right)dx\,dt\,dy\,ds\\ &\quad -\int \operatorname{sign}\left(T_l(u)-T_l(v)\right)\left(F(t,x)-G(s,y)\right)\\ &\qquad \times \omega_\rho\left(\frac{t-s}{2},\frac{x-y}{2}\right)\varphi\left(\frac{t+s}{2},\frac{x+y}{2}\right)dx\,dt\,dyds\\ &\quad \geq -\mu_l^u((0,T)\times\mathbf{R}^d) - \mu_l^v((0,T)\times\mathbf{R}^d),\end{aligned} \tag{3.17}$$

where

$$I_{\text{time}}(t,x,s,y) = \left|T_l(u(t,x)) - T_l(v(s,y))\right|\left(\partial_t+\partial_s\right)\varphi\left(\frac{t+s}{2},\frac{x+y}{2}\right),$$

$$I_{\text{conv}}(t,x,s,y) = \sum_{i=1}^{d} q_i(T_l(u(t,x)),T_l(v(s,y)))\left(\partial_{x_i}+\partial_{y_i}\right)\varphi\left(\frac{t+s}{2},\frac{x+y}{2}\right),$$

$$I_{\text{diff}}(t,x,s,y) = \sum_{i,j=1}^{d} r_{ij}(T_l(u),T_l(v))\left(\partial^2_{x_i x_j} + 2\partial^2_{x_i y_j} + \partial^2_{y_i y_j}\right)\varphi\left(\frac{t+s}{2},\frac{x+y}{2}\right).$$

Let us introduce the change of variables

$$\tilde{x} = \frac{x+y}{2}, \quad \tilde{t} = \frac{t+s}{2}, \quad z = \frac{x-y}{2}, \quad \tau = \frac{t-s}{2},$$

which maps $(0,T)\times\mathbf{R}^d\times(0,T)\times\mathbf{R}^d$ into

$$\Omega = \mathbf{R}^d\times\mathbf{R}^d\times\left\{(\tilde{t},\tau)\;\middle|\; 0\leq\tilde{t}+\tau\leq T,\quad 0\leq\tilde{t}-\tau\leq T\right\}.$$

Observe that

$$\left(\partial_t+\partial_s\right)\varphi\left(\frac{t+s}{2},\frac{x+y}{2}\right) = \varphi_{\tilde{t}}(\tilde{t},\tilde{x}), \qquad \left(\nabla_x+\nabla_y\right)\varphi(t,x,s,y) = \nabla_{\tilde{x}}\varphi(\tilde{t},\tilde{x}).$$

This change of variables diagonalizes also the operator $\nabla_{xx}+2\nabla_{xy}+\nabla_{yy}$:

$$\left(\nabla_{xx}+2\nabla_{xy}+\nabla_{yy}\right)\varphi\left(\frac{t+s}{2},\frac{x+y}{2}\right) = \nabla_{\tilde{x}\tilde{x}}\varphi(\tilde{t},\tilde{x}).$$

Keeping in mind that

$$x = \tilde{x} + z, \quad y = \tilde{x} - z, \quad t = \tilde{t} + \tau, \quad s = \tilde{t} - \tau,$$

we may now estimate (3.17) as

$$\begin{aligned} &\int_\Omega \Big(I_{\text{time}} + I_{\text{conv}} - I_{\text{diff}} \Big)(\tilde{t}, \tilde{x}, \tau, z) \omega_\rho(\tau, z)\, d\tilde{t}\, d\tilde{x}\, d\tau\, dz \\ &\qquad \geq - \int_\Omega \big| F(\tilde{t} + \tau, \tilde{x} + z) - G(\tilde{t} - \tau, \tilde{x} - z) \big| \, \omega_\rho(\tau, z)\, d\tilde{x}\, d\tilde{t}\, d\tau\, dz - o(1/l), \end{aligned} \tag{3.18}$$

where

$$\begin{aligned} I_{\text{time}}(\tilde{t}, \tilde{x}, \tau, z) &= \big| T_l(u(\tilde{t} + \tau, \tilde{x} + z)) - T_l(v(\tilde{t} - \tau, \tilde{x} - z)) \big| \, \varphi_{\tilde{t}}(\tilde{t}, \tilde{x}), \\ I_{\text{conv}}(\tilde{t}, \tilde{x}, \tau, z) &= \sum_{i=1}^d q_i \left(T_l(u(\tilde{t} + \tau, \tilde{x} + z)), T_l(v(\tilde{t} - \tau, \tilde{x} - z)) \right) \partial_{\tilde{x}_i} \varphi(\tilde{t}, \tilde{x}), \\ I_{\text{diff}}(\tilde{t}, \tilde{x}, \tau, z) &= \sum_{i,j=1}^d r_{ij} \left(T_l(u(\tilde{t} + \tau, \tilde{x} + z)), T_l(v(\tilde{t} - \tau, \tilde{x} - z)) \right) \partial^2_{\tilde{x}_i \tilde{x}_j} \varphi. \end{aligned}$$

Sending $\rho \downarrow 0$ in (3.18) yields

$$\begin{aligned} &\int \Bigg(|T_l(u) - T_l(v)| \, \partial_t \varphi + \sum_{i=1}^d q_i \left(T_l(u), T_l(v) \right) \partial_{x_i} \varphi \\ &\qquad + \sum_{i,j=1}^d r_{ij} \left(T_l(u), T_l(v) \right) \partial^2_{x_i x_j} \varphi \Bigg) \, dx\, dt \\ &\qquad\qquad \geq - \int |F(t,x) - G(t,x)| \, \varphi \, dx\, dt - o(1/l). \end{aligned} \tag{3.19}$$

By standard arguments (choosing a sequence of functions $0 \leq \varphi \leq 1$ from $\mathcal{D}((0,T) \times \mathbf{R}^d)$ that converges to $\mathbf{1}_{(0,t) \times \mathbf{R}^d}$ and using the initial conditions for u, v in the sense of, say, (D.6) in Definition 2.3, it follows from (3.19) that for a.e. $t \in (0,T)$

$$\begin{aligned} &\int_{\mathbf{R}^d} |T_l(u(t,x)) - T_l(v(t,x))| \, dx \\ &\qquad \leq \int_{\mathbf{R}^d} |T_l(u_0) - T_l(v_0)| \, dx + \int_0^t \int_{\mathbf{R}^d} |F(s,x) - G(s,x)| \, dx\, ds + o(1/l). \end{aligned} \tag{3.20}$$

Equipped with (D.5) in Definition 2.3 for u and v, sending $l \uparrow \infty$ in (3.20) finally yields the L^1 stability property (3.4). ∎

4. Existence of renormalized entropy solution. The purpose of this section is to prove the following theorem.

THEOREM 4.1 (existence). *Suppose that* (1.2), (1.3), *and* (1.4) *hold. Then there exists at least one renormalized entropy solution u of* (1.1).

We divide the proof into two steps.

Step 1 (bounded data). Suppose the data u_0 and F are bounded and integrable functions. Repeating the proof in [9] we find that there exists a unique entropy solution u to (1.1) (interpreted in the sense of Definition 2.2), and this entropy solution can

be constructed by the vanishing viscosity method [21]. For us it remains to prove that this entropy solution is also a renormalized entropy solution in the sense of Definition 2.3. To this end, let u_ρ be the unique classical (say $C^{1,2}$) solution to the uniformly parabolic problem

$$\begin{aligned} &\partial_t u_\rho + \operatorname{div} f(u_\rho) = \nabla \cdot (a(u_\rho)\nabla u_\rho) + \rho \Delta u_\rho + F, \qquad \rho > 0, \\ &u_\rho(x,0) = u_0(x). \end{aligned} \tag{4.1}$$

Equipped with the a priori estimates in [21], Chen and Perthame [9] proved

$$u_\rho \to u \quad \text{a.e. and in } C(0,T;L^1(\mathbf{R}^d)) \text{ as } \rho \downarrow 0, \tag{4.2}$$

where u is the unique entropy solution to (1.1).

For any C^2 function S and $\left(q_i^S\right)' = S' f_i'$, $\left(r_{ij}^S\right)' = S' a_{ij}$ for $1 \le i,j \le d$, multiplying the equation in (4.1) by $S'(u_\rho)$ yields

$$\begin{aligned} &\partial_t S(u_\rho) + \sum_{i=1}^d \partial_{x_i} q_i^S(u_\rho) - \sum_{i,j=1}^d \partial^2_{x_i x_j} r_{ij}^S(u_\rho) - \rho \Delta S(u_\rho) \\ &\quad - S'(u_\rho) F(u_\rho) = -\left(n_\rho^{S''} + m_\rho^{S''}\right)(t,x), \end{aligned} \tag{4.3}$$

where the parabolic dissipation measure $n_{n,\rho}^{S''}(t,x)$ is defined by

$$n_\rho^{S''}(t,x) = \sum_{k=1}^K \left(\sum_{i=1}^d \partial_{x_i} \zeta_{ik}^{S''}(u(t,x)) \right)^2,$$

and the entropy dissipation measure $m_\rho^{S''}(t,x)$ is defined by

$$m_\rho^{S''}(t,x) = \rho S''(u_\rho) \left|\nabla u_\rho\right|^2.$$

An easy approximation argument reveals that (4.3) continues to hold for any function $S \in W^{2,\infty}(\mathbf{R})$.

Inserting $S(u) = \frac{1}{l}\int_0^u T_l(\xi)\xi$ into (4.3) and then sending $l \downarrow 0$, we get the well-known estimate

$$\|u_\rho\|_{L^\infty(0,T;L^1(\mathbf{R}^d))} \le \|u_0\|_{L^1(\mathbf{R}^d)} + \|F\|_{L^1((0,T)\times\mathbf{R}^d)}. \tag{4.4}$$

We need to derive some additional a priori estimates (involving (2.4)) that are independent of ρ and $\|u_0\|_{L^\infty(\mathbf{R}^d)}$, $\|F\|_{L^\infty((0,T)\times\mathbf{R}^d)}$.

LEMMA 4.1. *For any $l > 0$, we have*

$$\int_{(0,T)\times\mathbf{R}^d} \left(\sum_{k=1}^K \left(\sum_{i=1}^d \partial_{x_i} \zeta_{ik}(T_l(u_\rho)) \right)^2 + \rho \left|\nabla T_l(u_\rho)\right|^2 \right) dx\, dt \le C_l$$

for some constant C_l that is independent of ρ but not l. More precisely,

$$C_l = l\left(\|u_0\|_{L^1(\mathbf{R}^d)} + \|F\|_{L^1((0,T)\times\mathbf{R}^d)} \right).$$

Proof. Introduce the function

$$S(u) = \int_0^u T_l(\xi)\, d\xi = \begin{cases} \frac{|u|^2}{2}, & |u| \le l, \\ l\,|u| - \frac{l^2}{2}, & |u| > l. \end{cases}$$

The lemma follows by choosing this $S(\cdot)$ in (4.3). ∎

LEMMA 4.2. *For any $l > 0$ and any $\delta > 0$,*

$$\frac{1}{\delta}\int_{\{l<|u_\rho|<l+\delta\}}\left(\sum_{k=1}^{K}\left(\sum_{i=1}^{d}\partial_{x_i}\zeta_{ik}(u_\rho)\right)^2+\rho\,|\nabla u_\rho|^2\right)dx\,dt\le\mathcal{E}(l) \tag{4.5}$$

for some bounded function $\mathcal{E}(\cdot)$ on $\mathbf{R}_+$ that is independent of ρ,δ and satisfies $\lim_{l\uparrow\infty}\mathcal{E}(l)=0$.

If the data u_0, F are bounded and $l > M := \|u_0\|_{L^\infty(\mathbf{R}^d)} + \|F\|_{L^\infty((0,T)\times\mathbf{R}^d)}$, then $\mathcal{E}(l) = 0$.

Proof. Let us define the function $S(\cdot)$ by $S(0) = 0$ and

$$S'(u)=\frac{1}{\delta}\left(T_{l+\delta}(u)-T_l(u)\right)=\begin{cases}-1, & u<-l-\delta,\\ \frac{-u-l}{\delta}, & -l-\delta<u<-l,\\ 0, & -l<u<l,\\ \frac{u-l}{\delta}, & l<u<l+\delta,\\ 1, & u>l+\delta.\end{cases}$$

Inserting this S into (4.3) gives

$$\begin{aligned}&\frac{1}{\delta}\int_{\{l<|u_\rho|<l+\delta\}}\left(\sum_{k=1}^{K}\left(\sum_{i=1}^{d}\partial_{x_i}\zeta_{ik}(u_\rho)\right)^2+\rho\,|\nabla u_\rho|^2\right)dx\,dt\\ &\quad\le\int_{\{|u_0|>l\}}|u_0|\,dx+\int_{\{|u_\rho|>l\}}|F|\,dx\,dt:=\mathcal{E}(l).\end{aligned} \tag{4.6}$$

Since $u_0\in L^1(\mathbf{R}^d)$, $F\in L^1((0,T)\times\mathbf{R}^d)$, and, thanks to (4.4), u_ρ is uniformly (in ρ) bounded in $L^1((0,T)\times\mathbf{R}^d)$, we have $\mathcal{E}(l)\to 0$ as $l\uparrow\infty$.

If the data u_0, F are bounded, then it is well known that $\|u_\rho\|_{L^\infty((0,T)\times\mathbf{R}^d)}\le M$, where M is defined in the lemma. We observe that if $l > M$, then $S(u_0) = 0$ and $S'(u_\rho) = 0$. Hence we deduce $\mathcal{E}(l) = 0$. ∎

Let us choose a particular $S = S_{\eta,h}$ in (4.3) of the form

$$\begin{aligned}&S_{\eta,h}(0)=0, && S'_{\eta,h}=\eta' h',\\ &\eta\in C^2(\mathbf{R}), && \eta''\ge 0, \qquad |\eta'|\le K,\\ &h\in C^2(\mathbf{R}), && \operatorname{supp}(h')\subset[-l,l].\end{aligned}$$

This gives

$$\begin{aligned}&\partial_t S_{\eta,h}(u_\rho)+\sum_{i=1}^{d}\partial_{x_i}q_i^{S_{\eta,h}}(u_\rho)-\sum_{i,j=1}^{d}\partial^2_{x_ix_j}r_{ij}^{S_{\eta,h}}(u_\rho)-\rho\Delta S_{\eta,h}(u_\rho)\\ &\quad-S'_{\eta,h}(u_\rho)F(u_\rho)=-\left(n_\rho^{\eta''h'}+\mu_\rho^{\eta'h''}\right)(t,x),\end{aligned} \tag{4.7}$$

where

$$\begin{aligned}\mu_\rho^{\eta'h''}(t,x)&:=-\left(n_\rho^{\eta'h''}+m_\rho^{\eta'h''}\right)(t,x)\\ &=-\eta'(u_\rho)h''(u_\rho)\left(\sum_{k=1}^{K}\left(\sum_{i=1}^{d}\partial_{x_i}\zeta_{ik}(u_\rho)\right)^2+\rho\,|\nabla u_\rho|^2\right).\end{aligned}$$

Let $h_{l,\delta} : \mathbf{R} \to \mathbf{R}$ denote the function defined by $h_{l,\delta}(0) = 0$ and

$$h'_{l,\delta}(u) = \begin{cases} 1, & |u| < l, \\ \frac{l+\delta-|u|}{\delta}, & l < |u| < l+\delta, \\ 0, & |u| > l+\delta. \end{cases}$$

Clearly,

$$h_{l,\delta}(u) \to T_l(u), \qquad h'_{l,\delta}(u) \to \mathbf{1}_{\{|u|<l\}} \tag{4.8}$$

for any $u \in \mathbf{R}$. The idea is to choose $h = h_{n,l}$ in (4.7) and then let $\delta \downarrow 0$. To this end, let us first define the Radon measure $\mu^K_{l,\rho,\delta}$ on $(0,T) \times \mathbf{R}^d$ by

$$d\mu^K_{l,\rho,\delta}(t,x) := \frac{K}{\delta} \mathbf{1}_{\{l<|u_\rho|<l+\delta\}} \left(\sum_{k=1}^K \left(\sum_{i=1}^d \partial_{x_i} \zeta_{ik}(u_\rho) \right)^2 + \rho \, |\nabla u_\rho|^2 \right) dx \, dt,$$

that is, for any Borel set $E \subset (0,T) \times \mathbf{R}^d$,

$$\mu^K_{l,\rho,\delta}(E) = \frac{K}{\delta} \int_{E \cap \{l<|u_\rho|<l+\delta\}} \left(\sum_{k=1}^K \left(\sum_{i=1}^d \partial_{x_i} \zeta_{ik}(u_\rho) \right)^2 + \rho \, |\nabla u_\rho|^2 \right) dx \, dt.$$

Then, by Lemma 4.2, $\mu^K_{l,\rho,\delta}((0,T) \times \mathbf{R}^d) \le \mathcal{E}(l)$. Consequently, we may assume that

$$\begin{aligned} &\mu^K_{l,\rho,\delta} \overset{\star}{\rightharpoonup} \mu^K_{l,\rho} \qquad \text{in the sense of measures on } (0,T) \times \mathbf{R}^d \text{ as } \delta \downarrow 0, \\ &\mu^K_{l,\rho} \overset{\star}{\rightharpoonup} \mu^K_l \qquad \text{in the sense of measures on } (0,T) \times \mathbf{R}^d \ \rho \downarrow 0 \end{aligned} \tag{4.9}$$

for some nonnegative bounded Radon measure μ^K_l satisfying

$$\mu^K_l((0,T) \times \mathbf{R}^d) \le \mathcal{E}(l) \to 0 \quad \text{as } l \uparrow \infty. \tag{4.10}$$

For any $0 \le \phi \in \mathcal{D}((0,T) \times \mathbf{R}^d)$, thanks to (4.8) and the convexity of η,

$$\begin{aligned} &\lim_{\delta \downarrow 0} \int_{(0,T)\times\mathbf{R}^d} n_\rho^{\eta'' h'_{l,\delta}}(t,x) \phi \, dx \, dt \\ &\qquad \ge \int_{(0,T)\times\mathbf{R}^d} \eta''(T_l(u_\rho)) \sum_{k=1}^K \left(\sum_{i=1}^d \partial_{x_i} \zeta_{ik}(T_l(u_\rho)) \right)^2 \phi \, dx \, dt. \end{aligned} \tag{4.11}$$

Again because of (4.8), it can be easily checked that as $\delta \downarrow 0$ (recall $q' = \eta' f'$ and $r' = \eta' a$)

$$\begin{aligned} &S_{\eta,h_{l,\delta}}(u) \to \eta(T_l(u)), \qquad S'_{\eta,h_{l,\delta}}(u) \to \eta'(T_l(u)), \\ &q^{S_{\eta,h_{l,\delta}}}(u) \to q(T_l(u)), \qquad r^{S_{\eta,h_{l,\delta}}}(u) \to r(T_l(u)) \end{aligned} \tag{4.12}$$

for any $u \in \mathbf{R}$.

Inserting $h = h_{l,\delta}$ into (4.7) and using $|\eta'| \le K$, (4.9), (4.11), (4.12) when sending $\delta \downarrow 0$, we get

$$\begin{aligned}
&\partial_t \eta(T_l(u_\rho)) + \sum_{i=1}^{d} \partial_{x_i} q_i(T_l(u_\rho)) - \sum_{i,j=1}^{d} \partial^2_{x_i x_j} r_{ij}(T_l(u_\rho)) \\
&\qquad - \rho \Delta \eta(T_l(u_\rho)) - \eta'(T_l(u_\rho)) F \\
&\le -\eta''(T_l(u_\rho)) \sum_{k=1}^{K} \left(\sum_{i=1}^{d} \partial_{x_i} \zeta_{ik}(u_\rho) \right)^2 + \mu^K_{l,\rho} \quad \text{in } \mathcal{D}'((0,T)\times \mathbf{R}^d).
\end{aligned} \tag{4.13}$$

Equipped with (4.9), passing to the limit $\rho \downarrow 0$ in (4.13) yields that u satisfies the entropy condition (2.5).

It remains to prove that the chain rule (D.3) in Definition 2.3 holds. For any $\psi \in C(\mathbf{R})$, the classical chain rule gives for $k = 1, \dots, K$

$$\sum_{i=1}^{d} \partial_{x_i} \zeta^{\psi}_{ik}(T_l(u_\rho)) = \psi(T_l(u_\rho)) \sum_{i=1}^{d} \partial_{x_i} \zeta_{ik}(T_l(u_\rho)) \qquad \forall l > 0.$$

As in [9], the proof is to observe that this equality continues to hold in the limit as $\rho \downarrow 0$ since u_ρ converges strongly and $\sum_{i=1}^{d} \partial_{x_i} \zeta_{ik}(T_l(u_\rho))$ weakly.

Step 2 (unbounded data). Suppose the data u_0 and F satisfy (1.2). For $n > 1$, introduce the truncated data $u_{0,n} = T_n(u_0)$ and $F_n = T_n(F)$. We have $u_{0,n} \to u_0$, $F_n \to F$ in L^1 as $n \uparrow \infty$. Thanks to the L^1 contraction property of the solution operator to (4.1), $\{u_n\}_{n>1}$ is a Cauchy sequence in $C(0,T;L^1(\mathbf{R}^d))$ and has a limit point u. From Step 1 we know that each u_n is a renormalized entropy solution of (1.1) with u_0 and F replaced by $u_{0,n}$ and F_n, respectively. Denote by $\mu^K_{l,n}$ the corresponding renormalization measure. Lemma 4.1 and (4.10) imply that the following n-independent a priori estimates hold for each $l > 0$:

$$\begin{aligned}
&\|u_n\|_{L^\infty(0,T;L^1(\mathbf{R}^d))} \le \|u_0\|_{L^1(\mathbf{R}^d)} + \|F\|_{L^1((0,T)\times\mathbf{R}^d)}, \\
&\sum_{k=1}^{K} \Big(\mathrm{div} \zeta_k(T_l(u_n)) \Big)^2 = \sum_{k=1}^{K} \left(\sum_{i=1}^{d} \partial_{x_i} \zeta_{ik}(T_l(u_n)) \right)^2 \le C_l, \\
&\mu^K_{l,n}((0,T)\times\mathbf{R}^d) \le \int_{\{|u_0|>l\}} |u_0| \, dx + \int_{\{|u_n|>l\}} |F| \, dx\, dt
\end{aligned}$$

for some constant C_l depending on l but not n. Equipped with these estimates and the strong convergence $u_n \to u$, we can repeat the steps in the above limiting process for the viscous approximations $\{u_\rho\}_{\rho>0}$ and prove that the limit point u of $\{u_n\}_{n>1}$ is a renormalized entropy solution of (1.1), with the renormalization measure $\mu^{u,K}_l$ being a limit point of $\{\mu^K_{l,n}\}_{n>1}$. This completes the proof of Theorem 4.1.

Acknowledgment. This work was done while MB visited the Centre of Mathematics for Applications (CMA) at the University of Oslo, Norway, and he is grateful for the hospitality.

REFERENCES

[1] P. Bénilan, J. Carrillo, and P. Wittbold, *Renormalized entropy solutions of scalar conservation laws*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 29 (2000), pp. 313–327.

[2] P. Bénilan and H. Touré, *Sur l'équation générale $u_t = a(\cdot, u, \phi(\cdot, u)_x)_x + v$ dans L^1. II. Le problème d'évolution*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 12 (1995), pp. 727–761.

[3] D. Blanchard, F. Murat, and H. Redwane, *Existence and uniqueness of a renormalized solution for a fairly general class of nonlinear parabolic problems*, J. Differential Equations, 177 (2001), pp. 331–374.

[4] R. Bürger, S. Evje, and K. H. Karlsen, *On strongly degenerate convection-diffusion problems modeling sedimentation-consolidation processes*, J. Math. Anal. Appl., 247 (2000), pp. 517–556.

[5] J. Carrillo, *On the uniqueness of the solution of the evolution dam problem*, Nonlinear Anal., 22 (1994), pp. 573–607.

[6] J. Carrillo, *Entropy solutions for nonlinear degenerate problems*, Arch. Ration. Mech. Anal., 147 (1999), pp. 269–361.

[7] G.-Q. Chen and E. DiBenedetto, *Stability of entropy solutions to the Cauchy problem for a class of nonlinear hyperbolic-parabolic equations*, SIAM J. Math. Anal., 33 (2001), pp. 751–762.

[8] G.-Q. Chen and K. H. Karlsen, *L^1 framework for continuous dependence and error estimates for quasi-linear degenerate parabolic equations*, Trans. Amer. Math. Soc., to appear.

[9] G.-Q. Chen and B. Perthame, *Well-posedness for non-isotropic degenerate parabolic-hyperbolic equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 20 (2003), pp. 645–668.

[10] B. Cockburn and G. Gripenberg, *Continuous dependence on the nonlinearities of solutions of degenerate parabolic equations*, J. Differential Equations, 151 (1999), pp. 231–251.

[11] R. J. DiPerna and P.-L. Lions, *On the Cauchy problem for Boltzmann equations: Global existence and weak stability*, Ann. of Math. (2), 130 (1989), pp. 321–366.

[12] R. Eymard, T. Gallouët, and R. Herbin, *Existence and uniqueness of the entropy solution to a nonlinear hyperbolic equation*, Chinese Ann. Math. Ser. B, 16 (1995), pp. 1–14.

[13] R. Eymard, T. Gallouët, R. Herbin, and A. Michel, *Convergence of a finite volume scheme for nonlinear degenerate parabolic equations*, Numer. Math., 92 (2002), pp. 41–82.

[14] K. H. Karlsen and M. Ohlberger, *A note on the uniqueness of entropy solutions of nonlinear degenerate parabolic equations*, J. Math. Anal. Appl., 275 (2002), pp. 439–458.

[15] K. H. Karlsen and N. H. Risebro, *On the uniqueness and stability of entropy solutions of nonlinear degenerate parabolic equations with rough coefficients*, Discrete Contin. Dyn. Syst., 9 (2003), pp. 1081–1104.

[16] S. N. Kružkov, *First order quasi-linear equations in several independent variables*, Mat. Sb. (N.S.), 10 (1970), pp. 217–243.

[17] C. Mascia, A. Porretta, and A. Terracina, *Nonhomogeneous Dirichlet problems for degenerate parabolic-hyperbolic equations*, Arch. Ration. Mech. Anal., 163 (2002), pp. 87–124.

[18] A. Michel and J. Vovelle, *Entropy formulation for parabolic degenerate equations with general Dirichlet boundary conditions and application to the convergence of FV methods*, SIAM J. Numer. Anal., 41 (2003), pp. 2262–2293.

[19] É. Rouvre and G. Gagneux, *Solution forte entropique de lois scalaires hyperboliques-paraboliques dégénérées*, C. R. Acad. Sci. Paris Sér. I Math., 329 (1999), pp. 599–602.

[20] A. I. Vol'pert, *The spaces BV and quasi-linear equations*, Mat. Sb. (N.S.), 2 (1967), pp. 225–267.

[21] A. I. Vol'pert and S. I. Hudjaev, *Cauchy's problem for degenerate second order quasilinear parabolic equations*, Mat. Sb. (N.S.), 7 (1969), pp. 365–387.

[22] Z. Wu and J. Yin, *Some properties of functions in BV_x and their applications to the uniqueness of solutions for degenerate quasilinear parabolic equations*, Northeastern Math. J., 5 (1989), pp. 395–422.

SIAM J. MATH. ANAL.
Vol. 36, No. 2, pp. 423–440

Γ-CONVERGENCE THROUGH YOUNG MEASURES*

PABLO PEDREGAL†

Abstract. We present a general framework to treat Γ-convergence of functionals through Young measures and through slicing decomposition. After dealing with a general situation where functionals are defined in Lebesgue spaces, we concentrate on the gradient case. Explicit computations are possible, in this case, when the sequence of functions determining the functionals has a special property. We illustrate the method by examining several well-known situations. More than these examples, emphasis in this paper is placed on the method itself and its generality.

Key words. slicing decomposition, joint Young measure, lower semicontinuity

AMS subject classifications. 49J45, 74Q05

DOI. 10.1137/S0036141003425696

1. Introduction. Γ-convergence of functionals is an important method for dealing with sequences of functionals and for understanding their limiting properties. It was originally introduced in the pioneering works [9], [10]. See [8] for a formal and more complete analysis. It is closely related to homogenization, G- and H-convergence (see [6], [11], [15], [16]), as well as variational problems and techniques (see [3], [7]). Many examples and applications are scattered throughout the literature, but [1] is a nice account of the application of all these ideas to optimal design and shape optimization. See also [5]. In this work, we would like to start a systematic treatment of Γ-convergence through the study of the underlying Young measures associated to relevant sequences. A major tool is the slicing measure decomposition (see [2], [12]).

To explain our perspective, suppose we have a sequence of integral functionals

$$I_j(u) = \int_\Omega W(a_j(x), u(x))\, dx, \quad u \in \mathcal{A},$$

where $\Omega \subset \mathbf{R}^N$ is a bounded, regular domain, $a_j : \Omega \to \mathbf{R}^m$, and $\mathcal{A}$ is some weakly closed subset of a certain reflexive Lebesgue space. The integrand

$$W(\lambda, \rho) : \mathbf{R}^m \times \mathbf{R}^d \to \mathbf{R}$$

is assumed to be continuous to begin with. Under coercivity for W with respect to ρ, the Γ-limit of the sequence of functional $\{I_j\}$ is defined by putting

$$I(u) = \inf \left\{ \liminf_{j\to\infty} I_j(u_j) : u_j \rightharpoonup u \right\}, \tag{1.1}$$

and it represents the right notion of variational convergence of functionals. A prime objective in many situations is to provide an explicit, integral form for $I(u)$, and doing so amounts to describing how the new integrand can be determined and computed from the sequence $\{a_j\}$. The procedure, carried out in many cases which can be

*Received by the editors March 28, 2003; accepted for publication (in revised form) September 26, 2003; published electronically July 29, 2004. This work is supported by research projects BFM2001-0738 of the MCyT and GC-02-001 of Castilla-La Mancha (Spain).
http://www.siam.org/journals/sima/36-2/42569.html
†ETSI Industriales, Universidad de Castilla-La Mancha, 13071, Ciudad Real, Spain (pablo.pedregal@uclm.es).

found in the references given above, essentially consists in finding a lower bound in (1.1), in the form of an integral functional, set up in such a way that will eventually become the Γ-limit, because that lower bound will in fact be an equality for a certain, cleverly chosen sequence $u_j \rightharpoonup u$ for any u. We would like to describe how this whole procedure can be done in a more-or-less systematic way through Young measures.

The key idea is to work with the joint Young measure corresponding to pairs $\{(a_j, u_j)\}$, where $u_j \rightharpoonup u$. But since $\{a_j\}$ is given and cannot be changed in any way, by means of the slicing decomposition we keep the Young measure associated with $\{a_j\}$ and work with the part of the joint measure coming from $\{u_j\}$. Notice that the joint Young measure will not, in general, be a product measure, and therefore a main issue is to understand the connection and relationship between their respective Young and joint Young measures.

Let

$$W(\lambda, \rho) : \mathbf{R}^m \times \mathbf{R}^d \to \mathbf{R}$$

be a continuous integrand such that

$$\begin{aligned} c\left(|\rho|^p - 1\right) \leq W(a_j(x), \rho) \leq C\left(|\rho|^p + 1\right), \\ |W(\lambda_1, \rho) - W(\lambda_2, \rho)| \leq w(|\lambda_1 - \lambda_2|)\, |\rho|^p \end{aligned}$$

for some $C > c > 0$, $p > 1$, a.e. $x \in \Omega$ and all j. The function w is continuous and $w(0) = 0$. Notice that an explicit x dependence of W can be incorporated into a_j so that, without loss of generality, we can assume no explicit dependence on x.

Let $\{a_j\}$ be weakly convergent in $L^q(\Omega)$ for some $q > 1$ and let $\sigma = \{\sigma_x\}_{x\in\Omega}$ be its underlying Young measure. Define the integrand

$$\psi(x, \rho) : \Omega \times \mathbf{R}^d \to \mathbf{R}$$

by putting

$$\psi(x, \rho) = \min_{\varphi} \left\{ \int_{\mathbf{R}^m} CW(x, \varphi(\lambda))\, d\sigma_x(\lambda) : \rho = \int_{\mathbf{R}^m} \varphi(\lambda)\, d\sigma_x(\lambda) \right\}.$$

We will show that ψ is well-defined in the sense that this infimum is indeed a minimum. CW is the convexification of W with respect to ρ. Notice that ψ is defined through a variational problem with respect to the Young measure corresponding to the sequence $\{a_j\}$, which determines the sequence of functionals $\{I_j\}$.

THEOREM 1.1. *Under the above hypotheses, the* Γ*-limit of* $\{I_j\}$ *is given by*

$$I(u) = \int_\Omega \psi(x, u(x))\, dx.$$

In section 3 we include various typical examples where one can explicitly calculate the density ψ by exploiting optimality conditions.

The situation where functionals depend on gradients is much more interesting,

$$I_j(u) = \int_\Omega W(a_j(x), \nabla u(x))\, dx, \quad u \in \mathcal{A},$$

and this time $\mathcal{A}$ is a weakly closed subset of a certain reflexive Sobolev space. A result similar to the previous one is valid only under a main, additional, structural

assumption on the sequence $\{a_j\}$. This property, called the "average gradient property" (AGP), roughly says that averages of gradients over level sets of a_j are gradients themselves. A more rigorous treatment can be found in section 4. The integrand W is assumed to enjoy the same properties as above. Suppose that $\{a_j\}$ is weakly convergent in $L^q(\Omega)$ for some $q > 1$ and that it verifies AGP, and let $\sigma = \{\sigma_x\}_{x\in\Omega}$ be its underlying Young measure. Define the integrand

$$\psi(x, \rho) : \Omega \times \mathbf{R}^d \to \mathbf{R}$$

by putting

$$\psi(x,\rho) = \min_{\varphi} \left\{ \int_{\mathbf{R}^m} CW(x, \varphi(\lambda))\, d\sigma_x(\lambda) : \rho = \int_{\mathbf{R}^m} \varphi(\lambda)\, d\sigma_x(\lambda), \right.$$
$$\left. \varphi(a_j(y)) \text{ is a gradient in } y \right\}.$$

A more precise definition of ψ is discussed later (section 4).

THEOREM 1.2. *If the sequence $\{a_j\}$ verifies AGP, the Γ-limit of $\{I_j\}$ is given by*

$$I(u) = \int_\Omega \psi(x, \nabla u(x))\, dx.$$

Some typical examples are explored in section 5 to illustrate the method.

2. The case without derivatives. In this section we treat and prove our main result for the most simple situation where the functionals do not depend explicitly on derivatives so that

$$I_j(u) = \int_\Omega W(a_j(x), u(x))\, dx,$$

where u belongs to some weakly closed subset of a certain Sobolev space. Ω is assumed to be a bounded, regular domain. More explicitly our assumptions on

$$W(\lambda, \rho) : \mathbf{R}^m \times \mathbf{R}^d \to \mathbf{R}$$

and the sequence $\{a_j\}$ are as follows:

1. $\{a_j\}$ is a weakly convergent sequence in $L^q(\Omega)$ for some $q > 1$.
2. W is uniformly coercive in ρ for all j with an exponent $p > 1$ and is uniformly bounded from above by the same power so that

$$c\,(|\rho|^p - 1) \le W(a_j(x), \rho) \le C\,(|\rho|^p + 1)$$

for some $C > c > 0$, all j, and a.e. $x \in \Omega$. Under this hypothesis, every sequence $\{u_j\}$ such that $\{I_j(u_j)\}$ is bounded from above will be bounded in $L^p(\Omega)$, and thus, possibly for a subsequence, it will converge weakly to some $u \in L^p(\Omega)$.

3. W is uniformly continuous in λ as indicated earlier:

$$|W(\lambda_1, \rho) - W(\lambda_2, \rho)| \le w(|\lambda_1 - \lambda_2|)\, |\rho|^p\,,$$

where w is continuous and $w(0) = 0$.

We have assumed no explicit, inhomogeneous dependence of W on $x \in \Omega$ but have pointed out that there is no loss of generality. Our aim is a description of the Γ-limit

$$I(u) = \inf \left\{ \liminf_{j\to\infty} I_j(u_j) : u_j \rightharpoonup u \text{ in } L^p(\Omega) \right\}.$$

Let u be given in $L^p(\Omega)$, and let $\{u_j\}$ be such that $\{I_j(u_j)\}$ is a nonincreasing sequence of numbers and $u_j \rightharpoonup u$ in $L^p(\Omega)$. Let $\sigma = \{\sigma_x\}_{x\in\Omega}$ be the Young measure associated with $\{a_j\}$ and let $\nu = \{\nu_x\}_{x\in\Omega}$ be the Young measure corresponding to the pairs $\{(a_j, u_j)\}$. Notice that

$$\operatorname{supp}(\sigma_x) \subset \mathbf{R}^m, \quad \operatorname{supp}(\nu_x) \subset \mathbf{R}^m \times \mathbf{R}^d.$$

At this point we invoke the slicing measure decomposition or disintegration.

THEOREM 2.1 (see [2], [12]). *Let ν be a nonnegative, finite Radon measure on $\mathbf{R}^{n+m}$, and let σ be its canonical projection onto $\mathbf{R}^n$ ($\sigma(E) = \nu(E \times \mathbf{R}^m)$). For σ-a.e. $x \in \mathbf{R}^n$ there exists a probability measure μ_x on $\mathbf{R}^m$ such that*

1. *the map*

$$x \mapsto \int_{\mathbf{R}^m} f(x,y)\, d\mu_x(y)$$

is σ-measurable for every bounded, continuous f;

2. *for every bounded, continuous function f*

$$\int_{\mathbf{R}^{n+m}} f(x,y)\, d\nu(x,y) = \int_{\mathbf{R}^n} \left(\int_{\mathbf{R}^m} f(x,y)\, d\mu_x(y) \right) d\sigma(x).$$

An enlightened way of shortening the statement of this theorem is to write

$$\nu(x,y) = \mu_x(y) \otimes \sigma(x).$$

Hence, in our situation, we can write

$$\nu_x = \mu_{\lambda,x} \otimes \sigma_x \tag{2.1}$$

for a.e. $x \in \Omega$ and every $\lambda \in \operatorname{supp}(\sigma_x)$. Each $\mu_{\lambda,x}$ is a certain probability measure with support contained in $\mathbf{R}^d$.

On the other hand, the representation in terms of Young measures always yields something smaller.

THEOREM 2.2 (see [13]). *If $\{z_j\}$ is a sequence of measurable functions with associated Young measure $\nu = \{\nu_x\}_{x\in\Omega}$, then*

$$\liminf_{j\to\infty} \int_E \psi(x, z_j(x))\, dx \geq \int_E \int_{\mathbf{R}^m} \psi(x,\lambda)\, d\nu_x(\lambda)\, dx$$

for every Carathéodory function ψ, bounded from below, and every measurable subset $E \subset \Omega$.

By using this fact, we obtain

$$\begin{aligned} \lim_{j\to\infty} \int_\Omega W(a_j(x), u_j(x))\, dx &\geq \int_\Omega \int_{\mathbf{R}^m \times \mathbf{R}^d} W(\lambda,\rho)\, d\nu_x(\lambda,\rho)\, dx \\ &= \int_\Omega \int_{\mathbf{R}^m} \int_{\mathbf{R}^d} W(\lambda,\rho)\, d\mu_{\lambda,x}(\rho)\, d\sigma_x(\lambda)\, dx. \end{aligned}$$

We also have the constraint on the first moment

$$u(x) = \int_{\mathbf{R}^m \times \mathbf{R}^d} \rho \, d\nu_x(\lambda, \rho) = \int_{\mathbf{R}^m} \int_{\mathbf{R}^d} \rho \, d\mu_{\lambda,x}(\rho) \, d\sigma_x(\lambda).$$

If we set

$$\varphi(\lambda, x) = \int_{\mathbf{R}^d} \rho \, d\mu_{\lambda,x}(\rho), \tag{2.2}$$

so that

$$u(x) = \int_{\mathbf{R}^m} \varphi(\lambda, x) \, d\sigma_x(\lambda), \tag{2.3}$$

we can go further down in our lower estimate by putting

$$\int_\Omega \int_{\mathbf{R}^m} \int_{\mathbf{R}^d} W(\lambda, \rho) \, d\mu_{\lambda,x}(\rho) \, d\sigma_x(\lambda) \, dx \geq \int_\Omega \int_{\mathbf{R}^m} CW(\lambda, \varphi(\lambda, x)) \, d\sigma_x(\lambda) \, dx,$$

where CW indicates the convex hull of W with respect to ρ. We can therefore write

$$I(u) \geq \inf_\varphi \left\{ \int_\Omega \int_{\mathbf{R}^m} CW(\lambda, \varphi(\lambda, x)) \, d\sigma_x(\lambda) \, dx : \varphi \text{ verifies (2.2) and (2.3) for some admissible } \nu \text{ as in (2.1)} \right\}.$$

One main step is to isolate the key requirements on the vector fields φ in the previous infimum without any reference to the measures $\mu_{\lambda,x}$ in (2.1), bearing in mind that we will later have to produce a sequence converging weakly to u for which all these inequalities are indeed equalities. In this general context, there is essentially no structural condition on φ so that we go further down by estimating

$$I(u) \geq \inf_\varphi \left\{ \int_\Omega \int_{\mathbf{R}^m} CW(\lambda, \varphi(\lambda, x)) \, d\sigma_x(\lambda) \, dx : u(x) = \int_{\mathbf{R}^m} \varphi(\lambda, x) \, d\sigma_x(\lambda) \right\}.$$

Because of the local nature of the Young measure, we see that this lower bound for $I(u)$ is an integral functional. Indeed if we define the integrand

$$\psi(x, \rho) : \Omega \times \mathbf{R}^d \to \mathbf{R}$$

by putting

$$\psi(x, \rho) = \inf_\varphi \left\{ \int_{\mathbf{R}^m} CW(x, \varphi(\lambda)) \, d\sigma_x(\lambda) : \rho = \int_{\mathbf{R}^m} \varphi(\lambda) \, d\sigma_x(\lambda) \right\}, \tag{2.4}$$

then, by interchanging the inf operation with the integral over Ω,

$$I(u) \geq \int_\Omega \psi(x, u(x)) \, dx.$$

Our goal is to show that we have, in fact, equality

$$I(u) = \int_\Omega \psi(x, u(x)) \, dx.$$

LEMMA 2.3. *The infimum in* (2.4) *is always attained.*

Proof. Notice that this lemma amounts to showing that a variational principle of the type

$$\text{Minimize in } \varphi: \quad \int_{\mathbf{R}^m} F(\lambda, \varphi(\lambda))\, d\sigma(\lambda)$$

subject to

$$\rho = \int_{\mathbf{R}^m} \varphi(\lambda)\, d\sigma(\lambda)$$

always has optimal solutions where σ is a given probability measure supported in $\mathbf{R}^m$ and F is a continuous integrand that is convex and coercive in the second variable. One could invoke Lebesgue spaces with respect to measures different from the usual Lebesgue measures (see [2]). Instead, in this simple context, it is easier to use the same slicing measure technique.

Let $\{\varphi_j\}$ be minimizing, and let φ be its weak limit. Define a measure ν supported on $\mathbf{R}^m \times \mathbf{R}^d$ by putting

$$\langle G, \nu \rangle = \lim_{j\to\infty} \int_{\mathbf{R}^m} G(\lambda, \varphi_j(\lambda))\, d\sigma(\lambda) \tag{2.5}$$

for any continuous G. It is clear that the projection of ν over $\mathbf{R}^m$ is σ. By the slicing measure decomposition we claim that, by the convexity of F,

$$\begin{aligned}
\lim_{j\to\infty} \int_{\mathbf{R}^m} F(\lambda, \varphi_j(\lambda))\, d\sigma(\lambda) &= \int_{\mathbf{R}^m} \int_{\mathbf{R}^d} F(\lambda, \rho)\, d\mu^\lambda(\rho)\, d\sigma(\lambda) \\
&\geq \int_{\mathbf{R}^m} F\left(\lambda, \int_{\mathbf{R}^d} \rho\, d\mu^\lambda(\rho)\right) d\sigma(\lambda) \\
&= \int_{\mathbf{R}^m} F(\lambda, \varphi(\lambda))\, d\sigma(\lambda).
\end{aligned}$$

Notice that applying (2.5) to $G(\lambda, \rho) = g(\lambda)\rho$ for arbitrary g, we conclude that the first moment of μ^λ is precisely the weak limit φ. □

Proof of Theorem 1.1. Let u be given. For a.e. $x \in \Omega$, by Lemma 2.3, we find $\varphi_0(\lambda, x)$ such that

$$\begin{aligned}
\psi(x, u(x)) &= \int_{\mathbf{R}^m} CW(\lambda, \varphi_0(\lambda, x))\, d\sigma_x(\lambda), \\
u(x) &= \int_{\mathbf{R}^m} \varphi_0(\lambda, x)\, d\sigma_x(\lambda).
\end{aligned}$$

It is also true that for every pair $(\lambda, x) \in \mathbf{R}^m \times \Omega$ we can find a probability measure $\mu_{\lambda,x}$ supported in $\mathbf{R}^d$ such that

$$\begin{aligned}
CW(\lambda, \varphi_0(\lambda, x)) &= \int_{\mathbf{R}^d} W(\lambda, \rho)\, d\mu_{\lambda,x}(\rho), \\
\varphi_0(\lambda, x) &= \int_{\mathbf{R}^d} \rho\, d\mu_{\lambda,x}(\rho).
\end{aligned}$$

In fact, it is a direct consequence of Carathéodory's theorem (see [7]) that we can take

$$\mu_{\lambda,x} = \sum_{l=0}^{d} t_{\lambda,x,l}\delta_{z_{\lambda,x,l}}, \qquad t_{\lambda,x,l} \geq 0, \sum_{l=0}^{d} t_{\lambda,x,l} = 1$$

for certain vectors $z_{\lambda,x,l}$. We conclude that

$$\int_{\Omega} \psi(x,u(x))\,dx = \int_{\Omega}\int_{\mathbf{R}^m}\int_{\mathbf{R}^d} W(\lambda,\rho)\,d\mu_{\lambda,x}(\rho)\,d\sigma_x(\lambda)\,dx.$$

Take the family of probability measures

$$\nu_x = \mu_{\lambda,x} \otimes \sigma_x. \tag{2.6}$$

Our main task consists in finding a sequence $\{u_j\}$ so that $\nu = \{\nu_x\}_{x\in\Omega}$ is the Young measure associated with $\{(a_j,u_j)\}$ and

$$\lim_{j\to\infty}\int_{\Omega} W(a_j(x),u_j(x))\,dx = \int_{\Omega}\psi(x,u(x))\,dx, \tag{2.7}$$

as desired. Notation becomes cumbersome but the leading argument is quite transparent, we believe.

An important issue here is that we are not entitled to change the sequence $\{a_j\}$ so that it is the sequence $\{u_j\}$ that must adapt itself to $\{a_j\}$. We will be using the next lemma. It is modeled after Lemma 7.9 in [13], but this time for a general, positive Radon measure. Its proof is based on the Lebesgue differentiation theorem, which is also valid for Radon measures (see, for instance, Corollary 2.23 in [2]). $B(\lambda,r)$ is, as usual, the ball centered at λ and radius r.

LEMMA 2.4. *Let σ be a positive, Radon measure in an open set D in $\mathbf{R}^d$. Let $N \subset D$ be a subset of σ-null measure. For $r_j : D\setminus N \to \mathbf{R}^+$ and $\{f_i\} \subset L^1(D,\sigma)$, there exist a set of points $\{\lambda_k^{(j)}\} \subset D\setminus N$ and positive numbers $\{\epsilon_k^{(j)}\}$, $\epsilon_k^{(j)} \leq r_j(\lambda_k^{(j)})$, such that*

$$\left\{B(\lambda_k^{(j)},\epsilon_k^{(j)})\right\} \text{ are pairwise disjoint for each } j,$$

$$\sigma\left(D\setminus \cup_k B(\lambda_k^{(j)},\epsilon_k^{(j)})\right) = 0 \quad \text{for each } j,$$

$$\int_D \xi(\lambda) f_i(\lambda)\,d\sigma(\lambda) = \lim_{j\to\infty}\sum_k f_i(\lambda_k^{(j)})\int_{B(\lambda_k^{(j)},\epsilon_k^{(j)})}\xi(\lambda)\,d\sigma(\lambda)$$

for every i and every $\xi \in L^\infty(D,\sigma)$.

Choose a dense, countable family of functions $\{W_i(\lambda,\rho)\}_{i=1,2,\ldots}$ vanishing at infinity. Put $W_0 \equiv W$. Consider the countable family of functions $\left\{\overline{W}_i(x)\right\}_{i=0,1,\ldots}$ defined by

$$\overline{W}_i(x) = \int_{\mathbf{R}^m}\tilde{W}_i(\lambda,x)\,d\sigma_x(\lambda),$$

$$\tilde{W}_i(\lambda,x) = \int_{\mathbf{R}^d} W_i(\lambda,\rho)\,d\mu_{\lambda,x}(\rho).$$

Apply the preceding lemma to Ω, the Lebesgue measure, $r_j(x)=1$, $N=\emptyset$, and the family $\{\overline{W}_i\}$ to conclude that we can find $\{x_k^{(j)}\}$ and $\{\epsilon_k^{(j)}\}$ such that

$$\left\{B(x_k^{(j)},\epsilon_k^{(j)})\right\} \text{ are pairwise disjoint for each } j,$$

$$\left|\Omega\setminus \cup_k B(x_k^{(j)},\epsilon_k^{(j)})\right|=0 \quad \text{for each } j,$$

$$\int_\Omega \xi(x)\int_{\mathbf{R}^m}\int_{\mathbf{R}^d} W_i(\lambda,\rho)\,d\mu_{\lambda,x}(\rho)\,d\sigma_x(\lambda)\,dx=\lim_{j\to\infty}\sum_k \overline{W}_i(x_k^{(j)})\int_{B(x_k^{(j)},\epsilon_k^{(j)})}\xi(x)\,dx$$

for all $\xi\in L^\infty(\Omega)$.

Apply Lemma 2.4 once again, this time to the Radon measures $\sigma_k^{(j)}\equiv\sigma_{x_k^{(j)}}$, $r_s(\lambda)=\epsilon_k^{(j)}$, $N=\emptyset$ and the family $\{\tilde{W}_i(\cdot,x_k^{(j)})\}$. Conclude that there exists a collection of points $\{\lambda_{k,r}^{(j,s)}\}$ and $\delta_{k,r}^{(j,s)}<\epsilon_k^{(j)}$ such that

$$\left\{B(\lambda_{k,r}^{(j,s)},\delta_{k,r}^{(j,s)})\right\} \quad \text{are pairwise disjoint for each } j,\ k, \text{ and } s,$$

$$\sigma_k^{(j)}\left(\mathbf{R}^m\setminus\cup_r B(\lambda_{k,r}^{(j,s)},\delta_{k,r}^{(j,s)})\right)=0 \quad \text{for every } j,\ k, \text{ and } s,$$

$$\int_{\mathbf{R}^m}\tilde{W}_i(\lambda,x_k^{(j)})\,d\sigma_k^{(j)}(\lambda)=\lim_{s\to\infty}\sum_r \tilde{W}_i(\lambda_{k,r}^{(j,s)},x_k^{(j)})\,\sigma_k^{(j)}\left(B(\lambda_{k,r}^{(j,s)},\delta_{k,r}^{(j,s)})\right)$$

for all k, j and i.

By our previous remark about Carathéodory's theorem,

$$\mu_{\lambda_{k,r}^{(j,s)},x_k^{(j)}}=\sum_{l=0}^d t_{k,r,l}^{(j,s)}\delta_{z_{k,r,l}^{(j,s)}},\quad t_{k,r,l}^{(j,s)}\ge 0, \sum_{l=0}^d t_{k,r,l}^{(j,s)}=1.$$

Take any partition of the set

$$\Lambda_{k,r}^{(j,s)}=\left\{a_j\in B(\lambda_{k,r}^{(j,s)},\delta_{k,r}^{(j,s)})\right\}\cap B(x_k^{(j)},\epsilon_k^{(j)})$$

in $d+1$ disjoint subsets, $\Lambda_{k,r,l}^{(j,s)}$, of relative (Lebesgue) measures $t_{k,r,l}^{(j,s)}$, and define $u_{j,s}$ on these subsets as $z_{k,r,l}^{(j,s)}$, respectively. Notice that

$$\lim_{j\to\infty}\sup_k\lim_{s\to\infty}\sup_r\left(\sigma_k^{(j)}\left(B(\lambda_{k,r}^{(j,s)},\delta_{k,r}^{(j,s)})\right)-\frac{\left|\Lambda_{k,r}^{(j,s)}\right|}{\left|B(x_k^{(j)},\epsilon_k^{(j)})\right|}\right)=0.$$

Take $\{\xi_n\}$ to be a dense, countable subset of continuous functions including the function identically 1 over Ω. For any such continuous ξ_n and any j and s, we have that the integrals

$$R_{n,i,j,s}=\int_\Omega \xi_n(x)W_i(a_j(x),u_{j,s}(x))\,dx$$

can be decomposed as follows:

$$\begin{aligned}
R_{n,i,j,s} &= \sum_k \sum_r \int_{\Lambda_{k,r}^{(j,s)}} \xi_n(x) W_i(\lambda_{k,r}^{(j,s)}, u_{j,s}(x))\, dx + d_{n,i,j,s} \\
&= \sum_k \sum_r \sum_l \int_{\Lambda_{k,r,l}^{(j,s)}} \xi_n(x)\, dx\, W_i(\lambda_{k,r}^{(j,s)}, z_{k,r,l}^{(j,s)}) + d_{n,i,j,s} \\
&= \sum_k \sum_r \int_{\Lambda_{k,r}^{(j,s)}} \xi_n(x)\, dx\, \tilde{W}_i(\lambda_{k,r}^{(j,s)}, x_k^{(j)}) + d_{n,i,j,s} \\
&= \sum_k \int_{B(x_k^{(j)}, \epsilon_k^{(j)})} \xi_n(x)\, dx \sum_r \tilde{W}_i(\lambda_{k,r}^{(j,s)}, x_k^{(j)}) \sigma_k^{(j)} \left(B(\lambda_{k,r}^{(j,s)}, \delta_{k,r}^{(j,s)}) \right) + d_{n,i,j,s}.
\end{aligned}$$

$d_{n,i,j,s}$ designates different sequences of numbers such that

$$\lim_{j\to\infty} \lim_{s\to\infty} d_{n,i,j,s} = 0$$

for all i and n. By our previous choices through Lemma 2.4, we can suitably choose $s = s(j)$ and have that the Young measure corresponding to $\{(a_j, u_j)\}$, $u_j = u_{j,s(j)}$, is precisely ν. Furthermore, by redoing all the previous computations for $\xi \equiv 1$ and W (for this we need the uniform continuity with respect to λ), it is elementary to check that (2.7) holds. In particular $\{u_j\}$ is bounded in $L^p(\Omega)$. □

An interesting consequence of our theorem is the following.

COROLLARY 2.5. *The Γ-limit of the initial functionals I_j depends upon the sequence $\{a_j\}$ only through its underlying Young measure.*

3. Some examples. What is quite remarkable is that in specific examples the computation of the density in (2.4) for the Γ-limit can be explicitly calculated. We will look at two typical, nontrivial examples.

Notice, to begin with, that if the sequence $\{a_j\}$ converges strongly to a, then the variational problem (2.4) is trivial since $\sigma_x = \delta_{a(x)}$, and we get

$$\psi(x, u(x)) = CW(a(x), u(x)).$$

On the other hand, solving (2.4) and overlooking the dependence on x amounts to being able to find the optimal solutions of problems of the type

$$\text{Minimize in } \varphi: \quad \int_{\mathbf{R}^m} F(\lambda, \varphi(\lambda))\, d\sigma(\lambda)$$

subject to

$$\rho = \int_{\mathbf{R}^m} \varphi(\lambda)\, d\sigma(\lambda),$$

where σ is a given probability measure supported in $\mathbf{R}^m$ and $F : \mathbf{R}^m \times \mathbf{R}^d \to \mathbf{R}$ is coercive and convex in the second variable. Since we know that there are always optimal solutions, these can be found by examining optimality conditions in many cases. The form of these will obviously depend on the probability measure σ.

Let us take

$$I_j(u) = \int_\Omega a_j(x)\, |u(x)|^2\, dx,$$

where $a_j : \Omega \to \mathbf{R}$ and $u : \Omega \to \mathbf{R}^d$ and $a_j \geq \alpha > 0$. The family of probability measures appearing in (2.4) is precisely the Young measure associated with $\{a_j\}$. In this example we know that those are supported in $\mathbf{R}$, more precisely in $(\alpha, +\infty)$. By Corollary 2.5 we can specify $\sigma = \{\sigma_x\}$ instead of a_j. The simplest nontrivial example corresponds, overlooking the dependence on x, to

$$\sigma = t\delta_a + (1-t)\delta_b$$

for $t \in (0,1)$ and $\alpha < a < b$. Since in this case the only relevant values in the optimization problem to find the density for the Γ-limit are $\varphi(a)$ and $\varphi(b)$, let us put for simplicity

$$A = \varphi(a), \quad B = \varphi(b).$$

Then we must solve

$$\text{Minimize in } (A,B): \quad ta\,|A|^2 + (1-t)b\,|B|^2$$

subject to

$$\rho = tA + (1-t)B.$$

The optimal solution is easily found to be

$$A = \frac{b}{tb + (1-t)a}\rho, \quad B = \frac{a}{tb + (1-t)a}\rho,$$

and hence the value of the infimum is

$$\frac{ab}{tb + (1-t)a}\,|\rho|^2.$$

In particular if $\{a_j\}$ generates the Young measure

$$\sigma_x = t(x)\delta_a + (1-t(x))\delta_b,$$

then the Γ-limit is

$$I(u) = \int_\Omega \frac{ab}{t(x)b + (1-t(x))a}\,|u(x)|^2\,dx.$$

Let us now assume that $\{a_j\}$ generates the Lebesgue measure restricted to the interval (a,b), where again $\alpha < a < b$. Then we should solve the variational problem

$$\text{Minimize in } \varphi: \quad \int_a^b y\,|\varphi(y)|^2\,dy$$

subject to

$$\rho = \int_a^b \varphi(y)\,dy.$$

By looking at optimality conditions, we find that the optimal $\varphi(y)$ is

$$\varphi(y) = \frac{1}{\log\frac{b}{a}}\frac{1}{y}\rho,$$

and the value of the infimum is

$$\frac{1}{\log \frac{b}{a}} |\rho|^2 .$$

In this case the Γ-limit is

$$I(u) = \frac{1}{\log \frac{b}{a}} \int_a^b |u(x)|^2 \, dx.$$

Another such interesting example corresponds to having

$$d\sigma_x(y) = \chi_{(a,b)}(y) f(x,y) \, dy$$

for a certain $f(x,y)$.

In connection with this last family of examples, consider the case of homogenization of integrals in the periodic case

$$I_j(u) = \int_Q W(jx, u(x)) \, dx,$$

where

$$W(y,\rho) : Q \times \mathbf{R}^d \to \mathbf{R}$$

is Q-periodic in y, and Q is the unit cube in $\mathbf{R}^d$. In this case we can take

$$a_j : Q \to Q, \quad a_j(x) = jx - [jx],$$

where brackets $[\cdot]$ indicate the integer part. It is well known (see Riemann–Lebesgue lemma [7], [13]) that the Young measure associated with $\{a_j\}$ is the Lebesgue measure restricted to Q. Thus the variational problem defining the homogenized functional

$$\psi(\rho) : \mathbf{R}^d \to \mathbf{R}$$

is

$$\text{Minimize in } \varphi : \quad \int_Q CW(y, \varphi(y)) \, dy$$

subject to

$$\rho = \int_Q \varphi(y) \, dy,$$

where CW is the convexification with respect to the u variable. When CW is smooth then optimality conditions may be used to determine explicitly the integrand $\psi(\rho)$ as in the examples above. Explicit dependence of W on x can also be allowed and nonperiodic examples can also be studied.

4. The gradient case. We would like to explore how the previous analysis may be adapted to deal with an explicit dependence on gradients of the integrand defining the functionals I_j. We thus assume

$$I_j(u) = \int_\Omega W(a_j(x), \nabla u(x)) \, dx. \tag{4.1}$$

The same technical assumptions on W as in the nongradient case hold:

$$c\left(|\rho|^p-1\right)\le W(a_j(x),\rho)\le C\left(|\rho|^p+1\right),$$
$$|W(\lambda_1,\rho)-W(\lambda_2,\rho)|\le w(|\lambda_1-\lambda_2|)\,|\rho|^p$$

for some $C>c>0$, all j, a.e. $x\in\Omega$, and some exponent $p>1$. We will restrict our attention here to the scalar case where $u:\Omega\to\mathbf{R}$ and leave the more complicated vector case for a future work [14]. Note that

$$a_j:\Omega\to\mathbf{R}^m,\quad W:\mathbf{R}^m\times\mathbf{R}^N\to\mathbf{R}$$

if Ω is a regular, bounded domain in $\mathbf{R}^N$. Assume that $\{a_j\}$ is uniformly bounded in some Lebesgue space.

Let $u\in W^{1,p}(\Omega)$ be given, and let $u_j\rightharpoonup u$ in $W^{1,p}(\Omega)$. Let

$$\nu=\{\nu_x\}_{x\in\Omega},\quad \nu_x=\mu_{\lambda,x}\otimes\sigma_x \tag{4.2}$$

be the Young measure associated with the pairs $\{(a_j,\nabla u_j)\}$, where $\sigma=\{\sigma_x\}_{x\in\Omega}$ is the one corresponding to $\{a_j\}$. Notice that the lower bound for the case without derivatives is again valid. Indeed if

$$\varphi(\lambda,x)=\int_{\mathbf{R}^N}\rho\,d\mu_{\lambda,x}(\rho),\qquad \nabla u(x)=\int_{\mathbf{R}^m}\varphi(\lambda,x)\,d\sigma_x(\lambda), \tag{4.3}$$

then a lower bound for the Γ-limit is

$$\inf_{\varphi}\left\{\int_\Omega\int_{\mathbf{R}^m}CW(\lambda,\varphi(\lambda,x))\,d\sigma_x(\lambda)\,dx\right\}, \tag{4.4}$$

where φ is admissible as indicated above. The key issue here is how to determine admissibility for the fields $\varphi(\lambda,x)$ in terms not only of σ, but also of the sequence a_j itself. Because of this fact, a general result such as Lemma 2.3 in this situation should be more involved. The difficulties are related to the ones we encounter when trying to adjust the gradients of several components (vector gradients) at the same time. This is due to the fact that we cannot modify the sequence $\{a_j\}$ in the least. Notice that the condition

$$\varphi(\lambda,x)=\int_{\mathbf{R}^N}\rho\,d\mu_{\lambda,x}(\rho)$$

essentially says that admissible vector fields φ should be "averages" of gradients over a partition of Ω related to the sequence $\{a_j\}$. As far as we can tell there is no explicit characterization of such property since we do not know how to reconstruct a gradient field from its averages over certain known sets so as to "patch" them together. This can only be done when there is some adjustment between the admissible fields φ in (4.4) and the sequence $\{a_j\}$. We would like, however, to explore how the proof of Theorem 1.1 can be redone in the gradient situation. The only case where we can proceed to compute the Γ-limit, always by reproducing the proof of Theorem 1.1, corresponds to the situation when the infimum in (4.4) over all admissible φ's equals the infimum over the fields satisfying a certain property related to the sequence $\{a_j\}$ itself. One such situation occurs when the sequence $\{a_j\}$ has the AGP. A rigorous way of formalizing this property follows. Formalities are related to the fact that this property needs to be defined locally. B stands for the unit ball in $\mathbf{R}^N$.

Definition 4.1. *We say that the sequence $\{a_j\}$ verifies the AGP (with respect to the exponent p) if $\sigma=\{\sigma_x\}_{x\in\Omega}$ is its corresponding Young measure and for a.e. $x\in\Omega$ whenever*

1. $r_j \searrow 0$ *and* $\{a_j(x + r_j y)\}$, $y \in B$, *generates* σ_x;
2. *for all* j

$$\left\{B(\lambda_k^{(j)}, r_k^{(j)})\right\} \text{ are pairwise disjoint,}$$
$$r_k^{(j)} < r_j \text{ for all } k,$$
$$\sigma_x\left(\mathbf{R}^m \setminus \cup_k B(\lambda_k^{(j)}, r_k^{(j)})\right) = 0;$$

3. $v \in W^{1,p}(B)$,

then if we define

$$\Omega_k^{(j)} = \left\{y \in B : a_j(x + r_j y) \in B(\lambda_k^{(j)}, r_k^{(j)})\right\}$$

and

$$V_j(y) = \frac{1}{\left|\Omega_k^{(j)}\right|} \int_{\Omega_k^{(j)}} \nabla v(z)\, dz, \quad y \in \Omega_k^{(j)},$$

it is true that

$$\| curl V_j\|_{W^{-1,q}(B)} \to 0$$

as $j \to \infty$.

This is a precise way (probably not the only one) of formalizing the heuristic idea that averages of gradients over "level sets" of $\{a_j\}$ are themselves gradients. As such, it has been specifically tailored to redo the proof of Theorem 1.1 in the gradient case. We will describe some important explicit examples in the next section.

If we define the density

$$\psi(x, \rho) : \Omega \times \mathbf{R}^N \to \mathbf{R}$$

by putting, just as before,

(4.5)
$$\psi(x, \rho) = \inf_{\varphi} \left\{ \int_{\mathbf{R}^m} CW(\lambda, \varphi(\lambda))\, d\sigma_x(\lambda) : \rho = \int_{\mathbf{R}^m} \varphi(\lambda)\, d\sigma_x(\lambda), \text{ and whenever} \right.$$
$$r_j \searrow 0 \text{ is such that } \{a_j(x + r_j y)\} \text{ generates } \sigma_x,$$
$$\left. \| \operatorname{curl} (\varphi(a_j(x + r_j y)))\|_{W^{-1,q}(B)} \to 0 \right\},$$

then we claim that the Γ-limit is

$$I(u) = \int_\Omega \psi(x, \nabla u(x))\, dx. \tag{4.6}$$

LEMMA 4.2. *If* $\{a_j\}$ *verifies the AGP, then the class of admissible fields in the infimum in* (4.5) *is identical to the class for the infimum in* (4.4) *for a.e.* $x \in \Omega$.

Proof. Let $\varphi(\lambda, x)$ be as in (4.3), where

$$\nu_x = \mu_{\lambda,x} \otimes \sigma_x$$

is the slicing measure decomposition of the Young measure associated to the pairs

$\{(a_j, \nabla u_j)\}$ for certain gradients $\{\nabla u_j\}$. The point x is regarded here as a parameter. We know that if $r_j \searrow 0$ is such that the sequence $\{a_j(x + r_j y)\}$, for $y \in B$, generates σ_x (the localization property of Young measures; see [13]) and

$$\Omega_k^{(j)} = \left\{ y \in B : a_j(x + r_j y) \in B(\lambda_k^{(j)}, r_k^{(j)}) \right\},$$
$$\sigma_x \left(\mathbf{R}^m \setminus \cup_k B(\lambda_k^{(j)}, r_k^{(j)}) \right) = 0,$$
$$\left\{ B(\lambda_k^{(j)}, r_k^{(j)}) \right\} \text{ pairwise disjoint for all } j, \quad r_k^{(j)} < r_j,$$

then

$$\sup_k \left| \frac{1}{\left| \Omega_k^{(j)} \right|} \int_{\Omega_k^{(j)}} \nabla u_j(z)\, dz - \varphi(\lambda_k^{(j)}, x) \right| \to 0 \quad \text{as } j \to \infty.$$

This is one important property of the slicing measure decomposition. Consequently if V_j is taken as in Definition 4.1 for the gradients ∇u_j, after a standard diagonal argument, we have

$$\| V_j(y) - \varphi(a_j(x + r_j y), x) \|_{L^p(B)} \to 0 \quad \text{as } j \to \infty,$$

where

$$\| \operatorname{curl}_y V_j(y) \|_{W^{-1,q}(B)} \to 0.$$

This certainly implies that

$$\| \operatorname{curl}_y \varphi(a_j(x + r_j y), x) \|_{W^{-1,q}(B)} \to 0. \qquad \square$$

In the next lemma we gather several elementary or well-known facts.

LEMMA 4.3.

1. *Every probability measure supported in* $\mathbf{R}^N$ *can be generated by a sequence of gradients (the scalar case; see* [13]*).*
2. *If* $\| curl V_j \|_{W^{-1,q}(\Omega)} \to 0$*, there exists a sequence* $\{U_j\}$*, bounded in* $W^{1,p}(\Omega)$*, such that*
$$\| \nabla U_j - V_j \|_{L^p(\Omega)} \to 0.$$
3. *If* $\| U_j - V_j \|_{L^p(\Omega)} \to 0$*, then the two sequences* $\{U_j\}$ *and* $\{V_j\}$ *generate the same Young measure* [13].

Proof of Theorem 1.2. We are now ready to prove Theorem 1.2. We follow closely the same procedure as in the proof of Theorem 1.1. The key change refers to the fact that by Lemma 4.2, the field φ_0 is locally (almost) a gradient in λ. Thus, by item 1 of Lemma 4.3, and for j, k fixed, the family of probability measures

$$\nu_k^{(j)}(y) = \mu_{\lambda_{k,r}^{(j,s)}, x_k^{(j)}} \otimes \sigma_{x_k^{(j)}}$$

for $y \in \Lambda_{k,r}^{(j,s)}$, where as before

$$\Lambda_{k,r}^{(j,s)} = \left\{ y \in B : a_j(x_k^{(j)} + \epsilon_k^{(j)})y) \in B(\lambda_{k,r}^{(j,s)}, \delta_{k,r}^{(j,s)}) \right\} \cap B,$$

is a gradient Young measure in B because its first moment is, by construction,

$$\varphi_0\left(a_j(x_k^{(j)}+\epsilon_k^{(j)})y), x_k^{(j)}\right),$$

and this field is essentially a gradient (the characterization of Young measures; see [13]). Hence, the sequence $\{u_{j,s}\}$ in the proof of Theorem 1.1 can be taken such that

$$\lim_{j\to\infty}\lim_{s\to\infty}\|\operatorname{curl} u_{j,s}\|_{W^{-1,q}(\Omega)}\to 0.$$

Keep in mind that the integral of φ_0 against σ is also a gradient. We conclude by applying items 2 and 3 of Lemma 4.3 to the pairs $\left\{(a_j, u_{j,s(j)})\right\}$ for an appropriate subsequence. □

Because x in (4.3) is like a parameter, finding explicitly the density for the Γ-limit (always under the AGP) amounts to solving problems of the type

$$\text{Minimize in } \varphi: \quad \int_{\mathbf{R}^m} F(\lambda,\varphi(\lambda))\,d\sigma(\lambda)$$

subject to

$$\rho=\int_{\mathbf{R}^m}\varphi(\lambda)\,d\sigma(\lambda),$$
$$\|\operatorname{curl}\,(\varphi(a_j(y))\|_{W^{-1,q}(D)}\to 0,$$

where a_j is defined in D, and it generates σ as a homogeneous Young measure.

Notice that a fact as Corollary 2.5 cannot hold in this situation because the underlying structure of the sequence $\{a_j\}$ is embedded in the definition of the density for the Γ-limit.

5. Some examples for the gradient case. One of the simplest examples we can consider is the gradient version of our first example in section 3 in two dimensions:

$$I_j(u)=\int_\Omega a_j(x)\,|\nabla u(x)|^2\,dx,$$

where $\Omega\subset\mathbf{R}^2$, $a_j(x)=\chi(jx\cdot n)a+(1-\chi(jx\cdot n))b$, n is a unit vector, χ is the characteristic function of the interval $(0,t)$ over $(0,1)$ extended by periodicity, and a and b are such that $0<a<b$. The Γ-limit can be computed explicitly if $\{a_j\}$ verifies the AGP. Since in this example the Young measure associated with $\{a_j\}$ is homogeneous and supported only in $\{a,b\}$, verifying AGP reduces to checking that if

$$V_j^a=\int_{\{a_j=a\}\cap B}\nabla v(z)\,dz,$$
$$V_j^b=\int_{\{a_j=b\}\cap B}\nabla v(z)\,dz$$

for any $v\in H^1(B)$, then

$$(V_j^a-V_j^b)\cdot Tn\to 0$$

as $j\to\infty$, where T is the counterclockwise $\pi/2$ rotation. In this particular case, because of the divergence theorem,

$$(V_j^a-V_j^b)\cdot Tn=\int_{\partial B_j}v(z)\otimes n(z)\,dS(z)-\int_{\partial B\setminus\partial B_j}v(z)\otimes n(z)\,dS(z),$$

where ∂B_j is the intersection of the region where $a_j = a$ with ∂B. It is clear that as j grows larger and larger,

$$(V_j^a - V_j^b) \cdot Tn \to 0.$$

In this way, since $\sigma = t\delta_a + (1-t)\delta_b$ is the (homogeneous) Young measure generated by $\{a_j\}$, the density for the Γ-limit is defined through the optimization problem

$$\text{Minimize in } (A, B): \quad ta\,|A|^2 + (1-t)b\,|B|^2$$

subject to

$$\rho = tA + (1-t)B, \quad (A-B) \cdot Tn = 0.$$

Notice that for an admissible vector field φ,

$$\varphi(a_j(x)) = \chi(jx \cdot n)\varphi(a) + (1 - \chi(jx \cdot n))\varphi(b),$$

and the condition of this field being a gradient amounts to having

$$(\varphi(a) - \varphi(b)) \cdot Tn = 0.$$

We have put $A = \varphi(a)$, $B = \varphi(b)$ as before.

After some elementary, algebraic computations, we find the optimal value

$$\psi(\rho) = (ta + (1-t)b)\,|\rho|^2 - \frac{(b-a)^2 t(1-t)}{(1-t)a + tb}(\rho \cdot n)^2.$$

As expected, we can also write

$$\psi(\rho) = \rho^T H \rho,$$

where

$$H = (ta + (1-t)b)\mathbf{1} - \frac{(b-a)^2 t(1-t)}{(1-t)a + tb} n \otimes n$$

is the associated, effective, or homogenized tensor. $\mathbf{1}$ is the identity matrix. We can iterate this procedure to find the Γ-convergence of higher-order laminates (see [1]).

Another typical example is

$$I_j(u) = \int_Q W(jx, \nabla u(x))\,dx,$$

where Q is the unit cube in $\mathbf{R}^2$, $W(y, \rho)$ is Q-periodic in y, and we have the bounds

$$c\left(|\rho|^2 - 1\right) \le W(y, \rho) \le C\left(|\rho|^2 + 1\right)$$

for $0 < c < C$ and all $y \in Q$ as well as the uniform continuity in the y variable. In this case we take, as in section 3,

$$a_j : Q \to Q, \quad a_j(x) = jx - [jx].$$

We also know that the Young measure associated with $\{a_j\}$ is the Lebesgue measure restricted to Q (homogeneous). In this example, the AGP is also easy to check since

for a given $v \in H^1(B)$, the fields V_j in Definition 4.1 (when $\lim_{j\to\infty} jr_j = +\infty$) converge strongly in $L^p(B)$ to the average of ∇v over B due to periodicity. Since this limit is a constant vector, the AGP holds.

To compute the density for the Γ-limit, we need to examine the requirement about

$$\operatorname{curl}_y(\varphi(a_j(x + r_j y), x))$$

converging to zero. In this case, due to homogeneity, the x-dependence is irrelevant. If r_j is such that $s_j = jr_j$ converges to $+\infty$, then we need to determine the fields $\varphi : \mathbf{R}^2 \to \mathbf{R}^2$ such that

$$\operatorname{curl}_y(\varphi(s_j y - [s_j y]) \to 0.$$

Clearly, we can consider φ as Q-periodic, and then if $z = s_j y$,

$$\operatorname{curl}_y(\varphi(s_j y)) = s_j \operatorname{curl}_z \varphi(z),$$

so that since $s_j \to +\infty$, we conclude that $\varphi = \nabla\xi$ for some (possibly nonperiodic) field ξ. The optimization problem defining the density for the Γ-limit is

$$\text{Minimize in } \xi : \quad \int_Q CW(y, \nabla\xi(y))\, dy$$

subject to

$$\rho = \int_Q \nabla\xi(y)\, dy, \quad \nabla\xi,\ Q\text{-periodic.}$$

By an elementary change, we can reformulate the problem as

$$\text{Minimize in } \xi : \quad \int_Q CW(y, \rho + \nabla\xi(y))\, dy$$

for all $\xi \in H^1(Q)$, Q-periodic. This is the typical cell problem in homogenization of multiple integrals for the scalar case [4].

REFERENCES

[1] G. Allaire, *Shape Optimization by the Homogenization Method*, Springer-Verlag, New York, 2002.

[2] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, Oxford Math. Monogr., Oxford University Press, New York, 2000.

[3] J. Ball and F. Murat, *$W^{1,p}$-quasiconvexity and variational problems for multiple integrals*, J. Funct. Anal., 58 (1984), pp. 225–253.

[4] A. Braides and A. Defranceschi, *Homogenization of Multiple Integrals*, Oxford Lecture Ser. Math. Appl. 12, Oxford University Press, New York, 1998.

[5] G. Buttazzo and G. DalMaso, *Γ-convergence and optimal control problems*, J. Optim. Theory Appl., 38 (1982), pp. 385–407.

[6] D. Cioranescu and J. Saint Jean Paulin, *Homogenization of Reticulated Structures*, Springer-Verlag, New York, 2002.

[7] B. Dacorogna, *Direct Methods in the Calculus of Variations*, Springer-Verlag, New York, 1989.

[8] G. DalMaso, *Introduction to Γ-Convergence*, Birkhäuser, Boston, 1993.

[9] E. De Giorgi, *Sulla convergenza di alcune successioni di integrali del tipo dell'area*, Rend. Mat. (6), 8 (1975), pp. 277–294.

[10] E. De Giorgi and T. Franzoni, *Su un tipo di convergenza variazionale*, Atti. Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur. (8), 58 (1975), pp. 842–850.
[11] E. De Giorgi and S. Spagnolo, *Sulla convergenza degli integrali dell'energia per operatori ellittici del secondo ordine*, Boll. Un. Mat. Ital. (4), 8 (1973), pp. 391–411.
[12] L. C. Evans, *Weak Convergence Methods for Nonlinear Partial Differential Equations*, CBMS Reg. Conf. Ser. Math., AMS, Providence, RI, 1990.
[13] P. Pedregal, *Parametrized Measures and Variational Principles*, Birkhäuser, Basel, 1997.
[14] P. Pedregal, in preparation, 2004.
[15] S. Spagnolo, *Sulla convergenza delle soluzioni di equazioni paraboliche ed ellittiche*, Ann. Scuola Norm. Sup. Pisa Cl. Sci., 22 (1968), pp. 571–597.
[16] L. Tartar, *H-measures, a new approach for studying homogenisation, oscillations and concentration effects in partial differential equations*, Proc. Roy. Soc. Edinburgh Sect. A, 115 (1990), pp. 193–230.

SIAM J. MATH. ANAL.
Vol. 36, No. 2, pp. 441–462

MULTILEVEL CHARACTERIZATIONS OF ANISOTROPIC FUNCTION SPACES*

GEORGE KYRIAZIS†

Abstract. We present a general method for extending decomposition systems of $L_2(\mathbb{R}^d)$ to decomposition systems for the anisotropic Triebel–Lizorkin and Besov spaces, $F_{p,q}^{\alpha,s}$ and $B_{p,q}^{\alpha,s}$, respectively, for the full range of the indexes. Our approach is based on techniques from harmonic analysis and relies on the boundedness of almost diagonal operators on appropriate sequence spaces. Typical examples of such decomposition systems are the various wavelet-type unconditional bases for $L_2(\mathbb{R}^d)$.

Key words. anisotropic function spaces, almost diagonal operators, unconditional bases, wavelets

AMS subject classifications. 41A17, 41A20, 42B25, 42C15, 46E35

DOI. 10.1137/S0036141003425684

1. Introduction.

Multilevel-basis characterizations of function spaces are important in many applications since they frequently lead to simple characterizations of the spaces in terms of discrete norms applied to the coefficients with respect to that basis. In this context, wavelet-type characterizations of the various isotropic Triebel–Lizorkin and Besov spaces have established themselves as a very useful tool in many fields such as statistics, image processing, and the numerical solutions of elliptic PDEs. In recent years, however, a renewed interest in pseudodifferential operators and Fourier multipliers acting on spaces with different degrees of smoothness in the various coordinate directions, as well as applications of nonlinear approximation, has lead to the study of the more general classes of anisotropic function spaces.

To describe our results we first introduce the standard multi-index notation. In particular, for every $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ and $\beta = (\beta_1, \dots, \beta_d) \in \mathbb{N}_0^d$ ($\mathbb{N}_0 := \mathbb{N} \cup \{0\}$, $d \geq 1$), we let $x^\beta := x_1^{\beta_1} \cdots x_d^{\beta_d}$, $|\beta| := \beta_1 + \cdots + \beta_d$, $\beta! := \beta_1! \cdots \beta_d!$, and $(\cdot)^{(\beta)} := \frac{\partial^{|\beta|}(\cdot)}{\partial^{\beta_1} x_1 \cdots \partial^{\beta_d} x_d}$. Also, if $x, y \in \mathbb{R}^d$, we define $xy := x_1 y_1 + \cdots + x_d y_d$.

We recall that an anisotropy on $\mathbb{R}^d$ is a vector $\alpha = (a_i, \dots, a_d)$ of positive numbers such that $a_1 + \cdots + a_d = d$ and we let $\alpha_{\min} := \min\{a_i : 1 \leq i \leq d\}$ and $\alpha_{\max} := \max\{a_i : 1 \leq i \leq d\}$. If $t \geq 0$ and $x \in \mathbb{R}^d$ the anisotropic dilation is defined by

$$t^\alpha x := (t^{a_1} x_1, \dots, t^{a_d} x_d),$$

and for every $s \in \mathbb{R}$ we use the notation $t^{s\alpha} x := (t^s)^\alpha x$.

Let α be an anisotropy on $\mathbb{R}^d$, which will be fixed for the rest of the article, and $\lambda > 1$. For every $k = (k_1, \dots, k_d) \in \mathbb{Z}^d$, $j \in \mathbb{Z}$, we define the parallelepiped $I_{j,k}^\alpha$ to be the image of the cube $k + [0,1)^d$ under anisotropic dilation by $\lambda^{-j\alpha}$, namely,

$$I_{j,k}^\alpha := \lambda^{-ja_1}[k_1, k_1 + 1) \times \cdots \times \lambda^{-ja_d}[k_d, k_d + 1),$$

*Received by the editors March 31, 2003; accepted for publication (in revised form) January 16, 2004; published electronically July 29, 2004.
http://www.siam.org/journals/sima/36-2/42568.html
†Department of Mathematics and Statistics, University of Cyprus, P.O. Box 20537, 1678 Nicosia, Cyprus (kyriazis@ucy.ac.cy).

and we use the notation $x_{I^\alpha_{j,k}} := (\lambda^{-ja_1}k_1, \dots, \lambda^{-ja_d}k_d)$ for its lower left corner. The volume of a parallelepiped I will be denoted by $|I|$ ($|I^\alpha_{j,k}| = \lambda^{-jd}$) and we define $\ell(I) := |I|^{1/d}$ its average sidelength. Then, for every $j \in \mathbb{N}_0$ the set $\mathcal{D}_j := \{I^\alpha_{j,k} : k \in \mathbb{Z}^d\}$ forms a disjoint partition of $\mathbb{R}^d$ and we define $\mathcal{D} := \cup_{j\in\mathbb{Z}}\mathcal{D}_j$ and $\mathcal{D}_+ := \cup_{j\in\mathbb{N}_0}\mathcal{D}_j$.

We denote by $\mathcal{S} := \mathcal{S}(\mathbb{R}^d)$ the Schwartz space of infinitely differentiable, rapidly decreasing functions on $\mathbb{R}^d$ and by $\mathcal{S}' := \mathcal{S}'(\mathbb{R}^d)$ its dual, the space of tempered distributions. The Fourier transform $\widehat{f}$ of an integrable function is defined by

$$\widehat{f}(\xi) = \int_{\mathbb{R}^d} f(x)e^{-ix\xi}\,dx,$$

while its inverse is defined by $\check{f}(\xi) = (2\pi)^{-d}\widehat{f}(-\xi)$. Duality now extends the Fourier transform and thus its inverse uniquely from $\mathcal{S}$ to $\mathcal{S}'$. Finally, we use $\langle f, \eta\rangle$ for the standard inner product $\int f\overline{\eta}$ of two functions, when this makes sense, and the same notation is employed for the action of a distribution $f \in \mathcal{S}'$ on $\overline{\eta} \in \mathcal{S}$.

Let now E be a finite set and $\Psi := \{\psi^e_I : e \in E, I \in \mathcal{D}_+\}$ be a decomposition system for $L_2(\mathbb{R}^d)$ with dual functionals $\tilde{\Psi} := \{\widetilde{\psi}^e_I : e \in E, I \in \mathcal{D}_+\}$, that is, for every $f \in L_2(\mathbb{R}^d)$

$$f = \sum_{e\in E}\sum_{I\in\mathcal{D}_+} \langle f, \widetilde{\psi}^e_I\rangle\psi^e_I. \tag{1.1}$$

Our goal is to study sufficient conditions on $\Psi, \tilde{\Psi}$ so that they form a decomposition system for the inhomogeneous anisotropic Triebel–Lizorkin and Besov spaces. We would like also to characterize the membership of a distribution f in these spaces by the size of the coefficients $\{\langle f, \widetilde{\psi}^e_I\rangle\}_{I,e}$. In particular, we shall prove that under certain smoothness and oscillation assumptions on the families $\Psi, \tilde{\Psi}$, depending on the parameters $s \in \mathbb{R}, 0 < p < \infty$, and $0 < q \le \infty$ (see Theorem 4.1), if $f \in F^{\alpha,s}_{p,q}$, then (1.1) holds in the distributional sense (and in the sense of $F^{\alpha,s}_{p,q}$ when $q \ne \infty$). In addition, we have

$$\|f\|_{F^{\alpha,s}_{p,q}} \approx \sum_{e\in E}\Big\|\Big(\sum_{I\in\mathcal{D}_+}(|I|^{-s/d}|\langle f, \widetilde{\psi}^e_I\rangle|\tilde{\chi}_I)^q\Big)^{1/q}\Big\|_{L_p}, \tag{1.2}$$

where $\tilde{\chi}_I := |I|^{-1/2}\chi_I$ is the characteristic function of I normalized in L_2. Here we have adopted the notation $A \approx B$, which means that there exist constants $C_1, C_2 > 0$ such that $C_1A \le B \le C_2A$. The equivalence constants C_1 and C_2 in (1.2) depend on d, p, q, and s. On other occasions, the reader will have to consult the text to understand the parameters on which the equivalence constants depend on. Throughout the paper, the constants are denoted by C and they may vary at every occurrence.

Similarly (for suitable $\Psi, \tilde{\Psi}$) we shall prove (see Theorem 4.2) that for every $f \in B^{\alpha,s}_{p,q}$, $s \in \mathbb{R}, 0 < p, q \le \infty$, the representation (1.1) holds in the distributional sense (and in the sense of $B^{\alpha,s}_{p,q}$ when $p, q \ne \infty$). Also,

$$\|f\|_{B^{\alpha,s}_{p,q}} \approx \sum_{e\in E}\Bigg(\sum_{m\in\mathbb{N}_0}\Bigg(\sum_{I\in\mathcal{D}_m}\big(|I|^{-s/d+1/p-1/2}|\langle f, \widetilde{\psi}^e_I\rangle|\big)^p\Bigg)^{q/p}\Bigg)^{1/q} \tag{1.3}$$

with the usual modifications when $q = \infty$ or $p = \infty$.

This type of question is well studied in the isotropic cases, especially within the wavelet theory; for a full account see [FJW], [M], [HW], and [K] and the references therein.

Multiscale characterizations for the anisotropic Besov $B^{\alpha,s}_{p,q}$ spaces were given in [GHT], [GT], [H], and [L], by means of compactly supported wavelet bases for $0 < p, q \le \infty$, and $s > 0$ and in the special case where α and s are related by (2.6). In these references Besov spaces are defined via the modulus of smoothness; consequently the techniques used have their roots in approximation theory, and they are not applicable to the Triebel–Lizorkin spaces. In our paper we prefer to define both scales of Besov and Triebel–Lizorkin spaces in a unified way by means of Calderon's reproducing formula. We note that in the case of Besov spaces the two definitions lead to the same spaces for $s > d(\frac{1}{p} - 1)_+$ (see [D]). In particular, our goal is to present a general method for extending decomposition systems (or unconditional bases) of $L_2(\mathbb{R}^d)$, which are usually easier to construct, to decomposition systems (or unconditional bases) for the inhomogeneous anisotropic Besov and Triebel–Lizorkin spaces. As a consequence we establish wavelet characterizations for the $F^{\alpha,s}_{p,q}$ and $B^{\alpha,s}_{p,q}$ spaces for the full range of the indexes and for any sufficiently decaying wavelet bases. As an additional byproduct of our results we point out that the whole apparatus of nonlinear approximation by bases functions (see [De]) becomes available for these spaces as well, with possible applications in numerical methods dealing with semielliptic differential operators. Finally, we note that our results hold also for the homogeneous versions of the anisotropic Besov and Triebel–Lizorkin spaces with minor modifications.

The outline of the paper is as follows. In section 2 we give the definitions of the anisotropic $F^{\alpha,s}_{p,q}$ and $B^{\alpha,s}_{p,q}$ spaces and we briefly review some of their main properties. In section 3 we study the boundedness of almost diagonal operators on the sequence spaces $f^{\alpha,s}_{p,q}$ and $b^{\alpha,s}_{p,q}$, a subject that is important by itself, since it can be used for the study of the boundedness of Calderon–Zygmund-type operators on the $F^{\alpha,s}_{p,q}$ and $B^{\alpha,s}_{p,q}$ spaces. In section 4 we present the main results of our paper, and finally in section 5 we apply these results within the framework of wavelet bases. Some technical lemmas have been included in the appendix section 6.

2. Anisotropic function spaces. Let $\alpha = (a_1, \dots, a_d)$ be our fixed anisotropy. An anisotropic distance associated to α is a continuous function $u : \mathbb{R}^d \to \mathbb{R}$ such that $u(x) > 0, x \neq 0$, and $u(t^\alpha x) = tu(x), t > 0$; typical examples are given by

$$u_p(x) = \left(\sum_{i=1}^{d} |x_i|^{p/a_i} \right)^{1/p}, \quad x \in \mathbb{R}^d,\ 0 < p < \infty.$$

It is well known that any two anisotropic distances are equivalent and that there exists a $C^\infty(\mathbb{R}^d \setminus \{0\})$ anisotropic distance function (see [Y]), which we denote by $|\cdot|_\alpha$. We also recall that any anisotropic distance satisfies a quasi-triangular inequality, and in particular there exists $c_\alpha > 1$ such that $|x + y|_\alpha \le c_\alpha(|x|_\alpha + |y|_\alpha), x, y \in \mathbb{R}^d$.

To define the various function spaces that we are interested in we consider $\lambda > 1$, which will be fixed for the rest of this section, and we further assume that

$$\{x \in \mathbb{R}^d : |x|_\alpha \le \lambda\} \subset [-\pi, \pi]^d. \tag{2.1}$$

(This assumption is needed to establish the anisotropic version of Calderon's reproducing formula (2.12).)

Let now $\varphi_0 \in \mathcal{S}$ be such that

$$\text{supp } \widehat{\varphi_0} \subset \{\xi \in \mathbb{R}^d : |\xi|_\alpha \leq \lambda\} \quad \text{and} \quad \widehat{\varphi_0}(\xi) = 1 \text{ if } |\xi|_\alpha \leq 1. \tag{2.2}$$

We let also

$$\widehat{\varphi}(\xi) := \widehat{\varphi}_0(\xi) - \widehat{\varphi}_0(\lambda^\alpha \xi), \quad \xi \in \mathbb{R}^d, \tag{2.3}$$

and we define $\varphi_\nu(x) := \lambda^{\nu d}\varphi(\lambda^{\nu\alpha}x), x \in \mathbb{R}^d$, $\nu \in \mathbb{N}$. Then it is easily seen that

$$\sum_{\nu\in\mathbb{N}_0} \widehat{\varphi_\nu}(\xi) = 1, \quad \xi \in \mathbb{R}^d. \tag{2.4}$$

Let $s \in \mathbb{R},\ 0 < p < \infty,\ 0 < q \leq \infty$; the anisotropic Triebel–Lizorkin space $F^{\alpha,s}_{p,q}$ is defined to be the set of all $f \in \mathcal{S}'$ such that

$$\|f\|_{F^{\alpha,s}_{p,q}} := \left\| \left(\sum_{\nu\in\mathbb{N}_0} (\lambda^{\nu s}|\varphi_\nu * f|)^q \right)^{1/q} \right\|_{L_p} < \infty, \tag{2.5}$$

(with the usual modification for $q = \infty$).

We note that by varying the indices s, α, p, q we recover most of the classical isotropic and anisotropic spaces. For instance, if $1 < p < \infty$ and $(s_1, \ldots, s_d) \in \mathbb{N}^d$, then the anisotropic Sobolev space

$$W_p^{(s_1,\ldots,s_d)}(\mathbb{R}^d) := \left\{ f \in \mathcal{S}' : \|f\|_{L_p(\mathbb{R}^d)} + \sum_{j=1}^d \left\| \frac{\partial^{s_j} f}{\partial x_j^{s_j}} \right\|_{L_p(\mathbb{R}^d)} \right\}$$

is identified with $F^{\alpha,s}_{p,2}$ (see [ST]), where α, s are defined by

$$\frac{1}{s} = \frac{1}{d}\left(\frac{1}{s_1} + \cdots + \frac{1}{s_d}\right), \quad \alpha = \left(\frac{s}{s_1}, \ldots, \frac{s}{s_d}\right). \tag{2.6}$$

It also trivially seen that if $s = s_1 = \cdots = s_d$, then $F^{\alpha,s}_{p,q}$ coincides with the isotropic space $F^s_{p,q}$ (see [T]).

In a similar vein, for $s \in \mathbb{R},\ 0 < p, q \leq \infty$, the anisotropic Besov space $B^{\alpha,s}_{p,q}$ is defined to be the set of all $f \in \mathcal{S}'$ such that

$$\|f\|_{B^{\alpha,s}_{p,q}} := \left(\sum_{\nu\in\mathbb{N}_0} (\lambda^{\nu s}\big\|\varphi_\nu * f\big\|_{L_p})^q \right)^{1/q} < \infty \tag{2.7}$$

(with the usual modification for $q = \infty$).

In the literature it is customary to use $\lambda = 2$; nevertheless, from standard estimates (similar to the ones in [T] section 2.3.2) it is not hard to prove that the above definitions are independent of $\lambda > 1$.

Associated to the Triebel–Lizorkin and Besov spaces are the sequence spaces $f^{\alpha,s}_{p,q}$ and the $b^{\alpha,s}_{p,q}$, respectively.

For $s \in \mathbb{R}$, $0 < p < \infty$, and $0 < q \leq \infty$, $f^{\alpha,s}_{p,q}$ is defined to be the space of all complex-valued sequences $h := (h_I)_{I\in\mathcal{D}_+}$ such that

$$\|h\|_{f^{\alpha,s}_{p,q}} := \left\| \left(\sum_{I\in\mathcal{D}_+} (|I|^{-s/d}|h_I|\tilde{\chi}_I)^q \right)^{1/q} \right\|_{L_p} < \infty,$$

where $\tilde{\chi}_I(x) := |I|^{-1/2}\chi_I(x)$, (with the usual modification for $q = \infty$).

Similarly, if $s \in \mathbb{R}$, and $0 < p, q \leq \infty$, $b_{p,q}^{\alpha,s}$ is defined to be the space of all complex-valued sequences $h := (h_I)_{I \in \mathcal{D}_+}$ such that

$$\|h\|_{b_{p,q}^{\alpha,s}} := \left(\sum_{j \in \mathbb{N}_0} \left(\sum_{I \in \mathcal{D}_j} \left(|I|^{-s/d+1/p-1/2} |h_I| \right)^p \right)^{q/p} \right)^{1/q} < \infty$$

(with the usual modification for $p = \infty$ or $q = \infty$).

Regarding the notation of the $F_{p,q}^{\alpha,s}, B_{p,q}^{\alpha,s}, f_{p,q}^{\alpha,s}$, and $b_{p,q}^{\alpha,s}$ spaces, we point out that the anisotropy α appears only implicitly in their definitions, either in the construction of φ or in the description of the parallelepipeds in $\mathcal{D}_+$. Therefore, as mentioned in the introduction, we consider that α is a fixed anisotropy throughout this article, and we are concerned only with the range of the other three indexes s, p, q.

Multiplying (2.4) by $\widehat{f}$ and inverting the Fourier transform we get that for every $f \in \mathcal{S}'$

$$f = \sum_{\nu \in \mathbb{N}_0} \varphi_\nu * f, \tag{2.8}$$

in the sense of $\mathcal{S}'$. We are interested in a discretized and more useful, for our purposes, version of (2.8). Working toward this we recall from [Di] that one can construct functions $\phi_0, \widetilde{\phi}_0, \phi, \widetilde{\phi} \in \mathcal{S}$ satisfying

$$\begin{aligned} &\operatorname{supp} \widehat{\phi_0}, \widehat{\widetilde{\phi}_0} \subset \{\xi : |\xi|_\alpha \leq \lambda\}, \\ &\operatorname{supp} \widehat{\phi}, \widehat{\widetilde{\phi}} \subset \left\{ \xi : \frac{1}{\lambda} \leq |\xi|_\alpha \leq \lambda \right\}, \end{aligned}$$

and such that

$$\sum_{\nu \in \mathbb{N}_0} \overline{\widehat{\widetilde{\phi}_\nu}(\xi)} \widehat{\phi_\nu}(\xi)(\xi) = 1, \quad \xi \in \mathbb{R}^d, \tag{2.9}$$

where as before $\phi_\nu(x) := \lambda^{\nu d} \phi(\lambda^{\nu\alpha} x)$ and $\widetilde{\phi}_\nu(x) := \lambda^{\nu d} \widetilde{\phi}(\lambda^{\nu\alpha} x)$, $x \in \mathbb{R}^d, \nu \in \mathbb{N}$. Similarly to (2.8) we get that for every $f \in \mathcal{S}'$

$$f = \sum_{\nu \in \mathbb{N}_0} \eta_\nu * \phi_\nu * f, \tag{2.10}$$

in the sense of $\mathcal{S}'$, where $\eta_\nu(x) := \overline{\widetilde{\phi}_\nu(-x)}$. This is the so-called Calderon's reproducing formula. The advantage of formula (2.10) over (2.8) is that we can further analyze the smooth terms $\eta_\nu * \phi_\nu * f, \nu \in \mathbb{N}_0$. Using techniques reminiscent of the Shannon sampling theorem, one can show (see [Di]) that for every $f \in \mathcal{S}'$

$$\eta_\nu * \phi_\nu * f(x) = \sum_{I \in \mathcal{D}_\nu} \langle f, \widetilde{\phi}_I \rangle \phi_I(x), \quad x \in \mathbb{R}^d, \nu \in \mathbb{N}_0, \tag{2.11}$$

where for every parallelepiped $I \in \mathcal{D}$,

$$\phi_I(\cdot) := |I|^{-1/2} \phi \left(\frac{\cdot - x_I}{\ell(I)^\alpha} \right), \quad \widetilde{\phi}_I(\cdot) := |I|^{-1/2} \widetilde{\phi} \left(\frac{\cdot - x_I}{\ell(I)^\alpha} \right).$$

It follows that every distribution $f \in F^{\alpha,s}_{p,q}$ (or $B^{\alpha,s}_{p,q}$) can be represented in the form

$$f = \sum_{\nu\in\mathbb{N}_0} \sum_{I\in\mathcal{D}_\nu} \langle f, \widetilde{\phi}_I\rangle \phi_I = \sum_{I\in\mathcal{D}_+} \langle f, \widetilde{\phi}_I\rangle \phi_I, \tag{2.12}$$

in the sense of $\mathcal{S}'$.

Moreover, the coefficients

$$s_I(f) := \langle f, \widetilde{\phi}_I\rangle, \quad I \in \mathcal{D}_+,$$

in (2.12) contain all the necessary information to determine whether a distribution belongs in the class of anisotropic Triebel–Lizorkin or Besov spaces. In particular, it was established in [Di] that if $s \in \mathbb{R}, 0 < p < \infty$, $0 < q \le \infty$, and $\varsigma := (s_I(f))_I$, then

$$\|f\|_{F^{\alpha,s}_{p,q}} \approx \|\varsigma\|_{f^{\alpha,s}_{p,q}}. \tag{2.13}$$

Similarly, if $s \in \mathbb{R}$ and $0 < p, q \le \infty$ (see [Di]), then

$$\|f\|_{B^{\alpha,s}_{p,q}} \approx \|\varsigma\|_{b^{\alpha,s}_{p,q}}. \tag{2.14}$$

We note that Dintelmann considered in [Di] only the case where $\lambda = 2$, along the lines of the isotropic cases established in [FJ]; however, his results follow almost verbatim for any $\lambda > 1$.

To establish (1.2) and (1.3) we first need to study the boundedness of almost diagonal operators on the spaces $f^{\alpha,s}_{p,q}$ and $b^{\alpha,s}_{p,q}$.

3. Almost diagonal matrices. In this section we are interested in giving sufficient conditions on a matrix

$$(a_{IJ})_{I,J\in\mathcal{D}_+}$$

so that it gives rise to a bounded operator $\mathbf{A}$ on $f^{\alpha,s}_{p,q}$ or $b^{\alpha,s}_{p,q}$. Similar to the isotropic cases (see [FJ]), we say that $\mathbf{A}$ is almost diagonal on $f^{\alpha,s}_{p,q}(b^{\alpha,s}_{p,q})$ if there exist $\epsilon > 0$ such that

$$|a_{IJ}| \le C\omega_{IJ}(\epsilon), \quad I, J \in \mathcal{D}_+,$$

with

$$\omega_{IJ}(\epsilon) = \left(\frac{\ell(I)}{\ell(J)}\right)^s \left(1 + \frac{|x_I - x_J|_\alpha}{\max(\ell(I),\ell(J))}\right)^{-\mathcal{J}-\epsilon} \min\left[\left(\frac{\ell(I)}{\ell(J)}\right)^{(d+\epsilon)/2}, \left(\frac{\ell(J)}{\ell(I)}\right)^{(d+\epsilon)/2+\mathcal{J}-d}\right],$$

where $\mathcal{J} := d/\min(1,p,q)$ for $f^{\alpha,s}_{p,q}$ and $\mathcal{J} := d/\min(1,p)$ for $b^{\alpha,s}_{p,q}$.

A basic tool in proving that almost diagonal matrices are bounded on $f^{\alpha,s}_{p,q}$ $(b^{\alpha,s}_{p,q})$ is the strong maximal operator $M_t, t > 0$, defined by

$$M_t(f)(x) := \left(\sup_{Q\ni x} |Q|^{-1} \int_Q |f(y)|^t \, dy\right)^{1/t}, \tag{3.1}$$

where the supremum is taken with respect to all rectangles with sides parallel to the coordinate axes. It is known that if $0 < p < \infty$, $0 < q \le \infty$, and $0 < t < \min\{p, q\}$, then for any sequence of functions $(f_j)_{j\in\mathbb{Z}}$

$$\left\| \left(\sum_{j\in\mathbb{Z}} M_t(f_j)^q \right)^{1/q} \right\|_{L_p} \le C \left\| \left(\sum_{j\in\mathbb{Z}} |f_j|^q \right)^{1/q} \right\|_{L_p}. \tag{3.2}$$

In the case where the sup in (3.1) is taken over, all cubes with sides parallel to the coordinate axes (3.2) is a well-known result of Fefferman and Stein [FS], while for the anisotropic case we refer the reader to [ST].

PROPOSITION 3.1. *Let $0 < p < \infty, 0 < q \le \infty$, and $s \in \mathbb{R}$. An almost diagonal operator on $f^{\alpha,s}_{p,q}$ is bounded.*

Proof. Let $\mathbf{A}$ be an almost diagonal operator on $f^{\alpha,s}_{p,q}$ associated with the matrix $(a_{IJ})_{I,J\in\mathcal{D}_+}$. We recall that

$$\|\mathbf{A}\|_{f^{\alpha,s}_{p,q}\to f^{\alpha,s}_{p,q}} := \sup_{\|h\|_{f^{\alpha,s}_{p,q}}\le 1} \|\mathbf{A}h\|_{f^{\alpha,s}_{p,q}},$$

where

$$(\mathbf{A}h)_I = \sum_{J\in\mathcal{D}_+} a_{IJ} h_J.$$

(The series is absolutely convergent; see proof below.) It follows that

$$\begin{aligned} \|\mathbf{A}h\|_{f^{\alpha,s}_{p,q}} &= \left\| \left(\sum_{I\in\mathcal{D}_+} (|I|^{-s/d} |(\mathbf{A}h)_I| \tilde{\chi}_I)^q \right)^{1/q} \right\|_{L_p} \\ &\le \left\| \left(\sum_{I\in\mathcal{D}_+} \left(|I|^{-s/d} \sum_{J\in\mathcal{D}_+} |a_{IJ}||h_J| \tilde{\chi}_I \right)^q \right)^{1/q} \right\|_{L_p} \\ &\le C(\sigma_1 + \sigma_2), \end{aligned} \tag{3.3}$$

where

$$\sigma_1 := \left\| \left(\sum_{I\in\mathcal{D}_+} \left(|I|^{-s/d} \sum_{|J|\le|I|} |a_{IJ}||h_J| \tilde{\chi}_I \right)^q \right)^{1/q} \right\|_{L_p}$$

and

$$\sigma_2 := \left\| \left(\sum_{I\in\mathcal{D}_+} \left(|I|^{-s/d} \sum_{|J|>|I|} |a_{IJ}||h_J| \tilde{\chi}_I \right)^q \right)^{1/q} \right\|_{L_p}.$$

To estimate σ_1, since $|J| \le |I|$,

$$|a_{IJ}| \le C \left(\frac{\ell(I)}{\ell(J)} \right)^{s+\frac{(d-\epsilon)}{2}-\mathcal{J}} \left(1 + \frac{|x_I - x_J|_\alpha}{\ell(I)} \right)^{-\mathcal{J}-\epsilon}. \tag{3.4}$$

Let $\mu_I := |I|^{-s/d}\tilde{\chi}_I$ and $0 < t < \min\{1, p, q\}$ be such that $\mathcal{J} + \frac{\epsilon}{2} > d/t$. Using Lemmas 6.4 and 6.8 we obtain

$$\begin{aligned}
\sigma_1 &\leq C\bigg\|\bigg(\sum_{I\in\mathcal{D}_+}\bigg(\sum_{|J|\leq|I|}\bigg(\frac{\ell(I)}{\ell(J)}\bigg)^{s+\frac{(d-\epsilon)}{2}-\mathcal{J}}\bigg(1+\frac{|x_I-x_J|_\alpha}{\ell(I)}\bigg)^{-\mathcal{J}-\epsilon}|h_J|\mu_I\bigg)^q\bigg)^{\frac{1}{q}}\bigg\|_{L_p}\\
&= C\bigg\|\bigg(\sum_{n\in\mathbb{N}_0}\sum_{I\in\mathcal{D}_n}\bigg(\sum_{m\geq n}\lambda^{(m-n)(s+\frac{(d-\epsilon)}{2}-\mathcal{J})}\sum_{J\in\mathcal{D}_m}(1+\lambda^n|x_I-x_J|_\alpha)^{-\mathcal{J}-\epsilon}|h_J|\mu_I\bigg)^q\bigg)^{\frac{1}{q}}\bigg\|_{L_p}\\
&\leq C\bigg\|\bigg(\sum_{n\in\mathbb{N}_0}\sum_{I\in\mathcal{D}_n}\bigg(\sum_{m\geq n}\lambda^{(m-n)(s+\frac{(d-\epsilon)}{2}-\mathcal{J}+\frac{d}{t})}M_t\bigg(\sum_{J\in\mathcal{D}_m}|h_J|\chi_J\bigg)\mu_I\bigg)^q\bigg)^{\frac{1}{q}}\bigg\|_{L_p}\\
&= C\bigg\|\bigg(\sum_{n\in\mathbb{N}_0}\bigg(\sum_{m\geq n}\lambda^{(m-n)(-\frac{\epsilon}{2}-\mathcal{J}+\frac{d}{t})}M_t\bigg(\sum_{J\in\mathcal{D}_m}|h_J|\mu_J\bigg)\bigg)^q\bigg)^{\frac{1}{q}}\bigg\|_{L_p}\\
&\leq C\bigg\|\bigg(\sum_{n\in\mathbb{N}_0}\bigg(M_t\bigg(\sum_{I\in\mathcal{D}_n}|h_I|\mu_I\bigg)\bigg)^q\bigg)^{\frac{1}{q}}\bigg\|_{L_p}\\
&\leq C\|h\|_{f^{\alpha,s}_{p,q}},
\end{aligned}$$

where in the last inequality we used the maximal inequality (3.2).

When $|J| > |I|$, we have that

$$|a_{IJ}| \leq C\bigg(\frac{\ell(I)}{\ell(J)}\bigg)^{s+\frac{(d+\epsilon)}{2}}\bigg(1+\frac{|x_I-x_J|_\alpha}{\ell(J)}\bigg)^{-\mathcal{J}-\epsilon}. \tag{3.5}$$

Employing Lemmas 6.4 and 6.8 once more we get

$$\begin{aligned}
\sigma_2 &\leq C\bigg\|\bigg(\sum_{I\in\mathcal{D}_+}\bigg(\sum_{|J|>|I|}\bigg(\frac{\ell(I)}{\ell(J)}\bigg)^{s+\frac{(d+\epsilon)}{2}}\bigg(1+\frac{|x_I-x_J|_\alpha}{\ell(J)}\bigg)^{-\mathcal{J}-\epsilon}|h_J|\mu_I\bigg)^q\bigg)^{\frac{1}{q}}\bigg\|_{L_p}\\
&= C\bigg\|\bigg(\sum_{n\in\mathbb{N}_0}\sum_{I\in\mathcal{D}_n}\bigg(\sum_{m<n}\lambda^{(m-n)(s+\frac{(d+\epsilon)}{2})}\sum_{J\in\mathcal{D}_m}\bigg(1+\frac{|x_I-x_J|_\alpha}{\ell(J)}\bigg)^{-\mathcal{J}-\epsilon}|h_J|\mu_I\bigg)^q\bigg)^{\frac{1}{q}}\bigg\|_{L_p}\\
&\leq C\bigg\|\bigg(\sum_{n\in\mathbb{N}_0}\sum_{I\in\mathcal{D}_n}\bigg(\sum_{m<n}\lambda^{(m-n)(s+\frac{(d+\epsilon)}{2})}M_t\bigg(\sum_{J\in\mathcal{D}_m}|h_J|\chi_J\bigg)\mu_I\bigg)^q\bigg)^{\frac{1}{q}}\bigg\|_{L_p}\\
&= C\bigg\|\bigg(\sum_{n\in\mathbb{N}_0}\bigg(\sum_{m<n}\lambda^{(m-n)\frac{\epsilon}{2}}M_t\bigg(\sum_{J\in\mathcal{D}_m}|h_J|\mu_J\bigg)\bigg)^q\bigg)^{\frac{1}{q}}\bigg\|_{L_p}\\
&\leq C\bigg\|\bigg(\sum_{n\in\mathbb{N}_0}\bigg(M_t\bigg(\sum_{I\in\mathcal{D}_n}|h_I|\mu_I\bigg)\bigg)^q\bigg)^{\frac{1}{q}}\bigg\|_{L_p}\\
&\leq C\|h\|_{f^{\alpha,s}_{p,q}}.
\end{aligned}$$

Putting the two estimates for σ_1 and σ_2 in (3.3) the result follows. □

Similar to the previous proposition for the Besov spaces we have the following.

PROPOSITION 3.2. *Let $0 < p, q \leq \infty$, and $s \in \mathbb{R}$. An almost diagonal operator on $b^{\alpha,s}_{p,q}$ is bounded.*

Proof. Let $\mathbf{A}$ be an almost diagonal operator on $b^{\alpha,s}_{p,q}$ associated with the matrix $(a_{IJ})_{I,J\in\mathcal{D}_+}$. As before, we need to prove that

$$\|\mathbf{A}\|_{b^{\alpha,s}_{p,q}\to b^{\alpha,s}_{p,q}} := \sup_{\|h\|_{b^{\alpha,s}_{p,q}}\le 1} \|\mathbf{A}h\|_{b^{\alpha,s}_{p,q}} < \infty.$$

We shall consider only $0 < p, q < \infty$; the cases $p = \infty$ or $q = \infty$ follow similarly.

Let $h \in b^{\alpha,s}_{p,q}$. To simplify our notation we define $\gamma := s/d - (1/p - 1/2)$ and $\tilde{h}_J := |J|^{-\gamma} h_J$. Since $(\mathbf{A}h)_I = \sum_{J\in\mathcal{D}_+} a_{IJ} h_J$,

$$\begin{aligned}
\|\mathbf{A}h\|^q_{b^{\alpha,s}_{p,q}} &= \sum_{m\in\mathbb{N}_0} \Bigg(\sum_{I\in\mathcal{D}_m} \big(|I|^{-\gamma} |(\mathbf{A}h)_I|\big)^p \Bigg)^{q/p} \\
&\le \sum_{m\in\mathbb{N}_0} \Bigg(\sum_{I\in\mathcal{D}_m} \Bigg(\sum_{J\in\mathcal{D}_+} (|J|/|I|)^\gamma |a_{IJ}||\tilde{h}_J| \Bigg)^p \Bigg)^{q/p} \\
&\le C(\sigma_1^q + \sigma_2^q)
\end{aligned}$$

with

$$\sigma_1 := \Bigg(\sum_{m\in\mathbb{N}_0} \Bigg(\sum_{I\in\mathcal{D}_m} \Bigg(\sum_{|J|\le|I|} (|J|/|I|)^\gamma |a_{IJ}||\tilde{h}_J| \Bigg)^p \Bigg)^{q/p} \Bigg)^{1/q}$$

and

$$\sigma_2 := \Bigg(\sum_{m\in\mathbb{N}_0} \Bigg(\sum_{I\in\mathcal{D}_m} \Bigg(\sum_{|J|>|I|} (|J|/|I|)^\gamma |a_{IJ}||\tilde{h}_J| \Bigg)^p \Bigg)^{q/p} \Bigg)^{1/q}.$$

Case I. $1 \le p < \infty$. For σ_1 using (3.4), Minkowski's inequality, and Lemmas 6.6 and 6.8

$$\begin{aligned}
\sigma_1^q &\le \sum_{m\in\mathbb{N}_0} \Bigg(\sum_{I\in\mathcal{D}_m} \Bigg(\sum_{n\ge m} \sum_{J\in\mathcal{D}_n} (|J|/|I|)^\gamma |a_{IJ}||\tilde{h}_J| \Bigg)^p \Bigg)^{q/p} \\
&\le \sum_{m\in\mathbb{N}_0} \Bigg(\sum_{I\in\mathcal{D}_m} \Bigg(\sum_{n\ge m} \sum_{J\in\mathcal{D}_n} (|J|/|I|)^{\gamma - \frac{s}{d} - \frac12 + \frac{\mathcal{J}}{d} + \frac{\epsilon}{2d}} \Bigg(1 + \frac{|x_I - x_J|_\alpha}{\ell(I)} \Bigg)^{-\mathcal{J}-\epsilon} |\tilde{h}_J| \Bigg)^p \Bigg)^{q/p} \\
&\le C \sum_{m\in\mathbb{N}_0} \Bigg(\sum_{n\ge m} \lambda^{(m-n)(\mathcal{J} - \frac{d}{p} + \frac{\epsilon}{2})} \Bigg(\sum_{I\in\mathcal{D}_m} \Bigg(\sum_{J\in\mathcal{D}_n} \Bigg(1 + \frac{|x_I - x_J|_\alpha}{\ell(I)} \Bigg)^{-\mathcal{J}-\epsilon} |\tilde{h}_J| \Bigg)^p \Bigg)^{\frac1p} \Bigg)^q \\
&\le C \sum_{m\in\mathbb{N}_0} \Bigg(\sum_{n\ge m} \lambda^{(m-n)(\mathcal{J} - \frac{d}{p} + \frac{\epsilon}{2} - \frac{d}{p'})} \Bigg(\sum_{J\in\mathcal{D}_n} |\tilde{h}_J|^p \Bigg)^{1/p} \Bigg)^q \\
&\le C \sum_{m\in\mathbb{N}_0} \Bigg(\sum_{I\in\mathcal{D}_m} \big(|I|^{-s/d + (1/p - 1/2)} |h_I|\big)^p \Bigg)^{q/p} = C\|h\|^q_{b^{\alpha,s}_{p,q}}
\end{aligned}$$

since $\mathcal{J} - \frac{d}{p} + \frac{\epsilon}{2} - \frac{d}{p'} = \mathcal{J} - d + \frac{\epsilon}{2} > 0$.

Similarly, using (3.5), Minkowski's inequality, and Lemma 6.7 we obtain

$$\begin{aligned}
\sigma_2^q &\le \sum_{m\in\mathbb{N}_0}\bigg(\sum_{I\in\mathcal{D}_m}\bigg(\sum_{n<m}\sum_{J\in\mathcal{D}_n}(|J|/|I|)^{\gamma}|a_{IJ}||\tilde{h}_J|\bigg)^p\bigg)^{q/p}\\
&\le \sum_{m\in\mathbb{N}_0}\bigg(\sum_{I\in\mathcal{D}_m}\bigg(\sum_{n<m}\sum_{J\in\mathcal{D}_n}(|J|/|I|)^{\gamma-\frac{s}{d}-\frac12-\frac{\epsilon}{2d}}\bigg(1+\frac{|x_I-x_J|_\alpha}{\ell(J)}\bigg)^{-\mathcal{J}-\epsilon}|\tilde{h}_J|\bigg)^p\bigg)^{q/p}\\
&\le C\sum_{m\in\mathbb{N}_0}\bigg(\sum_{n<m}\lambda^{(m-n)(-\frac{d}{p}-\frac{\epsilon}{2})}\bigg(\sum_{I\in\mathcal{D}_m}\bigg(\sum_{J\in\mathcal{D}_n}\bigg(1+\frac{|x_I-x_J|_\alpha}{\ell(J)}\bigg)^{-\mathcal{J}-\epsilon}|\tilde{h}_J|\bigg)^p\bigg)^{\frac1p}\bigg)^q\\
&\le C\sum_{m\in\mathbb{N}_0}\bigg(\sum_{n<m}\lambda^{(m-n)(-\frac{\epsilon}{2})}\bigg(\sum_{J\in\mathcal{D}_n}|\tilde{h}_J|^p\bigg)^{1/p}\bigg)^q\\
&\le C\sum_{m\in\mathbb{N}_0}\bigg(\sum_{I\in\mathcal{D}_m}\big(|I|^{-s/d+(1/p-1/2)}|h_I|\big)^p\bigg)^{q/p}=C\|h\|^q_{b^{\alpha,s}_{p,q}},
\end{aligned}$$

where in the last inequality we applied Lemma 6.8. Putting the estimates for σ_1 and σ_2 together we get the desired result for $1\le p<\infty$.

Case II. $p\le 1$. Similar to the previous case,

$$\begin{aligned}
\sigma_1^q &\le C\sum_{m\in\mathbb{N}_0}\bigg(\sum_{J\in\mathcal{D}_n}\sum_{n\ge m}\sum_{I\in\mathcal{D}_m}\lambda^{(m-n)(\mathcal{J}-\frac{d}{p}+\frac{\epsilon}{2})p}\bigg(1+\frac{|x_I-x_J|_\alpha}{\ell(I)}\bigg)^{-\mathcal{J}p-\epsilon p}|\tilde{h}_J|^p\bigg)^{q/p}\\
&\le C\sum_{m\in\mathbb{N}_0}\bigg(\sum_{J\in\mathcal{D}_n}\sum_{n\ge m}\lambda^{(m-n)(\mathcal{J}-\frac{d}{p}+\frac{\epsilon}{2})p}|\tilde{h}_J|^p\bigg)^{q/p}\\
&= C\sum_{m\in\mathbb{N}_0}\bigg(\sum_{n\ge m}\lambda^{(m-n)(\mathcal{J}-\frac{d}{p}+\frac{\epsilon}{2})p}\sum_{J\in\mathcal{D}_n}|\tilde{h}_J|^p\bigg)^{q/p}\\
&\le C\sum_{m\in\mathbb{N}_0}\bigg(\sum_{I\in\mathcal{D}_m}\big(|I|^{-s/d+(1/p-1/2)}|h_I|\big)^p\bigg)^{q/p}=C\|h\|^q_{b^{\alpha,s}_{p,q}},
\end{aligned}$$

where in the last inequality we used that $\mathcal{J}-\frac{d}{p}+\frac{\epsilon}{2}>0$.

Finally, from Lemmas 6.5 and 6.8,

$$\begin{aligned}
\sigma_2^q &\le C\sum_{m\in\mathbb{N}_0}\bigg(\sum_{n<m}\sum_{J\in\mathcal{D}_n}\sum_{I\in\mathcal{D}_m}\lambda^{(m-n)(\gamma-\frac{s}{d}-\frac12-\frac{\epsilon}{2d})dp}\bigg(1+\frac{|x_I-x_J|_\alpha}{\ell(J)}\bigg)^{-\mathcal{J}p-\epsilon p}|\tilde{h}_J|^p\bigg)^{q/p}\\
&\le C\sum_{m\in\mathbb{N}_0}\bigg(\sum_{n<m}\sum_{J\in\mathcal{D}_n}\lambda^{(m-n)(-\frac{\epsilon p}{2})}|\tilde{h}_J|^p\bigg)^{q/p}\\
&\le C\sum_{m\in\mathbb{N}_0}\bigg(\sum_{I\in\mathcal{D}_m}\big(|I|^{-s/d+(1/p-1/2)}|h_I|\big)^p\bigg)^{q/p}=C\|h\|^q_{b^{\alpha,s}_{p,q}}.
\end{aligned}$$

This concludes the proof of the proposition. □

Let now $r_1,r_2\in\mathbb{R}$, $M>0$, and $(\theta_I)_{I\in\mathcal{D}_+}$ and $(\eta_I)_{I\in\mathcal{D}_+}$ be families of functions on $\mathbb{R}^d$ that satisfy

$$\int_{\mathbb{R}^d}\theta_I(x)x^\beta\,dx=0,\ \beta\alpha\le r_1,\ |I|<1, \tag{3.6}$$

$$|\theta_I(x)| \le C|I|^{-\frac{1}{2}}\left(1+\frac{|x-x_I|_\alpha}{\ell(I)}\right)^{-M},\ I\in\mathcal{D}_+, \tag{3.7}$$

$$|\theta_I^{(\beta)}(x)| \le C|I|^{-\frac{1}{2}-\frac{\beta\alpha}{d}}\left(1+\frac{|x-x_I|_\alpha}{\ell(I)}\right)^{-M},\ \beta\alpha\le r_2+\alpha_{\max},\ I\in\mathcal{D}_+, \tag{3.8}$$

and

$$\int_{\mathbb{R}^d}\eta_I(x)x^\beta\,dx=0,\quad \beta\alpha\le r_2,\ |I|<1, \tag{3.9}$$

$$|\eta_I(x)| \le C|I|^{-\frac{1}{2}}\left(1+\frac{|x-x_I|_\alpha}{\ell(I)}\right)^{-M},\ I\in\mathcal{D}_+, \tag{3.10}$$

$$|\eta_I^{(\beta)}(x)| \le C|I|^{-\frac{1}{2}-\frac{\beta\alpha}{d}}\left(1+\frac{|x-x_I|_\alpha}{\ell(I)}\right)^{-M},\ \beta\alpha\le r_1+\alpha_{\max},\ I\in\mathcal{D}_+, \tag{3.11}$$

where (3.6), (3.11) and (3.8), (3.9) are void if $r_1<0$ or $r_2<0$, respectively.

Assuming that r_1, r_2 and M are sufficiently large and using Lemma 6.3 (or Remark 6.1 instead, if $|I|=|J|=1$ or when either of r_1, r_2 are negative), it is easily seen that the infinite matrix

$$\mathbf{A}:=(\langle\theta_I,\eta_J\rangle)_{I,J\in\mathcal{D}_+} \tag{3.12}$$

gives rise to a bounded operator on the f- and b-spaces. In particular, we have the following.

COROLLARY 3.1. *Let $0<p<\infty, 0<q\le\infty$, $s\in\mathbb{R}$, and $\mathcal{J}:=d/\min\{1,p,q\}$. Let also $(\theta_I)_{I\in\mathcal{D}_+},(\eta_I)_{I\in\mathcal{D}_+}$ be families of functions satisfying* (3.6)–(3.11) *for some $r_1,r_2\in\mathbb{R}$ and $M>0$. If $r_1>s$, $r_2>\mathcal{J}-d-s$, and $M>\max\{\mathcal{J},d+r_1,d+r_2\}$, then the matrix $\mathbf{A}$ in* (3.12) *defines a bounded operator on $f^{\alpha,s}_{p,q}$.*

COROLLARY 3.2. *Let $0<p,q\le\infty$, $s\in\mathbb{R}$, and $\mathcal{J}:=d/\min\{1,p\}$. We also assume that $(\theta_I)_{I\in\mathcal{D}_+},(\eta_I)_{I\in\mathcal{D}_+}$ are families of functions satisfying* (3.6)–(3.11) *for some $r_1,r_2\in\mathbb{R}$ and $M>0$. If $r_1>s$, $r_2>\mathcal{J}-d-s$, and $M>\max\{\mathcal{J},d+r_1,d+r_2\}$, then the matrix $\mathbf{A}$ in* (3.12) *defines a bounded operator on $b^{\alpha,s}_{p,q}$.*

4. Decomposition systems for function spaces. We start this section by giving two fundamental lemmas that will help us address the convergence of the series in (2.12).

LEMMA 4.1. *Let $s\in\mathbb{R}$, $0<p<\infty$, $0<q\le\infty$, and $\mathcal{J}:=d/\min\{1,p,q\}$. If $(\theta_I)_{I\in\mathcal{D}_+}$ satisfies* (3.6)–(3.8) *for some $r_1,r_2\in\mathbb{R}$ with $r_1>\mathcal{J}-d-s$, $r_2>s$, and $M>\max\{\mathcal{J},d+r_1,d+r_2\}$, then for every $d:=(d_I)_{I\in\mathcal{D}_+}\in f^{\alpha,s}_{p,q}$ the series $\sum_{I\in\mathcal{D}_+}d_I\theta_I$ converges in $\mathcal{S}'$ (and in $F^{\alpha,s}_{p,q}$ for $q\ne\infty$) and*

$$\Big\|\sum_{I\in\mathcal{D}_+}d_I\theta_I\Big\|_{F^{\alpha,s}_{p,q}}\le C\|d\|_{f^{\alpha,s}_{p,q}}. \tag{4.1}$$

Proof. To establish that the series $\sum_{I\in\mathcal{D}_+}d_I\theta_I$ converges in $\mathcal{S}'$ it is sufficient to prove that for every $\eta\in\mathcal{S}$ we have $|\sum_{I\in\mathcal{D}_+}d_I\langle\theta_I,\eta\rangle|<\infty$. For this, one has to use that

$$|\langle\theta_I,\eta\rangle|\le C\ell(I)^{r_1+d/2}(1+|x_I|_\alpha)^{-M}, \tag{4.2}$$

which is an immediate consequence of Lemma 6.3 (or Remark 6.1 if $|I|=1$ or $r_1<0$). We leave the details to the reader and we refer to [K], where we have worked out the full details for the isotropic case, which is similar.

To prove (4.1) we define $c_{IJ} := \langle \theta_I, \widetilde{\phi}_J \rangle$. Then

$$s_J(f) := \langle f, \widetilde{\phi}_J \rangle = \sum_{I \in \mathcal{D}_+} d_I \langle \theta_I, \widetilde{\phi}_J \rangle = \sum_{I \in \mathcal{D}_+} d_I c_{IJ}, \quad J \in \mathcal{D}.$$

In other words, if $\mathbf{C} := (c_{IJ})_{I,J \in \mathcal{D}}$ and $\varsigma := (s_J(f))_J$, we have

$$\varsigma = \mathbf{C}^T d,$$

where $\mathbf{C}^T$ is the transpose of the matrix $\mathbf{C}$. Applying Corollary 3.1 we get that $\mathbf{C}^T$ is an almost diagonal matrix on $f^{\alpha,s}_{p,q}$, and therefore bounded. Thus, from (2.13), it follows that

$$\|f\|_{F^{\alpha,s}_{p,q}} \approx \|\varsigma\|_{f^{\alpha,s}_{p,q}} = \|\mathbf{C}^T d\|_{f^{\alpha,s}_{p,q}} \leq C \|d\|_{f^{\alpha,s}_{p,q}}.$$

Finally we note that once (4.1) has been established it follows that for $q \neq \infty$ the series $\sum_{I \in \mathcal{D}_+} d_I \theta_I$ converges in the sense of $F^{\alpha,s}_{p,q}$, since its tail $\sum_{|I| \geq N} d_I \theta_I$ converges strongly to 0, as $N \to \infty$. □

Similarly, in the case of Besov spaces we have the following lemma.

LEMMA 4.2. *Let $s \in \mathbb{R}$, $0 < p, q \leq \infty$, and $\mathcal{J} := d/\min\{1, p\}$. If $(\theta_I)_{I \in \mathcal{D}_+}$ satisfies* (3.6)–(3.8) *for some $r_1, r_2 \in \mathbb{R}$, with $r_1 > \mathcal{J} - d - s$, $r_2 > s$, and $M > \max\{\mathcal{J}, d + r_1, d + r_2\}$, then for every $d := (d_I)_{I \in \mathcal{D}_+} \in B^{\alpha,s}_{p,q}$ the series $\sum_{I \in \mathcal{D}_+} d_I \theta_I$ converges in $\mathcal{S}'$ (and in $B^{\alpha,s}_{p,q}$ for $p, q \neq \infty$) and*

$$\left\| \sum_{I \in \mathcal{D}_+} d_I \theta_I \right\|_{B^{\alpha,s}_{p,q}} \leq C \|d\|_{b^{\alpha,s}_{p,q}}. \tag{4.3}$$

Proof. For the convergence of the series $\sum_{I \in \mathcal{D}_+} d_I \theta_I$ we note that since $d \in b^{\alpha,s}_{p,q}$, then $|d_I| \leq C|I|^{s/d+1/2-1/p}$, $I \in \mathcal{D}_+$. Using now this estimate it is not hard to see that the series converges absolutely. Again we refer the reader to [K], where we have worked out the details for the isotropic case. As far as the proof of (4.3) is concerned it is identical to the one of (4.1) since under our assumptions the matrix $\mathbf{C}^T$ is bounded on $b^{\alpha,s}_{p,q}$. □

Now let E be a finite set and $\Psi := \{\psi^e_I : e \in E, I \in \mathcal{D}_+\}$ be a decomposition system for $L_2(\mathbb{R}^d)$ with dual functionals $\tilde{\Psi} := \{\widetilde{\psi}^e_I : e \in E, I \in \mathcal{D}_+\}$; that is, for every $f \in L_2(\mathbb{R}^d)$

$$f = \sum_{e \in E} \sum_{I \in \mathcal{D}_+} \langle f, \widetilde{\psi}^e_I \rangle \psi^e_I.$$

We further assume that for every $e \in E$, the families $(\widetilde{\psi}^e_I)_{I \in \mathcal{D}_+}$ and $(\psi^e_I) I \in \mathcal{D}_+$ satisfy (3.6)–(3.8) and (3.9)–(3.11), respectively, with $r_{\tilde{\Psi}}$ instead of r_1 and r_Ψ instead of r_2. In particular for every $I \in \mathcal{D}_+$, we assume that

$$\int_{\mathbb{R}^d} x^\beta \psi^e_I(x)\, dx = 0, \quad \beta\alpha \leq r_{\tilde{\Psi}}, \ |I| < 1, \tag{4.4}$$

$$|\psi^e_I(x)| \leq C|I|^{-\frac{1}{2}} \left(1 + \frac{|x - x_I|_\alpha}{\ell(I)}\right)^{-M}, \tag{4.5}$$

$$|(\psi^e_I)^{(\beta)}(x)| \leq C|I|^{-\frac{1}{2} - \frac{\beta\alpha}{d}} \left(1 + \frac{|x - x_I|_\alpha}{\ell(I)}\right)^{-M}, \quad \beta\alpha \leq r_\Psi + \alpha_{\max}, \tag{4.6}$$

and

$$\int_{\mathbb{R}^d} x^\beta \widetilde{\psi}_I^e(x)\,dx = 0, \quad \beta\alpha \le r_\Psi,\ |I| < 1, \tag{4.7}$$

$$|\widetilde{\psi}_I^e(x)| \le C|I|^{-\frac{1}{2}}\left(1 + \frac{|x - x_I|_\alpha}{\ell(I)}\right)^{-M}, \tag{4.8}$$

$$|(\widetilde{\psi}_I^e)^{(\beta)}(x)| \le C|I|^{-\frac{1}{2}-\frac{\beta\alpha}{d}}\left(1 + \frac{|x - x_I|_\alpha}{\ell(I)}\right)^{-M}, \quad \beta\alpha \le r_{\widetilde{\Psi}} + \alpha_{\max}, \tag{4.9}$$

where $M > 0$ and $r_{\widetilde{\Psi}}, r_\Psi \in \mathbb{R}$. Of course (4.4), (4.9) and (4.6), (4.7) are void if $r_{\widetilde{\Psi}} < 0$ or $r_\Psi < 0$, respectively.

We note that for every $I \in \mathcal{D}_+$ since $\phi_I \in L_2(\mathbb{R}^d)$,

$$\phi_I = \sum_{e\in E}\sum_{J\in\mathcal{D}_+} \langle \phi_I, \widetilde{\psi}_J^e\rangle \psi_J^e. \tag{4.10}$$

Moreover, if $\psi_J^e, \widetilde{\psi}_J^e, J \in \mathcal{D}_+, e \in E$, satisfy (4.4)–(4.9) with $r_{\widetilde{\Psi}} > \mathcal{J} - d - s$ and $r_\Psi > s$, it is not hard to see that the sequence $(\langle \phi_I, \widetilde{\psi}_J^e\rangle)_{J\in\mathcal{D}_+} \in f_{p,q}^{\alpha,s}$ for $\mathcal{J} := d/\min\{1,p,q\}$ (or $b_{p,q}^{\alpha,s}$ for $\mathcal{J} := d/\min\{1,p\}$). It follows from Lemmas 4.1 and 4.2 that the convergence in (4.10) can be also considered in the sense of $F_{p,q}^{\alpha,s}$ or $B_{p,q}^{\alpha,s}$, respectively.

Also, for every $f \in F_{p,q}^{\alpha,s}$ (or $B_{p,q}^{\alpha,s}$) from (2.12) we get that

$$\begin{aligned} f &= \sum_{I\in\mathcal{D}_+} \langle f, \widetilde{\phi}_I\rangle \phi_I = \sum_{I\in\mathcal{D}_+}\sum_{e\in E}\sum_{J\in\mathcal{D}_+} \langle f, \widetilde{\phi}_I\rangle\langle \phi_I, \widetilde{\psi}_J^e\rangle \psi_J^e \\ &= \sum_{e\in E}\sum_{J\in\mathcal{D}_+}\sum_{I\in\mathcal{D}_+} \langle f, \widetilde{\phi}_I\rangle\langle \phi_I, \widetilde{\psi}_J^e\rangle \psi_J^e \\ &= \sum_{e\in E}\sum_{J\in\mathcal{D}_+} \langle f, \widetilde{\psi}_J^e\rangle \psi_J^e, \end{aligned}$$

where all identities above are considered in the distributional sense. To justify the third equality, we note that our assumptions guarantee that for every $e \in E$ the sequence

$$(d_J^e)_{J\in\mathcal{D}_+} := \Big(\sum_{I\in\mathcal{D}_+} |\langle f, \widetilde{\phi}_I\rangle||\langle \phi_I, \widetilde{\psi}_J^e\rangle|\Big)_{J\in\mathcal{D}_+}$$

belongs in $f_{p,q}^{\alpha,s}$ (or $b_{p,q}^{\alpha,s}$). Similar now to Lemma 4.1 (or Lemma 4.2), it follows that for every $\eta \in \mathcal{S}$

$$\sum_{J\in\mathcal{D}_+} |d_J^e||\langle \psi_J^e, \eta\rangle| < \infty,$$

which allows us to interchange the order of the summations.

THEOREM 4.1. *Let $s \in \mathbb{R}$, $0 < p < \infty$, $0 < q \le \infty$, and $\mathcal{J} := d/\min\{1,p,q\}$. Let also $\Psi, \widetilde{\Psi}$ be a decomposition system for $L_2(\mathbb{R}^d)$ satisfying (4.4)–(4.9) for some $r_\Psi, r_{\widetilde{\Psi}} \in \mathbb{R}$ with $r_\Psi > s$, $r_{\widetilde{\Psi}} > \mathcal{J} - d - s$, and $M > \max\{\mathcal{J}, d + r_\Psi, d + r_{\widetilde{\Psi}}\}$. Then, for every $f \in F_{p,q}^{\alpha,s}$,*

$$f = \sum_{e\in E}\sum_{I\in\mathcal{D}_+} \langle f, \widetilde{\psi}_I^e\rangle \psi_I^e, \tag{4.11}$$

in the sense of $\mathcal{S}'$ *(and in* $F^{\alpha,s}_{p,q}$ *for* $q \neq \infty$*). Moreover,*

$$\|f\|_{F^{\alpha,s}_{p,q}} \approx \sum_{e\in E} \|(\langle f, \widetilde{\psi}^e_I\rangle)_I\|_{f^{\alpha,s}_{p,q}}. \tag{4.12}$$

Proof. Taking into account our discussion above we only have to establish (4.12). From (4.11) we get that for every $I \in \mathcal{D}_+$

$$\langle f, \widetilde{\phi}_I\rangle = \sum_{e\in E}\sum_{J\in\mathcal{D}_+} \langle f, \widetilde{\psi}^e_J\rangle\langle \psi^e_J, \widetilde{\phi}_I\rangle = \sum_{e\in E}\sum_{J\in\mathcal{D}_+} a^e_{JI}\tilde{a}^e_J(f), \tag{4.13}$$

where

$$\tilde{a}^e_J(f) := \langle f, \widetilde{\psi}^e_J\rangle, \quad a^e_{JI} := \langle \psi^e_J, \widetilde{\phi}_I\rangle, \quad I, J \in \mathcal{D}_+, e \in E.$$

Then, if $\tilde{a}_e := (\tilde{a}^e_I(f))_{I\in\mathcal{D}_+}$ and $\mathbf{A}^e := (a^e_{IJ})_{I,J\in\mathcal{D}_+}$, we can express (4.13) in the form

$$\varsigma = \sum_{e\in E} \mathbf{A}^T_e \tilde{a}_e,$$

where as before $\varsigma := (s_I(f))_{I\in\mathcal{D}_+}$.

Similarly, if we define $\tilde{a}^e_{JI} := \langle \phi_J, \widetilde{\psi}^e_I\rangle, I, J \in \mathcal{D}_+$, then

$$\tilde{a}^e_I(f) = \langle f, \widetilde{\psi}^e_I\rangle = \sum_{J\in\mathcal{D}_+} \langle f, \widetilde{\phi}_J\rangle\langle \phi_J, \widetilde{\psi}^e_I\rangle = \sum_{J\in\mathcal{D}_+} \tilde{a}^e_{JI} s_J(f), \quad I \in \mathcal{D}_+.$$

Setting $\tilde{\mathbf{A}}^e := (\tilde{a}^e_{IJ})_{I,J\in\mathcal{D}_+}$ it follows that for every $e \in E$

$$\tilde{a}_e = \tilde{\mathbf{A}}^T_e \varsigma.$$

Employing now Corollary 3.1 we get that the matrices $\tilde{\mathbf{A}}^T_e, \mathbf{A}^T_e, e \in E$, are bounded on $f^{\alpha,s}_{p,q}$ and therefore

$$\begin{aligned}\|\varsigma\|_{f^{\alpha,s}_{p,q}} &= \left\| \sum_{e\in E} \mathbf{A}^T_e \tilde{a}_e \right\|_{f^{\alpha,s}_{p,q}} \leq C \sum_{e\in E} \|\mathbf{A}^T_e \tilde{a}_e\|_{f^{\alpha,s}_{p,q}} \leq C \sum_{e\in E} \|\tilde{a}_e\|_{f^{\alpha,s}_{p,q}} \\ &= C \sum_{e\in E} \|\tilde{\mathbf{A}}^T_e \varsigma\|_{f^{\alpha,s}_{p,q}} \leq C\|\varsigma\|_{f^{\alpha,s}_{p,q}}.\end{aligned}$$

This concludes the proof of the theorem. ∎

THEOREM 4.2. *Let* $s \in \mathbb{R}$, $0 < p, q < \infty$, *and* $\mathcal{J} := d/\min\{1, p\}$. *Let also* $\Psi, \widetilde{\Psi}$ *be a decomposition system for* $L_2(\mathbb{R}^d)$ *satisfying* (4.4)–(4.9) *for some* $r_\Psi, r_{\widetilde{\Psi}} \in \mathbb{R}$ *with* $r_\Psi > s$, $r_{\widetilde{\Psi}} > \mathcal{J} - d - s$, *and* $M > \max\{\mathcal{J}, d + r_\Psi, d + r_{\widetilde{\Psi}}\}$. *Then, for every* $f \in B^{\alpha,s}_{p,q}$,

$$f = \sum_{e\in E}\sum_{I\in\mathcal{D}_+} \langle f, \widetilde{\psi}^e_I\rangle \psi^e_I,$$

in the sense of $\mathcal{S}'$ *(and in* $B^{\alpha,s}_{p,q}$ *for* $p, q \neq \infty$*). Moreover,*

$$\|f\|_{B^{\alpha,s}_{p,q}} \approx \sum_{e\in E} \|(\langle f, \widetilde{\psi}^e_I\rangle)_I\|_{b^{\alpha,s}_{p,q}}. \tag{4.14}$$

Proof. Again we only have to demonstrate (4.14). One has to proceed as in the previous theorem and use Corollary 3.2 instead of Corollary 3.1 to establish that the matrices $\tilde{\mathbf{A}}^T_e, \mathbf{A}^T_e, e \in E$, are bounded on $b^{\alpha,s}_{p,q}$. We leave the details to the reader. ∎

5. Wavelet characterizations of function spaces. Let $\alpha = (a_1, \dots, a_d)$ be a given anisotropy and $\mathbf{M} = \operatorname{diag}(\lambda^{a_1}, \dots, \lambda^{a_d})$ be a diagonal matrix for some fixed $\lambda > 1$. Since $a_1 + \cdots + a_d = d$ we note that $\det \mathbf{M} = \lambda^d$ while for every $j \in \mathbb{N}$ the action of $\mathbf{M}^j$ on $\mathbb{R}^d$ can also be expressed by means of the anisotropic dilation by $\lambda^{j\alpha}$, i.e., for every $x = (x_1, \dots, x_d) \in \mathbb{R}^d$,

$$\mathbf{M}^j x = \lambda^{j\alpha} x = (\lambda^{ja_1} x_1, \dots, \lambda^{ja_d} x_d).$$

A biorthogonal wavelet basis associated to the matrix $\mathbf{M}$ is generated by a couple of scaling functions ψ^0, $\widetilde{\psi}^0$ which, among other assumptions, satisfy the dilation equations

$$\psi^0(x) = |\det \mathbf{M}|^{1/2} \sum_{k \in \mathbb{Z}^d} b_k \psi^0(\mathbf{M}x - k), \quad x \in \mathbb{R}^d,$$
$$\widetilde{\psi}^0(x) = |\det \mathbf{M}|^{1/2} \sum_{k \in \mathbb{Z}^d} c_k \widetilde{\psi}^0(\mathbf{M}x - k), \quad x \in \mathbb{R}^d,$$

for some sequence of complex numbers $(b_k)_k, (c_k)_k \in \ell_2(\mathbb{Z}^d)$, and they have biorthogonal shifts, i.e.,

$$\langle \psi^0(\cdot - k), \widetilde{\psi}^0(\cdot - n) \rangle = \delta_{k,n}, \quad k, n \in \mathbb{Z}.$$

Now let $E = (0, 1, \dots, \lambda^d - 1)$ and $E_0 = E \setminus \{0\}$. Associated to ψ^0 and $\widetilde{\psi}^0$ there exist two families of wavelet functions,

$$\Psi := \{\psi^e : e \in E_0\} \quad \text{and} \quad \widetilde{\Psi}_0 := \{\widetilde{\psi}^e : e \in E\}.$$

Following the wavelet literature for every $I \in \mathcal{D}, e \in E$, we also define

$$\psi_I^e(\cdot) := |I|^{-1/2} \psi^e \left(\frac{\cdot - x_I}{\ell(I)^\alpha} \right), \quad \widetilde{\psi}_I^e(\cdot) := |I|^{-1/2} \widetilde{\psi}^e \left(\frac{\cdot - x_I}{\ell(I)^\alpha} \right). \tag{5.1}$$

In particular, if $I := I_{j,k}^\alpha, k \in \mathbb{Z}^d, j \in \mathbb{Z}$, we note that

$$\psi_I^e(\cdot) = |\det \mathbf{M}^j|^{1/2} \psi^e(\mathbf{M}^j x \cdot - k), \quad \widetilde{\psi}_I^e(\cdot) = |\det \mathbf{M}^j|^{1/2} \widetilde{\psi}^e(\mathbf{M}^j x - k).$$

Then the collection of functions

$$W_0 := \{\psi_I^0, \widetilde{\psi}_I^0 : I \in \mathcal{D}_0, \} \cup \{\psi_I^e, \widetilde{\psi}_I^e : I \in \mathcal{D}_+, e \in E_0\}$$

constitutes a Riesz basis for $L_2(\mathbb{R}^d)$. In particular, for every $f \in L_2(\mathbb{R}^d)$ there exist unique coefficients $\langle f, \widetilde{\psi}_I^e \rangle, I \in \mathcal{D}_+,\ e \in E$, such that

$$f = \sum_{I \in \mathcal{D}_0} \langle f, \widetilde{\psi}_I^0 \rangle \psi_I^{\mathbf{0}} + \sum_{e \in E_0} \sum_{I \in \mathcal{D}_+} \langle f, \widetilde{\psi}_I^e \rangle \psi_I^e \tag{5.2}$$

and

$$\|f\|_{L_2(\mathbb{R}^d)} \approx \Big(\sum_{I\in\mathcal{D}_0} |\langle f, \widetilde{\psi}^0_I\rangle|^2 \Big)^{1/2} + \Big(\sum_{e\in E_0} \sum_{I\in\mathcal{D}_+} |\langle f, \widetilde{\psi}^e_I\rangle|^2 \Big)^{1/2}. \tag{5.3}$$

The construction of such a basis is a delicate matter and in the case of compactly supported wavelet bases it requires that

$$\lambda^{a_i} \in \mathbb{N}, \quad i = 1, \dots, d. \tag{5.4}$$

This in turn forces us to deal only with anisotropies $\alpha = (a_1, \dots, a_d)$ for which there exist $\lambda > 1$ such that (5.4) holds. As it turns out this is equivalent to requiring

$$(a_1, \dots, a_d) \in \mu \log \mathbb{N}^d \text{ for some } \mu > 0. \tag{5.5}$$

In particular, it is not hard to see that (5.5) holds for all anisotropies $\alpha \in \mathbb{Q}^d_+$. We refer the reader to [GT] for the proof of these facts and a complete analysis regarding the construction of anisotropic wavelet bases.

Standard assumptions on the functions $\{\psi^e, \widetilde{\psi}^e : e \in E\}$ include

$$\int_{\mathbb{R}^d} x^\beta \psi^e(x)\, dx = 0, \quad |\beta| \le r_1,\ e \in E_0,$$
$$\int_{\mathbb{R}^d} x^\beta \widetilde{\psi}^e(x)\, dx = 0, \quad |\beta| \le r_2,\ e \in E_0,$$

and

$$|(\psi^e)^{(\beta)}(x)| \le C(1 + |x|_\alpha)^{-M}, \quad |\beta| \le r_2 + m, e \in E,$$
$$|(\widetilde{\psi}^e)^{(\beta)}(x)| \le C(1 + |x|_\alpha)^{-M}, \quad |\beta| \le r_1 + m, e \in E,$$

where $M > 0$, $m \ge 0$, and $r_1, r_2 \in \mathbb{N}_0$. By requiring that r_1, r_2, m, and M are sufficiently large it is readily seen that the families $\{\psi^e_I : e \in E, I \in \mathcal{D}_+\}$, $\{\widetilde{\psi}^e_I : e \in E, I \in \mathcal{D}_+\}$ satisfy the assumptions of Theorems 4.1 and 4.2 and therefore form decomposition systems for the anisotropic Triebel–Lizorkin and Besov spaces. Moreover, the uniqueness of the wavelet coefficients in (5.2) shows that they constitute unconditional bases for these spaces.

6. Appendix: Inequalities. Throughout this section we assume that $\alpha = (a_1, \dots, a_d)$ is a fixed anisotropy. We start with the following version of Taylor's formula, which was given in [F].

LEMMA 6.1. *Let $r \ge 0$ and $f : \mathbb{R}^d \to C$ be a function such that $f^{(\beta)}$ exists for all $\beta \in \mathbb{N}^d_0$ with $\beta\alpha \le r + \alpha_{\max}$. Then, there is a constant C such that for every $x, y \in \mathbb{R}^d$*

$$f(x) = \sum_{\beta\alpha \le r} f^{(\beta)}(y) \frac{(x-y)^\beta}{\beta!} + R_r(x),$$

where

$$|R_r(x)| \le C \sum_{\beta\alpha > r}^{r+\alpha_{\max}} |x-y|^{\beta\alpha}_\alpha \sup_{|z-y|_\alpha \le |x-y|_\alpha} |f^{(\beta)}(z)|.$$

LEMMA 6.2. *Let $J \in \mathcal{D}_+$ and $x_1 \in \mathbb{R}^d$. If η, θ are functions on $\mathbb{R}^d$ such that for some $r \geq 0$ and $M > d + r$ satisfy*

$$\int_{\mathbb{R}^d} x^\beta \eta(x)\, dx = 0, \quad \beta\alpha \leq r, \tag{6.1}$$

$$|\eta(x)| \leq C|J|^{-\frac{1}{2}} \left(1 + \frac{|x - x_1|_\alpha}{\ell(J)}\right)^{-M}, \tag{6.2}$$

and

$$|\theta^{(\beta)}(x)| \leq C\big(1 + |x|_\alpha\big)^{-M}, \quad \beta\alpha \leq r + \alpha_{\max}; \tag{6.3}$$

then,

$$|\langle \theta, \eta \rangle| \leq C|J|^{r/d+1/2}(1 + |x_1|_\alpha)^{-M}.$$

Proof. Although the proof of the lemma is typical, we give a full account of it for the sake of completeness. We recall that there exists $c_\alpha \geq 1$ such that for every $x, y \in \mathbb{R}^d$, $|x + y|_\alpha \leq c_\alpha(|x|_\alpha + |y|_\alpha)$ and that for every $\beta \in \mathbb{N}_0^d$, $|x^\beta| = \prod_{i=1}^d (|x_i|^{1/a_i})^{a_i\beta_i} \leq \prod_{i=1}^d (\sum_{j=1}^d |x_j|^{1/a_j})^{a_i\beta_i} \leq (\sum_{j=1}^d |x_j|^{1/a_j})^{\beta\alpha} \leq C|x|_\alpha^{\beta\alpha}$.

From the moment condition of η we have

$$\begin{aligned}
|\langle \theta, \eta \rangle| &= \left| \int_{\mathbb{R}^d} \theta(y)\overline{\eta(y)}\, dy \right| \\
&= \left| \int_{\mathbb{R}^d} \left[\theta(y) - \sum_{\beta\alpha \leq r} \frac{(y - x_1)^\beta}{\beta!} \theta^{(\beta)}(x_1)\right] \overline{\eta(y)}\, dy \right| \\
&\leq \int_{\mathbb{R}^d} \left| \theta(y) - \sum_{\beta\alpha \leq r} \frac{(y - x_1)^\beta}{\beta!} \theta^{(\beta)}(x_1) \right| |\eta(y)|\, dy.
\end{aligned}$$

We will integrate over $A := \{y : |y - x_1|_\alpha \geq 1\}$ and A^c separately. For the integral over A, from (6.3)

$$\begin{aligned}
&\int_A |\theta(y) - \sum_{\beta\alpha \leq r} \frac{(y - x_1)^\beta}{\beta!} \theta^{(\beta)}(x_1)||\eta(y)|\, dy \\
&\quad \leq C|J|^{-1/2} \int_A (1 + |y|_\alpha)^{-M} \left(1 + \frac{|y - x_1|_\alpha}{\ell(J)}\right)^{-M} dy \\
&\qquad + C|J|^{-1/2} \int_A |y - x_1|_\alpha^r \big(1 + |x_1|_\alpha\big)^{-M} \left(1 + \frac{|y - x_1|_\alpha}{\ell(J)}\right)^{-M} dy \\
&\quad =: B_1 + B_2.
\end{aligned}$$

For B_1, we first consider the case where $|y|_\alpha \leq |x_1|_\alpha / 2c_\alpha$. Then, $|y - x_1|_\alpha \geq |x_1|_\alpha / 2c_\alpha$ and $\frac{|y - x_1|_\alpha}{\ell(J)} \geq \frac{1 + |y - x_1|_\alpha}{2\ell(J)} \geq C\frac{1 + |x_1|_\alpha}{\ell(J)}$. It follows that

$$\begin{aligned}
&|J|^{-1/2} \int_{A \cap \{|y|_\alpha \leq |x_1|_\alpha / 2c_\alpha\}} (1 + |y|_\alpha)^{-M} \left(1 + \frac{|y - x_1|_\alpha}{\ell(J)}\right)^{-M} dy \\
&\leq C|J|^{M/d-1/2}(1 + |x_1|_\alpha)^{-M} \int_A (1 + |y|_\alpha)^{-M}\, dy \qquad (6.4) \\
&\leq C|J|^{M/d-1/2}(1 + |x_1|_\alpha)^{-M}.
\end{aligned}$$

If $|y|_\alpha > |x_1|_\alpha/2c_\alpha$, then $(1+|y|_\alpha)^{-M} \le C(1+|x_1|_\alpha)^{-M}$, and hence we have

$$\begin{aligned} &|J|^{-1/2}\int_{A\cap\{|y|_\alpha>|x_1|_\alpha/2c_\alpha\}}(1+|y|_\alpha)^{-M}\left(1+\frac{|y-x_1|_\alpha}{\ell(J)}\right)^{-M}dy \\ &\le C|J|^{-1/2}(1+|x_1|_\alpha)^{-M}\int_{|y-x_1|_\alpha\ge 1}\left(\frac{|y-x_1|_\alpha}{\ell(J)}\right)^{-M}dy \qquad (6.5)\\ &\le C|J|^{M/d-1/2}(1+|x_1|_\alpha)^{-M}. \end{aligned}$$

For B_2, using that $|y-x_1|_\alpha \ge 1$, we have

$$\begin{aligned} B_2 &\le C|J|^{M/d-1/2}(1+|x_1|_\alpha)^{-M}\int_{|y-x_1|_\alpha\ge 1}|y-x_1|_\alpha^{-M+r}\,dy \qquad (6.6)\\ &\le C|J|^{M/d-1/2}(1+|x_1|_\alpha)^{-M}. \end{aligned}$$

Since $|J|\le 1$ and $M-d/2 > r+d/2$, from (6.4)–(6.6), we find

$$\int_A |\theta(y)-\sum_{\beta\alpha\le r}\frac{(y-x_1)^\beta}{\beta!}\theta^{(\beta)}(x_1)||\eta(y)|\,dy \le C|J|^{r/d+1/2}(1+|x_1|_\alpha)^{-M}.$$

Next, we estimate the integral over $A^c=\{y : |y-x_1|_\alpha < 1\}$. From Taylor's formula we know that

$$\begin{aligned} |\theta(y)-\sum_{\beta\alpha\le r}\frac{(y-x_1)^\beta}{\beta!}\theta^{(\beta)}(x_1)| &\le C\sum_{\beta\alpha>r}^{r+\alpha_{\max}}|y-x_1|_\alpha^{\beta\alpha}\sup_{|z-x_1|_\alpha\le|y-x_1|_\alpha}|\theta^{(\beta)}(z)|\\ &\le C|y-x_1|_\alpha^r\sup_{|z-x_1|_\alpha\le|y-x_1|_\alpha}(1+|z|_\alpha)^{-M}\\ &\le C|y-x_1|_\alpha^r(1+|x_1|_\alpha)^{-M}, \end{aligned}$$

where in the second inequality we used the fact that $|y-x_1|_\alpha\le 1$ and in the last that $|x_1|_\alpha \le c_\alpha(|z-x_1|_\alpha+|z|_\alpha)\le c_\alpha(1+|z|_\alpha)$. It follows that

$$\begin{aligned} &\int_{A^c}|\theta(y)-\sum_{\beta\alpha\le r}\frac{(y-x_1)^\beta}{\beta!}\theta^{(\beta)}(x_1)||\eta(y)|\,dy\\ &\le C|J|^{-1/2}\int_{A^c}|y-x_1|_\alpha^r(1+|x_1|_\alpha)^{-M}\left(1+\frac{|y-x_1|_\alpha}{\ell(J)}\right)^{-M}dy\\ &\le C|J|^{r/d-1/2}(1+|x_1|_\alpha)^{-M}\int_{A^c}\left(1+\frac{|y-x_1|_\alpha}{\ell(J)}\right)^{-M+r}dy\\ &\le C|J|^{r/d+1/2}(1+|x_1|_\alpha)^{-M}. \qquad \square \end{aligned}$$

Using dilations and translations we now easily get the following.

LEMMA 6.3. *Let $I,J\in\mathcal{D}$ with $|J|\le|I|$. We also assume that η_J,θ_I are functions on $\mathbb{R}^d$ such that for some $r\ge 0$ and $M>d+r$ satisfy*

$$\int_{\mathbb{R}^d}x^\beta\eta_J(x)\,dx=0,\quad \beta\alpha\le r, \qquad (6.7)$$

$$|\eta_J(x)|\le C|J|^{-\frac12}\left(1+\frac{|x-x_J|_\alpha}{\ell(J)}\right)^{-M}, \qquad (6.8)$$

and

$$|(\theta_I)^{(\beta)}(x)| \le C|I|^{-\frac{1}{2}-\frac{\beta\alpha}{d}}\left(1+\frac{|x-x_I|_\alpha}{\ell(I)}\right)^{-M}, \quad \beta\alpha \le r+\alpha_{\max}. \tag{6.9}$$

Then,

$$|\langle \theta_I, \eta_J\rangle| \le C\left(\frac{|J|}{|I|}\right)^{r/d+1/2}\left(1+\frac{|x_I-x_J|_\alpha}{\ell(I)}\right)^{-M}.$$

Remark 6.1. In the absence of zero moments, that is, if $|J| \le |I|$ and

$$|\eta_J(x)| \le C|J|^{-\frac{1}{2}}\left(1+\frac{|x-x_J|_\alpha}{\ell(J)}\right)^{-M},$$
$$|\theta_I(x)| \le C|I|^{-\frac{1}{2}}\left(1+\frac{|x-x_I|_\alpha}{\ell(I)}\right)^{-M},$$

$M > d$, then using similar arguments as in the proof of the previous lemma it is easy to show that

$$|\langle \theta_I, \eta_J\rangle| \le C\left(\frac{|J|}{|I|}\right)^{1/2}\left(1+\frac{|x_I-x_J|_\alpha}{\ell(I)}\right)^{-M}. \tag{6.10}$$

LEMMA 6.4. *Let $0 < t \le 1$ and $M > d/t$. For any sequence of complex numbers $(h_J)_{J\in\mathcal{D}_m}$, $m \in \mathbb{Z}$, and $x \in I \in \mathcal{D}$, we have*

$$\sum_{J\in\mathcal{D}_m}|h_J|\left(1+\frac{|x_I-x_J|_\alpha}{\max(\ell(I),\ell(J))}\right)^{-M} \le C\max\left\{\left(\frac{|I|}{|J|}\right)^{\frac{1}{t}},1\right\}M_t\Big(\sum_{J\in\mathcal{D}_m}|h_J|\chi_J\Big)(x).$$

Proof. Without loss of generality we assume that $x_I = 0$.

Case I. $|I| \le 2^{-md}$. We let $\delta := M/d - 1/t > 0$, and for each $j \in \mathbb{N}$ we define $\Omega_j := \{J \in \mathcal{D}_m : \lambda^{j-1} < \lambda^m|x_J|_\alpha \le \lambda^j\}$, while $\Omega_0 := \{J \in \mathcal{D}_m : \lambda^m|x_J|_\alpha \le 1\}$. If $x \in I$, then

$$\sum_{J\in\mathcal{D}_m}|h_J|\big(1+\lambda^m|x_J|_\alpha\big)^{-M} = \sum_{j=0}^{\infty}\sum_{J\in\Omega_j}|h_J|\big(1+\lambda^m|x_J|_\alpha\big)^{-M}$$
$$\le C\sum_{j=0}^{\infty}\sum_{J\in\Omega_j}|h_J|\lambda^{-jM} = C\sum_{j=0}^{\infty}\lambda^{-jd/t-j\delta d}\sum_{J\in\Omega_j}|h_J|$$
$$\le C\sup_{j\ge0}\lambda^{-jd/t}\sum_{J\in\Omega_j}|h_J| \le C\Big(\sup_{j\ge0}\lambda^{-jd}\sum_{J\in\Omega_j}|h_J|^t\Big)^{1/t}$$
$$= C\Big(\sup_{j\ge0}\lambda^{-jd}\lambda^{md}\int\Big(\sum_{J\in\Omega_j}|h_J|\chi_J\Big)^t\Big)^{1/t}$$
$$\le C\Big(\sup_{j\ge0}\frac{1}{|\cup_{J\in\Omega_j}J|}\int_{\cup_{J\in\Omega_j}J}\Big(\sum_{J\in\Omega_j}|h_J|\chi_J\Big)^t\Big)^{1/t}$$
$$\le CM_t\Big(\sum_{J\in\mathcal{D}_m}|h_J|\chi_J\Big)(x).$$

Case II. $|I| > \lambda^{-md}$. Let us assume that $\ell(I) = \lambda^{-n}$, $n < m$. For $j \in \mathbb{N}_+$ we define $\Omega_j := \{J \in \mathcal{D}_m : \lambda^{j-1} < \lambda^n |x_J|_\alpha \leq \lambda^j\}$, while for $j = 0$ we set $\Omega_0 := \{J \in \mathcal{D}_m : \lambda^n |x_J|_\alpha \leq 1\}$. Then for every $x \in I$ we have

$$\begin{aligned}
&\sum_{J\in\mathcal{D}_m} |h_J|\big(1+\lambda^n|x_J|_\alpha\big)^{-M} = \sum_{j=0}^{\infty}\sum_{J\in\Omega_j} |h_J|\big(1+\lambda^n|x_J|_\alpha\big)^{-M}\\
&\leq C\sum_{j=0}^{\infty}\sum_{J\in\Omega_j} |h_J|\lambda^{-jM} = C\sum_{j=0}^{\infty}\lambda^{-jd/t-j\delta d}\sum_{J\in\Omega_j}|h_J|\\
&\leq C\sup_{j\geq 0}\lambda^{-jd/t}\sum_{J\in\Omega_j}|h_J| \leq C\bigg(\sup_{j\geq 0}\lambda^{-jd}\sum_{J\in\Omega_j}|h_J|^t\bigg)^{1/t}\\
&= C\bigg(\sup_{j\geq 0}\lambda^{-jd}\lambda^{md}\int\Big(\sum_{J\in\Omega_j}|h_J|\chi_J\Big)^t\bigg)^{1/t}\\
&\leq C\lambda^{(m-n)d/t}\bigg(\sup_{j\geq 0}\frac{1}{|\cup_{J\in\Omega_j}J|}\int_{\cup_{J\in\Omega_j}J}\Big(\sum_{J\in\Omega_j}|h_J|\chi_J\Big)^t\bigg)^{1/t}\\
&\leq C\lambda^{(m-n)d/t}M_t\Big(\sum_{J\in\mathcal{D}_m}|h_J|\chi_J\Big)(x). \qquad \square
\end{aligned}$$

LEMMA 6.5. *Let $m, n \in \mathbb{Z}$ with $m \geq n$. If $J \in \mathcal{D}_n$ and $M > d$, then*

$$\sum_{I\in\mathcal{D}_m}\bigg(1+\frac{|x_I-x_J|_\alpha}{\ell(J)}\bigg)^{-M} \leq C\lambda^{(m-n)d}.$$

Proof. We have

$$\sum_{I\in\mathcal{D}_m}\bigg(1+\frac{|x_I-x_J|_\alpha}{\ell(J)}\bigg)^{-M} = \lambda^{(m-n)M}\sum_{j\in\mathbb{Z}^d}\big(\lambda^{m-n}+|\lambda^m x_J - j|_\alpha\big)^{-M}.$$

Using now the fact that for every $\rho \geq 1$ and $M > d$,

$$\sum_{j\in\mathbb{Z}^d}(\rho+|j|_\alpha)^{-M} \leq C\rho^{d-M},$$

the result follows. $\square$

LEMMA 6.6. *Let $M > d$, $1 \leq p \leq \infty$, and $m, n \in \mathbb{Z}$ be such that $n \geq m$. If $(d_J)_{J\in\mathcal{D}_n}$ is a sequence of complex numbers, then*

$$\bigg(\sum_{I\in\mathcal{D}_m}\bigg(\sum_{J\in\mathcal{D}_n}\bigg(1+\frac{|x_I-x_J|_\alpha}{\ell(I)}\bigg)^{-M}|d_J|\bigg)^p\bigg)^{1/p} \leq C\lambda^{(n-m)d/p'}\bigg(\sum_{J\in\mathcal{D}_n}|d_J|^p\bigg)^{1/p},$$

where $1/p + 1/p' = 1$.

Proof. We note that for every $I, \Delta \in \mathcal{D}_m$, and $J \in \mathcal{D}_n$ with $J \subset \Delta$,

$$\bigg(1+\frac{|x_I-x_\Delta|_\alpha}{\ell(I)}\bigg) \leq C\bigg(1+\frac{|x_I-x_J|_\alpha}{\ell(I)}\bigg).$$

Also, for every $I \in \mathcal{D}_m$, $\mathcal{D}_m = \{\Delta : \Delta = I + j\lambda^{-m\alpha}, j \in \mathbb{Z}^d\}$. Using these two facts we find

$$\begin{aligned}
&\Big(\sum_{I\in\mathcal{D}_m} \Big(\sum_{J\in\mathcal{D}_n} \Big(1 + \frac{|x_I - x_J|_\alpha}{\ell(I)}\Big)^{-M} |d_J|\Big)^p\Big)^{1/p} \\
&= \Big(\sum_{I\in\mathcal{D}_m} \Big(\sum_{\Delta\in\mathcal{D}_m} \sum_{\substack{J\in\mathcal{D}_n \\ J\subset\Delta}} \Big(1 + \frac{|x_I - x_J|_\alpha}{\ell(I)}\Big)^{-M} |d_J|\Big)^p\Big)^{1/p} \\
&\le C\Big(\sum_{I\in\mathcal{D}_m} \Big(\sum_{\Delta\in\mathcal{D}_m} \sum_{\substack{J\in\mathcal{D}_n \\ J\subset\Delta}} \Big(1 + \frac{|x_I - x_\Delta|_\alpha}{\ell(I)}\Big)^{-M} |d_J|\Big)^p\Big)^{1/p} \\
&= C\Big(\sum_{I\in\mathcal{D}_m} \Big(\sum_{j\in\mathbb{Z}^d} (1+|j|_\alpha)^{-M} \sum_{\substack{J\in\mathcal{D}_n \\ J\subset I+j\lambda^{-m\alpha}}} |d_J|\Big)^p\Big)^{1/p} \\
&\le C \sum_{j\in\mathbb{Z}^d} (1+|j|_\alpha)^{-M} \Big(\sum_{I\in\mathcal{D}_m} \Big(\sum_{\substack{J\in\mathcal{D}_n \\ J\subset I+j\lambda^{-m\alpha}}} |d_J|\Big)^p\Big)^{1/p} \\
&\le C\lambda^{(n-m)d/p'} \sum_{j\in\mathbb{Z}^d} (1+|j|_\alpha)^{-M} \Big(\sum_{I\in\mathcal{D}_m} \sum_{\substack{J\in\mathcal{D}_n \\ J\subset I+j\lambda^{-m\alpha}}} |d_J|^p\Big)^{1/p} \\
&\le C\lambda^{(n-m)d/p'} \Big(\sum_{J\in\mathcal{D}_n} |d_J|^p\Big)^{1/p},
\end{aligned}$$

where we used Minkowski's and Hölder's inequalities. □

In a similar vein we have the following lemma. (We leave the proof to the reader.)

LEMMA 6.7. *Let $M > d$, $1 \le p \le \infty$, and $m, n \in \mathbb{Z}$ be such that $m \ge n$. If $(d_J)_{J\in\mathcal{D}_n}$ is a sequence of complex numbers, then*

$$\Big(\sum_{I\in\mathcal{D}_m} \Big(\sum_{J\in\mathcal{D}_n} \Big(1 + \frac{|x_I - x_J|_\alpha}{\ell(J)}\Big)^{-M} |d_J|\Big)^p\Big)^{1/p} \le C\lambda^{(m-n)d/p} \Big(\sum_{J\in\mathcal{D}_n} |d_J|^p\Big)^{1/p}.$$

Finally, we close the appendix by stating a very useful result, used repeatedly in the proofs of section 3.

LEMMA 6.8. *Let $\lambda > 1$, $\theta > 0$, and $0 < q \le \infty$. If $a_n, b_n \ge 0$, $n \in \mathbb{Z}$, satisfy*

$$0 \le b_n \le \sum_{m\le n} \lambda^{(m-n)\theta} a_m,$$

then

$$\Big(\sum_{n\in\mathbb{Z}} b_n^q\Big)^{1/q} \le C\Big(\sum_{n\in\mathbb{Z}} a_n^q\Big)^{1/q}.$$

REFERENCES

[D] S. DACHKOVSKI, *Anisotropic function spaces and related semi-linear hypoelliptic equations*, Math. Nachr., 248–249 (2003), pp. 40–61.

[De] R. DEVORE, *Nonlinear Approximation*, in Acta Numerica, Cambridge University Press, Cambridge, UK, 1998, pp. 51–150.

[Di] P. Dintelmann, *Classes of Fourier multipliers and Besov-Nikoskij spaces*, Math. Nachr., 173 (1995), pp. 115–130.

[F] W. Farkas, *Atomic and subatomic decompositions in anisotropic function spaces*, Math. Nachr., 209 (2000), pp. 83–113.

[FJ] M. Frazier and B. Jawerth, *A discrete transform and decompositions of distribution*, J. Funct. Anal., 93 (1990), pp. 34–170.

[FJW] M. Frazier, B. Jawerth, and G. Weiss, *Littlewood-Paley Theory and the Study of Function Spaces*, CBMS Reg. Conf. Ser. Math. 79, AMS, Providence, RI, 1991.

[FS] C. Fefferman and E. Stein, *Some maximal inequalities*, Amer. J. Math., 93 (1971), pp. 107–115.

[GHT] G. Garrigos, R. Hochmuth, and A. Tabacco, *Wavelet characterizations for anisotropic Besov spaces: Cases* $0 < p < 1$, Proc. Edinburgh Math. Soc. (2), to appear.

[GT] G. Garrigos and A. Tabacco, *Wavelet decompositions of anisotropic Besov spaces*, Math. Nachr., 239/240 (2002), pp. 80–102.

[H] R. Hochmuth, *Wavelet characterizations for anisotropic Besov spaces*, Appl. Comp. Harm. Anal., 12 (2002), pp. 179–208.

[HW] E. Hernandez and G. Weiss, *A First Course on Wavelets*, Studies in Advanced Mathematics, CRC Press, Boca Raton, FL, 1996.

[K] G. Kyriazis, *Decomposition systems for Function spaces*, Studia Math., 157 (2003), pp. 133–169.

[L] C. Leisner, *Nonlinear wavelet approximation in anisotropic Besov spaces*, Indiana Univ. Math. J., 52 (2003), pp. 437–455.

[M] Y. Meyer, *Ondelettes et Opérateurs* I: *Ondelettes*, Hermann, Paris, 1990.

[ST] B. Stöckert and H. Triebel, *Decomposition methods for function spaces of* B^s_{pq} *type and* F^s_{pq} *type*, Math. Nachr., 89 (1979), pp. 247–267.

[T] H. Triebel, *Theory of Function Spaces*, Birkhäuser, Basel, 1983.

[Y] M. Yamazaki, *A quasi-homogeneous version of paradifferential operators*, I. *Boundedness on spaces of Besov type*, J. Fac. Sci. Univ. Tokyo, 33 (1986), pp. 131–174.

SIAM J. MATH. ANAL.
Vol. 36, No. 2, pp. 463–478

ON THE EGUCHI–OKI–MATSUMURA EQUATION FOR PHASE SEPARATION IN ONE SPACE DIMENSION*

TAKAO HANADA†, NAOYUKI ISHIMURA‡, AND MASAAKI NAKAMURA§

Abstract. Eguchi–Oki–Matsumura equations are introduced to describe the dynamics of pattern formation that arises from phase separation in some binary alloys. The model extends the well-known Cahn–Hilliard equation and consists of coupled two functions; one is the local concentration and the other is the local degree of order. We show the existence of a solution, its asymptotic profile, and in part the structure of steady state solutions. Computational studies are also given.

Key words. pattern formation, phase separation, Eguchi–Oki–Matsumura model

AMS subject classifications. 80A30, 35K45

DOI. 10.1137/S0036141003400124

1. Introduction.

There has been much interest in the dynamics of pattern formation resulting from phase separation, which is commonly observed in many physical contexts; we recall, for example, certain binary alloys and polymer mixtures. Cahn and Hilliard [4], based on a continuum model in thermodynamics, made a phenomenological approach to explain such kinetics and derive the fourth-order partial differential equations (PDEs), known as the Cahn–Hilliard equation. Many studies have been performed on this equation and much progress has been achieved so far from various points of view. For more information and background materials, see [1], [2], [3], [6], [7], [8], [9], [11], [21], [22], [23], [24], [25] and the references therein.

Eguchi, Oki, and Matsumura [10], in an attempt to theoretically investigate such pattern formation, introduced a system of equations, referred to here as EOM equations. This motion law is derived from the first principles of thermodynamics of irreversible process under appropriate assumptions on the free energy, and it generalizes the formulation settled by Cahn and Hilliard. The EOM equation extends the Cahn–Hilliard equation and consists of coupled two phase fields, one the local concentration and the other the local degree of order. After performance of a suitable scaling of parameters, presented later, EOM equations in one space dimension, with which we are mainly concerned, are expressed as follows:

$$
\begin{cases}
u_t = -\varepsilon^2 u_{xxxx} + ((a+v^2)u)_{xx} & \text{in } 0<x<l, \quad t>0,\\
v_t = v_{xx} + (b-u^2-v^2)v & \text{in } 0<x<l, \quad t>0,\\
u_x = u_{xxx} = v_x = 0 & \text{at } x=0 \text{ and } l, \quad t>0,\\
u|_{t=0} = u_0, \quad v|_{t=0} = v_0 & \text{on } 0\le x\le l,
\end{cases}
\tag{1.1}
$$

*Received by the editors March 3, 2003; accepted for publication (in revised form) January 16, 2004; published electronically July 29, 2004. This work was partially supported by Grants-in-Aids for Scientific Research 10555023, 12640223, 13555021, and 13640206 from the Japan Ministry of Education, Science, Sports and Culture.

http://www.siam.org/journals/sima/36-2/40012.html

†Department of Mathematics, Chiba Institute of Technology, Narashino, Chiba 275-0023, Japan (hanada@pf.it-chiba.ac.jp).

‡Department of Mathematics, Graduate School of Economics, Hitotsubashi University, Kunitachi, Tokyo 186-8601, Japan (ishimura@math.hit-u.ac.jp).

§College of Science and Technology, Nihon University, Kanda-Surugadai, Tokyo 101-8308, Japan (nakamura@math.cst.nihon-u.ac.jp).

where $u = u(x,t)$ and $v = v(x,t)$ denote unknown functions related to the local concentration and the local degree of order, respectively. The total concentration of u is conserved under the evolution of (1.1). Namely, we have

$$\frac{1}{l}\int_0^l u(x,t)\,dx = m,$$

where m is a constant. Given initial data u_0, v_0 should satisfy required compatibility conditions:

$$(u_0)_x = (u_0)_{xxx} = (v_0)_x = 0 \qquad \text{at } x = 0,\, l, \quad \text{and } \frac{1}{l}\int_0^l u_0(x)\,dx = m.$$

Positive constants ε, a depend on the temperature, and $b \in \boldsymbol{R}$ is the principal parameter which increases from negative to positive as the temperature decreases from above to below the critical temperature. We focus our attention, however, on the case of positive b, since the negative b turns out to enjoy rather trivial behaviors.

As a special case of EOM equations, we observe that if $v \equiv 0$, then (1.1) reduces to

$$\begin{cases} u_t = -\varepsilon^2 u_{xxxx} + a u_{xx} & \text{in } 0 < x < l, \quad t > 0, \\ u_x = u_{xxx} = 0 & \text{at } x = 0 \text{ and } l, \quad t > 0, \\ (1/l)\int_0^l u\,dx = m. \end{cases}$$

This is the famous Cahn–Hilliard equation in its simplest form (if we especially allow $a < 0$). If we put $u \equiv m$ in (1.1), then we recover

$$\begin{cases} v_t = v_{xx} + (b - m^2 - v^2)v & \text{in } 0 < x < l, \quad t > 0, \\ v_x = 0 & \text{at } x = 0 \text{ and } l, \quad t > 0, \end{cases}$$

which is deduced from the Ginzburg–Landau theory for superconductivity.

Motivated partly by works concerning the Cahn–Hilliard equation [8], [11], [25], we expect that the solution $(u(x,t), v(x,t))$ for (1.1) converges as $t \to \infty$ to the solution $(u(x), v(x))$ for the steady state problem

$$\begin{cases} -\varepsilon^2 u_{xxxx} + ((a + v^2)u)_{xx} = 0 & \text{in } 0 < x < l, \\ v_{xx} + (b - u^2 - v^2)v = 0 & \text{in } 0 < x < l, \\ u_x = u_{xxx} = v_x = 0 & \text{at } x = 0 \text{ and } l, \\ (1/l)\int_0^l u\,dx = m. \end{cases} \tag{1.2}$$

We remark that (1.2) always has a solution $u \equiv m$ and $v \equiv 0$. If $b \le 0$, then it can be seen that this is the only solution to (1.2) by virtue of the maximum principle. If $b > m^2$, (1.2) has another solution $u \equiv m$ and $v \equiv \pm\sqrt{b - m^2}$. We call these solutions trivial. Solutions that are different from trivial ones will be called nontrivial solutions of the EOM equations; in other words, solution (u, v) to (1.2), both of which are not simultaneously constants, will be referred to as nontrivial solutions.

In this article, we are concerned with the local solvability, the asymptotic behavior of solutions to (1.1), and the structure of steady state solutions to (1.2). It is exhibited that the local degree of order v plays a key role in producing phase separation in the EOM model. Our main analytical achievements are summarized as follows.

THEOREM 1.1. *Suppose that* $u_0, v_0 \in H^2(0,l)$ *with* $(u_0)_x = (v_0)_x = 0$ *at* $x = 0, l$ *and* $(1/l)\int_0^l u_0\,dx = m$. *Then, for each* $T > 0$, *there exists a unique solution* (u,v) *to* (1.1) *such that*

$$u \in L^2((0,T);H^4(0,l)) \cap L^\infty([0,T);H^2(0,l)),$$
$$v \in L^2((0,T);H^2(0,l)) \cap L^\infty([0,T);H^1(0,l)).$$

For any initial data above, the solution (u,v) *converges as* $t \to \infty$ *to a solution of the steady state problem* (1.2).

There is at least one monotone nontrivial steady state solution of (1.2) *if we assign suitably large* b *and* m^2. *Moreover, for any integer* $k \geq 2$ *and for appropriately chosen large* b *and* m^2 *depending on* k, (1.2) *has at least one nonmonotone nontrivial steady state solution, each of whose derivatives changes sign exactly* $(k-1)$ *times.*

The values of b and m^2 stated in the theorem will be clarified in the course of proof.

For related results concerning EOM equations, see [17], [20].

We briefly outline the idea of proof. To obtain the local in time solution, a standard Galerkin method is employed. The global existence then follows by a priori estimates, with the aid of a Lyapunov functional; the free energy of the system serves as a Lyapunov functional, which is given by

$$F[u,v] := \int_0^l \left(\frac{\varepsilon^2}{2}u_x^2 + \frac{1}{2}v_x^2 + \frac{a}{2}u^2 + \frac{1}{4}v^4 - \frac{b}{2}v^2 + \frac{1}{2}u^2v^2\right) dx. \tag{1.3}$$

Note that $F[u,v]$ is well defined for (u,v) of the function class specified in Theorem 1.1. A direct calculation leads to

$$\frac{d}{dt}F[u,v](t) = -\int_0^l \{-\varepsilon^2 u_{xxx} + ((a+v^2)u)_x\}^2\,dx - \int_0^l v_t^2\,dx \leq 0 \tag{1.4}$$

for any solution (u,v) to (1.1).

Every solution is proved to converge to a steady state. Depending on the value of parameters b and m, EOM equations have various steady state solutions, which numerical investigation in section 5 clearly illustrates. Section 4 shows analytically the existence of nontrivial steady state solutions of EOM equations.

We quickly review the derivation of (1.1) for completeness of our exposition. Following [10], we begin with the total free energy:

$$F_{\mathrm{EOM}}[u,v] = \int_\Omega \left(\frac{H}{2}|\nabla u|^2 + \frac{K}{2}|\nabla v|^2 + f(u,v)\right) dx,$$

where $\Omega \subset \boldsymbol{R}^3$ represents a bounded domain and physical constants H, K mean the surface energy per unit area, which depend on the temperature. The function $f(u,v)$ stands for the density of the bulk free energy assumed to be given by

$$f(u,v) = \frac{a}{2}u^2 + \frac{1}{4}v^4 - \frac{b}{2}v^2 + \frac{g}{2}u^2v^2,$$

where positive constants a, g depend on the temperature and b denotes the principal parameter. Here we confine ourselves to considering the one-dimensional case and put $\Omega = (0,l)$. We make a scaling of variables. Define

$$\sqrt{\frac{g}{K}}u \to u, \quad \frac{1}{\sqrt{K}}v \to v, \quad \frac{a}{gK} \to a, \quad \frac{b}{K} \to b$$

and divide F_{EOM} by K^2. Then the functional (1.3) is discovered with $\varepsilon^2 = H/gK$. We remark that we have just rearranged the constants while retaining the role of the principal parameter b.

2. Existence of solutions. First we deal with the solvability of the problem (1.1). To establish the local in time existence, we implement a standard Galerkin approximation method.

Let $\mathcal{F}$ denote the complete orthonormal system in $L^2(0,l)$ with the even periodic boundary condition:

$$\mathcal{F} := \left\{ \frac{1}{\sqrt{l}}, \sqrt{\frac{2}{l}} \cos\frac{\pi x}{l}, \sqrt{\frac{2}{l}} \cos\frac{2\pi x}{l}, \dots, \sqrt{\frac{2}{l}} \cos\frac{n\pi x}{l}, \dots \right\}.$$

For every integer $N > 0$, let W_N be a linear space spanned by $\{1/\sqrt{l}, \sqrt{2/l}\cos(\pi x/l), \dots, \sqrt{2/l}\cos(N\pi x/l)\}$ and P_N denote the orthogonal projector in $L^2(0,l)$ onto W_N; namely,

$$P_N : L^2(0,l) \to W_N := \mathrm{Span} \left\{ \frac{1}{\sqrt{l}}, \sqrt{\frac{2}{l}} \cos\frac{\pi x}{l}, \dots, \sqrt{\frac{2}{l}} \cos\frac{N\pi x}{l} \right\}.$$

To be precise, for every even periodic $f \in L^2(0,l)$, we define

$$P_N f(x) := \frac{1}{\sqrt{l}} f^0 + \sum_{n=1}^{N} f^n \sqrt{\frac{2}{l}} \cos\frac{n\pi x}{l},$$

where

$$f^0 := \frac{1}{\sqrt{l}} \int_0^l f(x)\,dx \quad \text{and} \quad f^n := \int_0^l f(x) \sqrt{\frac{2}{l}} \cos\frac{n\pi x}{l}\,dx.$$

We are then looking for an approximate solution $(u_N(x,t), v_N(x,t))$ to (1.1) given by

$$\begin{aligned} u_N(x,t) &= m + \sum_{n=1}^{N} u^n(t) \sqrt{\frac{2}{l}} \cos\frac{n\pi x}{l} \\ &\quad (u^0(t) = \sqrt{l}m \text{ is used without possible confusion}), \\ v_N(x,t) &= \frac{1}{\sqrt{l}} v^0(t) + \sum_{n=1}^{N} v^n(t) \sqrt{\frac{2}{l}} \cos\frac{n\pi x}{l}. \end{aligned}$$

That is, u_N, v_N are interpreted as W_N-valued functions on $[0,T)$ for some $T > 0$, which satisfy

$$\begin{cases} (u_N)_t = -\varepsilon^2 (u_N)_{xxxx} + a(u_N)_{xx} + P_N((v_N)^2 u_N)_{xx} & \text{in } 0 < x < l,\ 0 < t < T, \\ (v_N)_t = (v_N)_{xx} + b v_N - P_N(((u_N)^2 + (v_N)^2) v_N) & \text{in } 0 < x < l,\ 0 < t < T, \\ u_N|_{t=0} = P_N u_0,\ v_N|_{t=0} = P_N v_0 & \text{on } 0 \le x \le l. \end{cases}$$

If we write up the equations for each component, then we discover

$$
\begin{aligned}
\frac{du^n(t)}{dt} &= -\varepsilon^2 \left(\frac{n\pi}{l}\right)^4 u^n(t) - a\left(\frac{n\pi}{l}\right)^2 u^n(t) \\
&\quad - \frac{1}{2l}\left(\frac{n\pi}{l}\right)^2 \sum_{G(n_1,n_2,n_3;n)}^{*} \iota_{n_1}\iota_{n_2}\iota_{n_3} v^{n_1}(t)v^{n_2}(t)u^{n_3}(t), \\
\frac{dv^n(t)}{dt} &= -\left(\frac{n\pi}{l}\right)^2 v^n(t) + bv^n(t) \\
&\quad - \frac{1}{2l} \sum_{G(n_1,n_2,n_3;n)}^{*} \iota_{n_1}\iota_{n_2}\iota_{n_3}\iota_n (u^{n_1}(t)u^{n_2}(t) + v^{n_1}(t)v^{n_2}(t))v^{n_3}(t) \\
&\qquad (\iota_0 := 2^{-1/2} \quad \text{and} \quad \iota_n := 1 \text{ for } n \geq 1), \\
u^n(0) &= \int_0^l u_0(x)\sqrt{\frac{2}{l}}\cos\frac{n\pi x}{l}\,dx \qquad \text{for } n \geq 1, \\
v^n(0) &= \begin{cases} (1/\sqrt{l})\int_0^l v_0(x)\,dx & \text{for } n = 0, \\ \int_0^l v_0(x)\sqrt{2/l}\cos(n\pi x/l)\,dx & \text{for } n \geq 1. \end{cases}
\end{aligned} \tag{2.1}
$$

Here we have defined

$$
\begin{aligned}
G(n_1,n_2,n_3;n) := \{0 \leq n_1, n_2, n_3 \leq N \,|\, & n_1 + n_2 + n_3 = \pm n \\
\text{or } \ n_1 - n_2 + n_3 = \pm n \quad & \text{or } n_1 + n_2 - n_3 = \pm n \quad \text{or } n_1 - n_2 - n_3 = \pm n\},
\end{aligned}
$$

and the summation $\sum^*$ indicates that the multiplicity is taken into account; for example, if $(n_1,n_2,n_3;n) = (1,1,1;1)$, then $\sum^* v^1(t)v^1(t)u^1(t) = 3v^1(t)v^1(t)u^1(t)$, since three equations in the definition of G are satisfied. If $(n_1,n_2,n_3;n) = (0,0,0;0)$, then the multiplication by eight is made.

The system of equations (2.1) has a unique solution on $[0,T_N)$ for some $T_N > 0$. The passage to the limit $N \to \infty$ is based on a priori estimates on (u_N, v_N). Here we refer to various a priori estimates established for the solution (u,v) to (1.1) in the next section, which are principally applicable as well to the truncated systems with necessary modifications. For instance, an upper bound for v_N is estimated as

$$
\|v_N(t)\|^2 := \int_0^l v_N(x,t)^2\,dx = \sum_{i=0}^{N} v^i(t)^2 \leq \limsup_{N\to\infty} \|v_N(t)\|^2 \leq \|v(t)\|^2 \leq 2bl,
$$

where the last inequality is expressed in Lemma 3.1. Note also that norm $\|\cdot\|$ is consistent with those defined there.

In addition, the time evolution of the Lyapunov functional $F[u_N, v_N](t)$ is calculated to be

$$
\begin{aligned}
&\frac{d}{dt}F[u_N,v_N](t) \\
&\quad = -\int_0^l \{-\varepsilon^2 (u_N)_{xxx} + (au_N + P_N(u_N(v_N)^2))_x\}^2\,dx - \int_0^l ((v_N)_t)^2\,dx \leq 0.
\end{aligned}
$$

After performing an integration with respect to t, we infer that $\|u_N(t)\|$, $\|(u_N)_x(t)\|$, $\|(v_N)_x(t)\|$ and therefore $\max|u_N(\cdot,t)|$ and $\max|v_N(\cdot,t)|$ are uniformly bounded by constants which depend on $H^1(0,l)$-norms of the initial data u_0 and v_0 but are independent of N.

We are thus able to let $N \to \infty$; in particular, we have $\liminf_{N\to\infty} T_N \geq T > 0$ for some $T > 0$. Uniform bounds of $H^1(0,l)$-norms enable us to repeat the local solvability procedure and continue the solution. We remark that the linear parts of (2.1) are good terms. In summary, our existence results are formulated as follows.

PROPOSITION 2.1. *Suppose that* $u_0, v_0 \in H^1(0,l)$ *with* $(u_0)_x = (v_0)_x = 0$ *at* $x = 0, l$ *and* $(1/l)\int_0^l u_0\,dx = m$. *Then, for each* $T > 0$, *there exists a unique solution* (u,v) *to* (1.1) *such that*

$$u \in L^2((0,T); H^3(0,l)) \cap L^\infty([0,T); H^1(0,l)),$$
$$v \in L^2((0,T); H^2(0,l)) \cap L^\infty([0,T); H^1(0,l)).$$

Further regularities claimed in Theorem 1.1 are standard consequences of a priori estimates depicted as lemmas in the subsequent section.

We next deal with the long-term behavior of the solution (u,v) to (1.1). This is a rather routine inference by virtue of the Lyapunov functional $F[u,v]$.

Since the solution (u,v) exists for any $T > 0$, integration of (1.4) with respect to t yields

$$\limsup_{t\to\infty} F[u,v](t) + \int_0^\infty d\tau \int_0^l (\{-\varepsilon^2 u_{xxx} + ((a+v^2)u)_x\}^2 + v_t^2)\,dx \leq F[u_0,v_0].$$

Taking into account that $F[u,v]$ is bounded below, we conclude that there is a sequence $t_1 < t_2 < \cdots < t_n < \cdots \to \infty$ for which

$$\begin{cases} (-\varepsilon^2 u_{xxx} + ((a+v^2)u)_x)(t_n) \to 0 \\ (v_{xx} + (b - u^2 - v^2)v)(t_n) \to 0 \end{cases} \qquad \text{as } n \to \infty.$$

That is, (u,v) tends to an element of the ω-limit set of (u_0,v_0), on which $F[u,v]$ is constant; namely, (u,v) converges to an equilibrium solution of the steady state problem (1.2).

3. A priori estimates. We turn our attention to some a priori estimates, which are needed to prove the existence and to determine the asymptotic profile of the solution (u,v) to (1.1). For brevity of presentation, we introduce the following function spaces:

$$E_T := \{(u,v) \in L^2((0,T); H^4(0,l)) \times L^2((0,T); H^2(0,l)) \mid u_x = u_{xxx} = v_x = 0 \text{ at } x = 0, l\},$$

$$E_0 := \left\{ (u_0,v_0) \in (H^2(0,l))^2 \,\middle|\, (u_0)_x = (v_0)_x = 0 \text{ at } x = 0, l,\ \frac{1}{l}\int_0^l u_0\,dx = m \right\},$$

where $T > 0$. The norms of $L^p(0,l)$ $(1 \leq p \leq \infty)$ are denoted by $\|\cdot\|_p$ and $\|\cdot\| := \|\cdot\|_2$. The inequalities $\|\cdot\|_p \leq l^{(q-p)/q}\|\cdot\|_q$ $(1 < p \leq q \leq \infty)$ are easily checked. Furthermore, C_0 stand for various constants depending only on the initial data and constants ε^2, a, b, which may differ from line to line. We understand that C_0 is independent of t.

First we provide a bound for v.

LEMMA 3.1. *There holds*

$$\|v(t)\|_\infty \leq \max\{\|v_0\|_\infty, \sqrt{b}\} \qquad \textit{for } 0 < t < T.$$

We recall once again that $b \geq 0$ is assumed.

Proof. Take an arbitrarily small $\delta > 0$ and suppose there exists first $0 \le t < T$, $0 \le x \le l$ such that

$$\|v_0\|_\infty \le \sqrt{b} + \delta \le v(x,t) = \max_{0\le y\le l} v(y,t).$$

The other case $v(x,t) = \min_{0\le y\le l} v(y,t) \le -\sqrt{b} - \delta \le -\|v_0\|$ proceeds similarly. We then find that

$$0 \le v_t(x,t) = v_{xx}(x,t) + (b - u(x,t)^2 - v(x,t)^2)v(x,t) \le (b - (\sqrt{b}+\delta)^2)v(x,t) < 0,$$

a contradiction. Since δ is arbitrary, we are done. □

We can say further that for any initial data $(u_0, v_0) \in E_0$ and for every $\delta > 0$, there exists T_δ such that $\|v(t)\|_\infty \le (1+\delta)\sqrt{b}$ for $t \ge T_\delta$. We may thus assume without loss of generality that $\|v_0\|_\infty \le \sqrt{2b}$ from the beginning. Note also that this especially leads to $\|v(t)\|_p \le l^{1/p}\sqrt{2b}$ for any $1 \le p \le \infty$.

LEMMA 3.2. *For any initial data $(u_0, v_0) \in E_0$, the solution $(u, v) \in E_T$ to (1.1) verifies*

$$\|u(t)\|,\ \|u_x(t)\|,\ \|v_x(t)\| \le C_0$$

for $0 < t < T$, and moreover

$$\|u(t)\|_\infty \le C_0.$$

Proof. We employ the Lyapunov functional introduced in (1.3). We assert that for $0 < t < T$

$$F[u,v](t) + \int_0^t d\tau \int_0^l (\{-\varepsilon^2 u_{xxx} + ((a+v^2)u)_x\}^2 + v_t^2)\, dx \le F[u_0, v_0],$$

from which the first three estimates hold. Thanks to the inequality

$$\|w\|_\infty \le \frac{1}{\sqrt{l}}\|w\| + \sqrt{l}\|w_x\| \qquad \text{for } w \in H^1(0,l),$$

we arrive at the last estimate. □

LEMMA 3.3. *It follows that for any $0 \le t \le s \le T$*

$$\int_t^s (\|u_{xx}(\tau)\|^2 + \|u_x(\tau)\|^2 + \|v_x(\tau)\|^2)\, d\tau \le C_0(1 + (s-t)).$$

Proof. Multiplying the first and the second equation of (1.1) by u and v, respectively, and integrating by parts, we infer that

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}\|u(t)\|^2 + \varepsilon^2\|u_{xx}(t)\|^2 &= -\int_0^l ((a+v^2)u)_x u_x\, dx \\
&= -a\|u_x(t)\|^2 - \|(u_x v)(t)\|^2 - 2\int_0^l u u_x v v_x\, dx \\
&\le -a\|u_x(t)\|^2 + \|(uv_x)(t)\|^2 \le -a\|u_x(t)\|^2 + C_0^2\|v_x(t)\|^2, \\
\frac{1}{2}\frac{d}{dt}\|v(t)\|^2 + \|v_x(t)\|^2 &= \int_0^l (b - u^2 - v^2)v^2\, dx \le b\|v(t)\|^2 \le 2lb^2,
\end{aligned}$$

where the use of $\|u(t)\|_\infty \le C_0$ and $\|v(t)\|_\infty \le \sqrt{2b}$ is made. We compute

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}(\|u(t)\|^2 + (C_0^2+1)\|v(t)\|^2) + \varepsilon^2\|u_{xx}(t)\|^2 + a\|u_x(t)\|^2 + \|v_x(t)\|^2 \\ \le b^2(C_0^2+1) =: C_0,\end{aligned}$$

and hence we have

$$\begin{aligned}\|u(s)\|^2 + (C_0^2+1)\|v(s)\|^2 + 2\int_t^s (\varepsilon^2\|u_{xx}(\tau)\|^2 + a\|u_x(\tau)\|^2 + \|v_x(\tau)\|^2)\,d\tau \\ \le \|u(t)\|^2 + (C_0^2+1)\|v(t)\|^2 + C_0(s-t) =: C_0(1+(s-t)),\end{aligned}$$

from where the lemma follows. □

Next we are going to derive higher derivative estimates. Our intention is to achieve bounds represented explicitly in terms of t.

LEMMA 3.4. *There holds for any* $0 \le t \le s \le T$,

$$\int_t^s (\|u_{xxx}(\tau)\|^2 + \|v_{xx}(\tau)\|^2)\,d\tau \le C_0(1+(s-t)).$$

Proof. This time we multiply the first and the second equation of (1.1) by u_{xx} and v_{xx}, respectively; we deduce that

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\|u_x(t)\|^2 + \varepsilon^2\|u_{xxx}(t)\|^2 \\ &\quad = -\int_0^l ((a+v^2)u)_{xx}u_{xx}\,dx = -a\|u_{xx}(t)\|^2 + \int_0^l (uv^2)_x u_{xxx}\,dx \\ &\quad \le -a\|u_{xx}(t)\|^2 + \frac{\varepsilon^2}{2}\|u_{xxx}(t)\|^2 + \frac{1}{2\varepsilon^2}\int_0^l (u_x v^2 + 2uvv_x)^2\,dx \\ &\quad \le -a\|u_{xx}(t)\|^2 + \frac{\varepsilon^2}{2}\|u_{xxx}(t)\|^2 + C_0(\|u_x(t)\|^2 + \|v_x(t)\|^2), \\ &\frac{1}{2}\frac{d}{dt}\|v_x(t)\|^2 + \|v_{xx}(t)\|^2 = -\int_0^l (b-u^2-v^2)vv_{xx}\,dx \\ &\quad = \int_0^l \{(b-u^2-v^2)v_x^2\,dx - (u^2+v^2)_x vv_x\}\,dx \\ &\quad = \int_0^l \{(b-3v^2)v_x^2 - u^2v_x^2 - 2uu_xvv_x\}\,dx \\ &\quad \le \int_0^l \{(b-3v^2)v_x^2 + u_x^2v^2\}\,dx \le 2b(\|u_x(t)\|^2 + \|v_x(t)\|^2).\end{aligned}$$

Adding these inequalities, we have

$$\frac{1}{2}\frac{d}{dt}(\|u_x(t)\|^2 + \|v_x(t)\|^2) + \frac{\varepsilon^2}{2}\|u_{xxx}(t)\|^2 + \|v_{xx}(t)\|^2 \le C_0(\|u_x(t)\|^2 + \|v_x(t)\|^2).$$

An integration with respect to t combined with Lemmas 3.2 and 3.3 finishes the proof of Lemma 3.4. □

Finally we formalize the next lemma.

LEMMA 3.5. *There hold for any* $0 \le t \le s \le T$

$$\|u_{xx}(t)\|^2 \le C_0(1+t), \quad \int_t^s \|u_{xxxx}(\tau)\|^2\,d\tau \le C_0(1+(s-t)).$$

Proof. We multiply the equation of u by u_{xxxx} and integrate by parts to obtain

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|u_{xx}(t)\|^2 + \varepsilon^2\|u_{xxxx}(t)\|^2 &= \int_0^l (au+uv^2)_{xx}u_{xxxx}\,dx\\ &\le -a\|u_{xxx}(t)\|^2 + \frac{\varepsilon^2}{2}\|u_{xxxx}(t)\|^2 + \frac{1}{2\varepsilon^2}\int_0^l ((uv^2)_{xx})^2\,dx\\ &\le -a\|u_{xxx}(t)\|^2 + \frac{\varepsilon^2}{2}\|u_{xxxx}(t)\|^2\\ &\qquad + C_0(\|u_{xx}(t)\|^2 + \|v_{xx}(t)\|^2 + \|(u_xv_x)(t)\|^2 + \|v_x(t)^2\|^2).\end{aligned}$$

To handle the last term, we appeal to the inequality

$$\|u_x(t)\|_\infty \le \sqrt{l}\|u_{xx}(t)\|, \quad \|v_x(t)\|_\infty \le \sqrt{l}\|v_{xx}(t)\|,$$

which is valid in light of $u_x = v_x = 0$ at the boundary; it then follows that

$$\begin{aligned}\|(u_xv_x)(t)\|^2 \le l\|u_{xx}(t)\|^2\|v_x(t)\|^2 \le C_0\|u_{xx}(t)\|^2,\\ \|v_x(t)^2\|^2 \le C_0\|v_{xx}(t)\|^2.\end{aligned}$$

To sum up, we have accomplished

$$\frac{1}{2}\frac{d}{dt}\|u_{xx}(t)\|^2 + \frac{\varepsilon^2}{2}\|u_{xxxx}(t)\|^2 + a\|u_{xxx}(t)\|^2 \le C_0(\|u_{xx}(t)\|^2 + \|v_{xx}(t)\|^2).$$

The integral estimates of the previous two lemmas now imply the conclusion we desired. □

4. Structure of steady solutions. In this section, the structure of steady state solutions to EOM equations is analyzed; we want to seek for solutions $u = u(x)$ and $v = v(x)$ which verify (1.2).

Numerical investigation presented in section 5 strongly indicates that there really exist nontrivial steady state solutions to EOM equations for certain parameter values. We recall that nontrivial solutions are defined as those for (1.2) other than trivial solutions mentioned in section 1; namely, they are solutions (u, v) to (1.2), both of which are not simultaneously constants. The aim of this section is to confirm analytically these numerical observations. Our results are as follows and extend our previous establishments [15].

PROPOSITION 4.1. *For suitably assigned large b and m^2, there exists at least one monotone nontrivial steady state solution for EOM equations. Furthermore, for any integer $k \ge 2$ and for appropriately large b and m^2 depending on k, EOM equations have nonmonotone nontrivial steady state solutions, each of whose derivatives changes sign exactly $(k-1)$ times.*

The large values of b and m^2 stated in Proposition 4.1 can be computed explicitly, for which we do not go into detail. We also remark that our result should be compared to [5], where the fact is described that nonmonotonic functions cannot be a local minimizer for a single Cahn–Hilliard-type functional.

The proof is carried out from the variational point of view and divided into several steps. First we notice that although (1.2) is a system of equations of fourth order, it has a second-order variational structure; the solution to (1.2) is given by the critical point of a functional (1.3) among the function space

$$\mathcal{A} := \left\{ (u,v) \in (H^1(0,l))^2 \;\middle|\; \frac{1}{l}\int_0^l u\,dx = m \right\}.$$

We have the next lemma, whose proof is given in a similar way as in Lemma 3.1 of [25], and we may safely omit it.

LEMMA 4.2. *Problem* (1.2) *is equivalent to the problem of finding the critical points of functional* $F[u,v]$ *defined by* (1.3) *over* $\mathcal{A}$.

Step 1. Existence of nontrivial global minimizer.

We immediately obtain

$$\begin{aligned}\frac{1}{l}F[m,0] &= \frac{a}{2}m^2,\\ \frac{1}{l}F[m,\pm\sqrt{b-m^2}] &= \frac{a}{2}m^2-\frac{1}{4}(b-m^2)^2 \quad \text{if } b>m^2.\end{aligned}$$

Since F is bounded below on $\mathcal{A}$, our task of constructing at least one nontrivial steady state solution is to find a test function $(u,v)\in\mathcal{A}$ such that

$$\frac{1}{l}F[u,v] < \frac{a}{2}m^2-\frac{1}{4}(b-m^2)^2. \tag{4.1}$$

The minimization procedure then works well to produce at least one nontrivial steady state solution, which is global minimizer of (1.3) over $\mathcal{A}$.

To accomplish this, we prepare

$$\begin{aligned}u(x) &= m-\delta\cos\frac{\pi x}{l},\\ v(x) &= \pm\sqrt{b-\left(m-\delta\cos\frac{\pi x}{l}\right)^2},\end{aligned} \tag{4.2}$$

where $\delta>0$ is a parameter and we assign $b>(m+\delta)^2$. Clearly $(u,v)\in\mathcal{A}$ and we compute

$$\begin{aligned}\frac{1}{l}F[u,v] = \frac{a}{2}\,m^2 &-\frac{1}{4}(b-m^2)^2+(\varepsilon^2(\pi/l)^2+a)\frac{\delta^2}{4}\\ &+\frac{\delta^2\pi^2}{2l^2}\int_0^l\frac{(m-\delta\cos(\pi x/l))^2\sin^2(\pi x/l)}{b-(m-\delta\cos(\pi x/l))^2}\,dx-\frac{m^2\delta^2}{2}-\frac{3\delta^4}{32}+\frac{\delta^2}{4}(b-m^2).\end{aligned}$$

If we further set $b=2m^2$ and $\delta=m/4$, then we infer that

$$\begin{aligned}\frac{1}{l}F[u,v] = \frac{1}{l}F[m,\pm\sqrt{b-m^2}] &-\frac{131}{2^{13}}m^4\\ &+\left(\frac{\varepsilon^2(\pi/l)^2+a}{64}+\frac{\pi^2}{32l^2}\int_0^l\frac{(4-\cos(\pi x/l))^2\sin^2(\pi x/l)}{32-(4-\cos(\pi x/l))^2}\,dx\right)m^2,\end{aligned}$$

from which we conclude that (4.1) holds, taking larger m if necessary.

Step 2. Monotonicity of the global minimizer.

Our intention is to show the monotonicity of the global minimizer (u,v) obtained in Step 1. Since there is a cross-term $\int_0^l 2^{-1}u^2v^2\,dx$ in the functional $F[u,v]$, a simple direct application of usual rearrangements is insufficient. We have, however, the next lemma.

LEMMA 4.3. *Suppose* $f,g\in H^1(0,l)$. *Let* f^i *and* g^d *denote the monotone increasing and decreasing rearrangement of* f *and* g, *respectively. Then*

$$\int_0^l f^i g^d\,dx \le \int_0^l fg\,dx,$$

where the equality holds if and only if $(f,g) \equiv (f^i, g^d)$ *or* (f^d, g^i).

Suppose additionally f, g *are both nonnegative (or nonpositive). Then*

$$\int_0^l (f^i)^2 (g^d)^2 \, dx \le \int_0^l f^2 g^2 \, dx,$$

where the equality holds if and only if $(f,g) \equiv (f^i, g^d)$ *or* (f^d, g^i). *If the sign of* (f,g) *is opposite, then* (f^i, g^i) *or* (f^d, g^d) *brings the same conclusion.*

We recall that the monotone increasing rearrangement f^i of $f \in H^1(0,l)$ is defined as follows. For $c \in \boldsymbol{R}$, put

$$I_c := \{x \in (0,l) \,|\, f(x) \ge c\},$$
$$I_c^i := \begin{cases} \{l - |I_c| \le x \le l\} & \text{if } I_c \ne \emptyset, \\ \emptyset & \text{if } I_c = \emptyset, \end{cases}$$

where $|I_c|$ denotes the Lebesgue measure of the interval I_c. We then have

$$f^i(x) = \sup\{c \in \boldsymbol{R} \,|\, x \in I_c^i\} \qquad \text{for } 0 \le x \le l.$$

The monotone decreasing rearrangement is defined similarly.

The next properties of rearrangements are well known [19]:

$$\begin{aligned} &\int_0^l (f^i)_x^2 \, dx \le \int_0^l f_x^2 \, dx, \\ &\int_0^l (f^i)^p \, dx = \int_0^l f^p \, dx \quad \text{if } f \ge 0,\ p > 0. \end{aligned} \tag{4.3}$$

The proof of our lemma is rather standard. Approximating f by step functions, we see that it suffices to deduce the lemma in the case of sequences, which is known to be true [18]. We omit the details.

Now we return to our problem. Taking account of (4.3) and Lemma 4.3, we assert that the monotonicity of the global minimizer (u,v) follows at once if (u,v) are both simultaneously taken to be nonnegative (or nonpositive). Indeed, suppose (u,v) are nonnegative. Performing the monotone increasing and decreasing rearrangement for u and v, respectively, we derive that

$$F[u^i, v^d] \le F[u,v] = \min_{\mathcal{A}} F,$$

from which we establish that $(u,v) \equiv (u^i, v^d)$ or (u^d, v^i).

It remains to prove that (u,v) are both taken to be nonnegative (or nonpositive). Suppose $m = l^{-1} \int_0^l u \, dx > 0$. Then $|\{x \,|\, u(x) > 0\}| > 0$ and we note that u solves

$$\begin{aligned} -\varepsilon^2 u_{xx} + (a + v^2)u &= \beta \qquad \text{in } 0 < x < l, \\ u_x &= 0 \qquad \text{at } x = 0 \text{ and } l \end{aligned}$$

for some constant $\beta \in \boldsymbol{R}$. Since $|\{x \,|\, u(x) > 0\}| > 0$, the sign of constant β must be positive, which certainly holds at the maximum point in $\{x \,|\, u(x) > 0\}$. We recall that $a > 0$. If there happens $|\{x \,|\, u(x) < 0\}| > 0$, then a similar reasoning as above applied to the minimum point of u (< 0) implies $\beta < 0$, which is a contradiction.

We therefore conclude that u has the same sign as m. If we further put $v^+(x) = \max\{v(x), -v(x)\} \geq 0$, then we observe

$$F[u, v^+] = F[u, v] = \min_{\mathcal{A}} F,$$

from which we find that (u, v) are nonnegative.

Step 3. Existence of multi-bump solutions.

Now, utilizing this monotone nontrivial solution, for every integer $k \geq 2$, we construct nontrivial solutions (u^k, v^k) to (1.2), whose derivatives u^k_x and v^k_x change sign exactly $(k-1)$ times, respectively. To obtain these solutions, we divide the interval $[0, l]$ into k subpieces $[(i-1)l/k, il/k]$ $(i = 1, 2, \ldots, k)$. On the first subinterval $[0, l/k]$, the minimization procedure formulated in Step 1 works if b and m^2 are suitably large enough and there exists at least one monotone nontrivial solution (u^k, v^k) to the equations of (1.2) with $u^k_x = v^k_x = 0$ at $x = 0$ and l/k. These are monotone functions. We define inductively on $i = 1, 2, \ldots, k$ that

$$u^k(x) = \begin{cases} u^k(x - ((i-1)l/k)), & i \text{ is odd}, \\ u^k(-x + (il/k)), & i \text{ is even}, \end{cases}$$

on each subintervals $[(i-1)l/k, il/k]$. v^k is extended similarly onto $[0, l]$. Such constructed (u^k, v^k) clearly solves (1.2) and fulfills our requirements.

The proof of Proposition 4.1 is finally completed. □

5. Computational study. In this section we treat the computational research for EOM equations. The detailed exposition is given in another work [16]; here we just incorporate several results. Note that our discretization scheme is motivated partially by [12], [13], [14], whose principal target is to cultivate a stable reasonable numerical scheme for computing the Cahn–Hilliard equation.

Let $x_k = k\Delta x$ $(k = 0, 1, \ldots, n)$ with $\Delta x = l/n$. The discretized free energy $F^n[U, V]$ for the approximations (U_k, V_k) of $(u(x_k, t), v(x_k, t))$ is expressed as

$$F^n[U, V] = \sum\nolimits_t \left(\frac{\varepsilon^2}{4}((\nabla_+ U_k)^2 + (\nabla_- U_k)^2) + \frac{1}{4}((\nabla_+ V_k)^2 + (\nabla_- V_k)^2) \right.$$
$$\left. + \frac{a}{2}U_k^2 + \frac{1}{4}V_k^4 - \frac{b}{2}V_k^2 + \frac{1}{2}U_k^2 V_k^2 \right) \Delta x.$$

Here ∇_+ and ∇_- denote the forward and backward difference in x, respectively:

$$\nabla_+ U_k = \frac{1}{\Delta x}(U_{k+1} - U_k), \quad \nabla_- U_k = \frac{1}{\Delta x}(U_k - U_{k-1}).$$

Furthermore, $\sum_t$ represents the trapezoidal summation formula defined by

$$\sum\nolimits_t U_k^2 = \frac{1}{2}U_0^2 + \sum_{k=1}^{n-1} U_k^2 + \frac{1}{2}U_n^2.$$

Now, for the approximations $(\bar{U}_k, \bar{V}_k)$ of $(u(x_k, t + \Delta t), v(x_k, t + \Delta t))$, we mainly adopt the implicit scheme as follows:

$$\frac{\bar{U}_k - U_k}{\Delta t} = \nabla^2 \left(-\varepsilon^2 \nabla^2 \frac{\bar{U}_k + U_k}{2} + a\frac{\bar{U}_k + U_k}{2} + \frac{\bar{U}_k \bar{V}_k + U_k V_k}{2} \frac{\bar{V}_k + V_k}{2} \right),$$
$$\frac{\bar{V}_k - V_k}{\Delta t} = \nabla^2 \frac{\bar{V}_k + V_k}{2} + b\frac{\bar{V}_k + V_k}{2} - \frac{\bar{U}_k \bar{V}_k + U_k V_k}{2} \frac{\bar{U}_k + U_k}{2} - \frac{\bar{V}_k{}^2 + V_k^2}{2} \frac{\bar{V}_k + V_k}{2},$$

FIG. 5.1. *Convergence to a monotone steady state solution.*

where $\nabla^2 := \nabla_+\nabla_-$ stands for the second-order central difference in x. With this implicit scheme, we deduce that the discretized total concentration is conserved and the discretized free energy is decreasing:

$$\sum\nolimits_t \bar{U}_k = \sum\nolimits_t U_k, \qquad F^n[\bar{U}, \bar{V}] \le F^n[U, V].$$

We supplement our solver, however, by the explicit scheme, since our current problem is in a sense stable by virtue that $a > 0$ and the implicit one is a little complicated to implement.

$$\frac{\bar{U}_k - U_k}{\Delta t} = -\varepsilon^2\nabla^4 U_k + \nabla^2((a + V_k^2)U_k),$$
$$\frac{\bar{V}_k - V_k}{\Delta t} = \nabla^2 V_k + (b - U_k^2 - V_k^2)V_k.$$

In this case, the dissipation of the free energy holds only approximately; nevertheless, it is whole enough for our purposes.

The discretized boundary conditions should be fixed as

$$\begin{aligned}
&U_{-1} = U_1, \quad U_{n-1} = U_{n+1} \quad \text{in place of } u_x = 0 \text{ at } x = 0 \text{ and } l,\\
&V_{-1} = V_1, \quad V_{n-1} = V_{n+1} \quad \text{in place of } v_x = 0 \text{ at } x = 0 \text{ and } l,\\
&U_{-2} = U_2, \quad U_{n-2} = U_{n+2} \quad \text{in place of } u_{xxx} = 0 \text{ at } x = 0 \text{ and } l.
\end{aligned}$$

We focus our interest on the question of whether the variety of steady state solutions exists. In the following, constants are taken to be

$$l = 1, \quad \varepsilon = 1, \quad a = m = \frac{1}{4}.$$

Several steady solutions are illustrated in the following figures.

Figure 5.1 depicts the convergence of a solution (u, v) for (1.1) to a monotone steady state solution. We set $b = 16/25$ and as initial function we employ

$$u_0(x) = m,$$
$$v_0(x) = \sqrt{b - m^2} - \frac{1}{1000}\cos(\pi x).$$

The computation is implemented under the mesh size $\Delta x = 1/64$ and $\Delta t = 1/256$ up to the time interval $0 \le t \le 4096$.

The monotone steady state solution corresponds to the case $k = 1$ in Theorem 1.1. It is numerically unstable with respect to the perturbation on initial data.

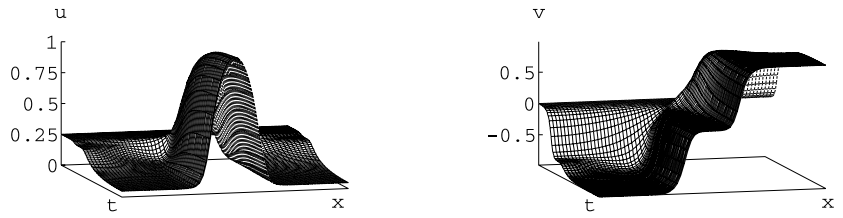

FIG. 5.2. *Convergence to a nonmonotone steady state solution.*

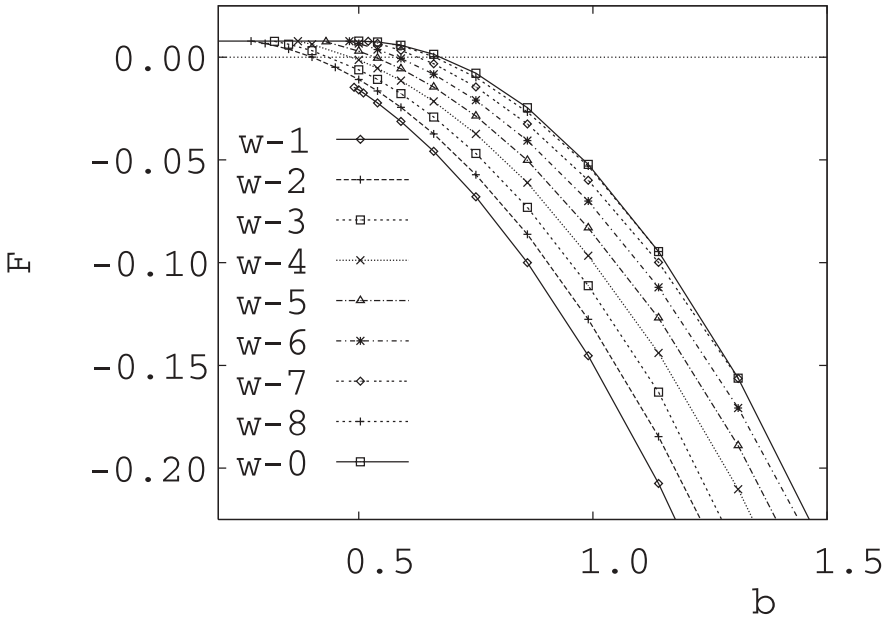

FIG. 5.3. *Energy diagram of steady state solutions.*

Figure 5.2, on the other hand, illustrates the convergence to a nonmonotone steady state solution. We set $b = 0.99$, and as initial functions we take

$$u_0(x) = m,$$
$$v_0(x) = -\frac{1}{1000}\cos(\pi x).$$

The implementation data are the same as those of Figure 5.1, performed during the time interval $0 \le t \le 128$.

The limiting function is related to the case $k = 2$ in Theorem 1.1; however, the function v is monotone increasing in Figure 5.2. This apparent discrepancy is easily reconciled because the sign of v is irrespective to the problem. We hasten to remark that the nonmonotone steady solution, which is constructed in Theorem 1.1 with $k = 2$, also is numerically realized.

Finally, we exhibit the energy diagram of various steady state solutions in Figure 5.3. Here the notation $w{-}i$ $(i = 0, 1, \ldots, 8)$ means the steady state solution to

(1.1), which is akin to the one with $k = i + 1$ described in Theorem 1.1.

It is observed that the energy of the monotone steady state solution dominates others. The detailed investigation, including the problem of the stability, will be deferred to another work.

Acknowledgments. We are grateful to Professor Hiroshi Fujita for his interest in this research. Thanks are also due to anonymous referees for pointing out mistakes in the first version of the manuscript and for valuable comments. In addition, the discussion in the Creation and Sustenance of Diversity workshop held at the International Institute for Advanced Study, Kyoto, was very fruitful.

REFERENCES

[1] N. D. Alikakos, P. W. Bates, and G. Fusco, *Slow motion for the Cahn-Hilliard equation in one space dimension*, J. Differential Equations, 90 (1991), pp. 81–135.

[2] J. F. Blowey and C. M. Elliott, *The Cahn-Hilliard gradient theory for phase separation with non-smooth free energy, Part* I: *Mathematical analysis*, European J. Appl. Math., 2 (1991), pp. 233–279.

[3] J. F. Blowey and C. M. Elliott, *The Cahn-Hilliard gradient theory for phase separation with non-smooth free energy, Part* II: *Numerical analysis*, European J. Appl. Math., 3 (1992), pp. 147–179.

[4] J. W. Cahn and J. E. Hilliard, *Free energy of a nonuniform system,* I*, Interfacial free energy*, J. Chem. Phys., 28 (1958), pp. 258–267.

[5] J. Carr, M. E. Gurtin, and M. Slemrod, *Structured phase transitions on a finite interval*, Arch. Ration. Mech. Anal., 86 (1984), pp. 317–351.

[6] J. Carr and R. Pego, *Metastable patterns in solutions of* $u_t = \varepsilon^2 u_{xx} - f(u)$, Comm. Pure Appl. Math., 42 (1989), pp. 523–576.

[7] X. Chen, *Global asymptotic limit of solutions of the Cahn-Hilliard equation*, J. Differential Geom., 44 (1996), pp. 262–311.

[8] C. M. Elliott and D. A. French, *Numerical studies of the Cahn-Hilliard equation for phase separation*, IMA J. Appl. Math., 38 (1987), pp. 97–128.

[9] C. M. Elliott and H. Garcke, *On the Cahn–Hilliard equation with degenerate mobility*, SIAM J. Math. Anal., 27 (1996), pp. 404–423.

[10] T. Eguchi, K. Oki, and S. Matsumura, *Kinetics of ordering with phase separation*, in Phase Transformations in Solids, Mat. Res. Soc. Symp. Proc. 21, T. Tsakalakos, ed., Elsevier, New York, 1984, pp. 589–594.

[11] C. M. Elliott and S. Zheng, *On the Cahn-Hilliard equation*, Arch. Ration. Mech. Anal., 96 (1986), pp. 339–357.

[12] D. Furihata, *Finite difference schemes for* $\partial u/\partial t = (\partial/\partial)^\alpha \delta G/\partial u$ *that inherit energy conservation or dissipation property*, J. Comput. Phys., 156 (1999), pp. 181–205.

[13] D. Furihata and T. Matsuo, *A stable, convergent, conservative and linear finite difference scheme for the Cahn-Hilliard equation*, Japan J. Indust. Appl. Math., 20 (2003), pp. 65–85.

[14] D. Furihata and M. Mori, *A stable finite difference scheme for the Cahn-Hilliard equation based on a Lyapunov functional*, Z. Angew. Math. Mech., 76 (1996), pp. 405–406.

[15] T. Hanada, N. Ishimura, and M. A. Nakamura, *Note on steady solutions of the Eguchi-Oki-Matsumura equation*, Proc. Japan Acad. Ser. A Math. Sci., 76 (2000), pp. 146–148.

[16] T. Hanada, N. Ishimura, and M. A. Nakamura, *Stable finite difference scheme for a model equation of phase separation*, Appl. Math. Comput., 151 (2004), pp. 95–104.

[17] T. Hanada, M. A. Nakamura, and C. Shima, *On Eguchi-Oki-Matsumura equations*, in Advances in Numerical Mathematics, Proceedings of the Fourth Japan-China Joint Seminar on Numerical Mathematics, H. Kawarada, M. A. Nakamura, and Z.-C. Shi, eds., Gakuto, Tokyo, 1999, pp. 213–222.

[18] G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, 2nd ed., Cambridge University Press, Cambridge, UK, 1952.

[19] B. Kawohl, *Rearrangements and Convexity of Level Sets in PDE*, Lecture Notes in Math. 1150, Springer-Verlag, New York, 1985.

[20] M. Kurokiba, N. Tanaka, and A. Tani, *Existence of solution for Eguchi-Oki-Matsumura equation describing phase separation and order-disorder transition in binary alloys*, J. Math. Anal. Appl., 272 (2002), pp. 448–457.

[21] A. Novick-Cohen, *Energy methods for the Cahn-Hilliard equation*, Quart. Appl. Math., 46 (1988), pp. 681–690.
[22] A. Novick-Cohen and L. A. Peletier, *The steady states of the one-dimensional Cahn-Hilliard equation*, Appl. Math. Lett., 5 (3) (1992), pp. 45–46.
[23] A. Novick-Cohen and L. A. Segel, *Nonlinear aspects of the Cahn-Hilliard equation*, Phys. D, 10 (1984), pp. 277–298.
[24] R. Temam, *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, Appl. Math. Sci. 68, Springer-Verlag, New York, 1988.
[25] S. Zheng, *Asymptotic behavior of solution to the Cahn-Hilliard equation*, Appl. Anal., 23 (1986), pp. 165–184.

SIAM J. MATH. ANAL.
Vol. 36, No. 2, pp. 479–497

STABLE STATIONARY PATTERNS AND INTERFACES ARISING IN REACTION-DIFFUSION SYSTEMS*

YOSHIHITO OSHITA†

Abstract. We study reaction-diffusion systems with FitzHugh–Nagumo-type nonlinearity. We consider the rich structures of stable stationary solutions for two different parameter scalings with the corresponding limiting problems. We study the complex phase separation patterns and derive the stationary interface equation for the limiting problems.

Key words. sharp interface, microscopic structure, Young measure

AMS subject classifications. 35J50, 35K57, 82B24

DOI. 10.1137/S0036141002406722

1. Introduction.

In this paper, we study the singular limiting behavior of some stable steady states of a class of reaction-diffusion systems:

$$(E)^{\mu}_{\varepsilon,D} \qquad \begin{cases} u_t = \varepsilon^2 \Delta u + f(u) - \dfrac{\varepsilon}{\mu} v & \text{in } \Omega \times \mathbb{R}^+, \\ \tau v_t = D\Delta v + u - m - \gamma v & \text{in } \Omega \times \mathbb{R}^+, \\ \dfrac{\partial u}{\partial n} = \dfrac{\partial v}{\partial n} = 0 & \text{on } \partial\Omega \times \mathbb{R}^+, \end{cases}$$

where $\Omega \subset \mathbb{R}^N$ $(N \geq 2)$ is a bounded domain with the smooth boundary $\partial\Omega$; $\partial/\partial n$ is the normal derivative on $\partial\Omega$; $\varepsilon, \mu \ll 1$ and $D \gg 1$ are positive parameters; γ and τ are positive constants; $f(u) = -W'(u)$ ($W \in C^2(\mathbb{R})$ is a double-well potential such that $W(\pm 1) = 0$ and $W(s) > 0$ for $s \neq \pm 1$); and m is a constant between two global minima of W, namely,

$$m \in (-1, 1). \tag{A1}$$

The system $(E)^{\mu}_{\varepsilon,D}$ describes, for instance, wave propagation in excitable media such as a Belousov–Zhabotinskii reaction, pattern formation in population genetics, propagation of signals along a nerve axon or cardiac tissue, etc. They are referred to as FitzHugh–Nagumo equations and have been studied extensively. (See, for instance, [3, 4, 5, 10, 15] and the references therein.)

We construct stable stationary solutions of $(E)^{\mu}_{\varepsilon,D}$ and study a behavior of them in two cases:

(i) $\varepsilon \to 0$ for fixed D, μ and

(ii) $\varepsilon \to 0, D \to \infty$ for fixed μ.

Although the Allen–Cahn equation

$$u_t = \Delta u + f(u)$$

with homogeneous Neumann data does not have nonconstant stable stationary solutions, $(E)^{\mu}_{\varepsilon,D}$ provides a rich structure, that is, a formation of nontrivial stationary

*Received by the editors May 2, 2002; accepted for publication (in revised form) January 16, 2004; published electronically July 29, 2004. This work was partially supported by Research Fellowships of the Japan Society for the Promotion of Science.

http://www.siam.org/journals/sima/36-2/40672.html

†Graduate School of Mathematical Sciences, University of Tokyo, 3-8-1, Komaba, Meguro-ku, Tokyo 153-8914, Japan (oshita@ms.u-tokyo.ac.jp).

patterns. We construct stationary solutions of $(E)^{\mu}_{\varepsilon,D}$ by using variational methods although $(E)^{\mu}_{\varepsilon,D}$ is not a gradient flow associated to some functional. For a proof of stability of such stationary solutions, we can apply the result of [16]. This paper is mostly concerned with variational problems for functionals with parameters ε, D, and μ and asymptotic behavior of minimizers.

Here we compare our case with one in [16]. In [16], it was shown that there are stable stationary solutions (u, v) such that the total variation of u is very large when $\varepsilon \sim \mu \ll 1$ and D is of order 1. On the other hand, we see that in both (i) and (ii), there is a limiting variational problem with a positive parameter μ. Then we study the μ-parametrized limiting functionals in the limit $\mu \to 0$, and we see that, in the case of (i), minimizers of the μ-parametrized limiting functional have the large total variation when μ is small.

Finally we derive the Euler–Lagrange equation of the singular limiting problem and it turns out to be a balance condition on the interface between the local effects of curvature and the effect of the nonlocal term. We will see that, in the case of (ii), the above balance condition on the interface coincides with the result of [5]. (See Remark 2.3 below.)

We make the following conditions on f:

(f1) $f \in C^1(\mathbb{R})$.

(f2) f has just three zeros, $f(\pm 1) = f(a) = 0$, $a \in (-1, 1)$, and satisfies $f'(\pm 1) < 0$, $f'(a) > 0$.

(f3) $\int_{-1}^{1} f(v)\, dv = 0$.

(f4) f has the polynomial growth at infinity, that is, there exist constants $s \in [1, \infty)$ and $c_1, c_2, R > 0$ such that

$$c_1|u|^s \le |f(u)| \le c_2|u|^s \quad \text{for all } |u| \ge R.$$

We use the following notation. $(\cdot, \cdot)$ denotes L^2-inner product, $\langle u \rangle$ is the mean value of $u \in L^1(\Omega)$ on Ω,

$$\int_\Omega |Du| := \sup \left\{ \int_\Omega \sum_{i=1}^{N} u D_i g_i \, dx \,;\, g = (g_1, \dots, g_N) \in C_0^1(\Omega, \mathbb{R}^N), |g| \le 1 \right\}$$

is the total variation of $u \in L^1(\Omega)$, $BV(\Omega)$ is the Banach space of functions $u \in L^1(\Omega)$ with

$$\|u\|_{BV} := \int_\Omega u \, dx + \int_\Omega |Du| < \infty,$$

$|\cdot|$ is the n-dimensional Lebesgue measure, and $P_\Omega(G)$ denotes the perimeter of $G \subset \Omega$ with respect to Ω. Denote by ∂' the relative boundary with respect to Ω.

For an oriented smooth hypersurface, we define the second fundamental form so that the principal curvatures and mean curvature of S^{N-1} are negative if the normal vector orients to the center. In this paper, the mean curvature denotes the sum of all principal curvatures.

2. Main results. The original problem has three independent parameters ε, D, and μ with given constants τ, γ, and m. Theorem 2.1 concerns two scalings for fixed μ, that is, $\varepsilon \to 0$ with fixed D, and $\varepsilon \to 0$, $D \to \infty$. To state the main results, define

$$c_0 := \frac{\sqrt{2}}{\int_{-1}^{1} \sqrt{W(s)}\, ds}.$$

Remark 2.1. The constant c_0 can be characterized as follows when f is smooth. Indeed, it is known that if, in addition, f is smooth, then there exist a constant $\delta > 0$ and smooth functions $U : \mathbb{R} \times (-\delta, \delta) \to \mathbb{R}$, $c : (-\delta, \delta) \to \mathbb{R}$ such that

$$U_{xx}(x,V) + c(V)U_x(x,V) + f(U(x,V)) = V, \quad x \in \mathbb{R}, V \in (-\delta,\delta),$$
$$U(-\infty,V) = h_-(V), \quad U(+\infty,V) = h_+(V), \quad c(0) = 0,$$

where $h_\pm(V)$, $V \in (-\delta,\delta)$ are two solutions of $f(\cdot) = V$ which are smooth in V and satisfy $h_\pm(0) = \pm 1$. This means that for each $V \in (-\delta,\delta)$, $u(x,t) = U(x - c(V)t, V)$ is a traveling wave solution of the equation

$$u_t = u_{xx} + f(u) - V, \quad x \in \mathbb{R}, \ \ t \in \mathbb{R},$$

which converges to $h_\pm(V)$ as $x \to \pm\infty$ ($c = c(V)$ represents the speed of the traveling wave solution). See [2, 5, 8, 9] and references therein. Then it is easy to see that there holds

$$c_0 = \left.\frac{dc}{dV}\right|_{V=0}.$$

Our first result is the following.

THEOREM 2.1. *Assume* (A1) *and* (f1)–(f4). *Then the following hold:*

(i) *Let $D = 1$. Then for any sequence $\varepsilon_n \to 0$, there exists a subsequence $\varepsilon_k = \varepsilon_{n_k} \to 0$ and stable stationary solutions (u_k, v_k) of $(E)^\mu_{\varepsilon_k,D}$ such that u_k converges strongly in $L^1(\Omega)$ to a solution of*

$$(P)^\mu \qquad \begin{cases} \min\limits_{u\in\mathcal{G}} B^\mu(u), \quad B^\mu(u) := \left[\dfrac{2}{c_0}P_\Omega(\{u=1\}) + \dfrac{1}{2\mu}(K(u-m),u-m)\right], \\ \mathcal{G} := \{u \in BV(\Omega)\,;\, |u(x)| = 1 \quad \textit{for almost all } x \in \Omega\}, \end{cases}$$

where

$$K := (-\Delta + \gamma)^{-1} \tag{2.1}$$

is the Green operator with a homogeneous Neumann boundary condition.

(ii) *For any sequence $\varepsilon_n \to 0$, $D_n \to \infty$, there exist subsequences $\varepsilon_k = \varepsilon_{n_k} \to 0$, $D_k = D_{n_k} \to \infty$ such that for each k, $(E)^\mu_{\varepsilon_k,D_k}$ has a stable stationary solution (u_k, v_k) which has the property that u_k converges strongly in $L^1(\Omega)$ to a solution of*

$$(\widetilde{P})^\mu \qquad \min_{u\in\mathcal{G}} \widetilde{B}^\mu(u), \quad \widetilde{B}^\mu(u) := \left[\frac{2}{c_0}P_\Omega(\{u=1\}) + \frac{1}{2\mu\gamma}|\Omega|(\langle u\rangle - m)^2\right],$$

where $\mathcal{G}$ is the same as in $(P)^\mu$.

The method used to prove Theorem 2.1 is based on the Γ-convergence result for a sequence of functionals of Modica–Mortola type. Such a method was first applied to such problems as the van der Waals–Cahn–Hilliard model or the phase transition model in the gradient theory of fluids (see [11]). Note that such problems are gradient flows associated to some functionals, but our problem $(E)^\mu_{\varepsilon,D}$ is not.

$(P)^\mu$ and $(\widetilde{P})^\mu$ are the geometric minimization problems with a parameter dependence, which determine the location of interior boundary layers. (Here we call $\Gamma = \partial'\{u = 1\}$ a sharp interface for $u \in \mathcal{G}$.) We are interested in the following two problems: the asymptotic behavior, as $\mu \to 0$, of solutions of $(P)^\mu$ and $(\widetilde{P})^\mu$,

and finding a geometric interface equation for minimizers. Theorem 2.2 concerns the parameter dependence of solutions of $(P)^\mu$ and $(\widetilde{P})^\mu$. In Theorem 2.3, we derive the geometric interface equation associated with the solutions of $(P)^\mu$ and $(\widetilde{P})^\mu$. Note that we obtain it without the technique of the matched asymptotic expansion.

THEOREM 2.2. (i) *Let u^μ be a solution of $(P)^\mu$. Then for any sequence $\mu_k \to 0$, (u^μ) generates the Young measure $\nu = (\nu_x)_{x\in\Omega}$ such that*

$$\nu_x = \frac{1-m}{2}\delta_{-1} + \frac{1+m}{2}\delta_1 \quad \textit{for almost all } x \in \Omega.$$

Furthermore, there holds

$$C_1\mu^{-1/3} \le P_\Omega(\{u^\mu = 1\}) \le C_2\mu^{-1/3}, \quad \mu \in (0,1), \tag{2.2}$$

where C_1, C_2 are positive constants (independent of μ).

(ii) *Let $\widetilde{u}^\mu$ be a solution of $(\widetilde{P})^\mu$. There holds*

$$P_\Omega(\{\widetilde{u}^\mu = 1\}) \le C_3, \quad \mu \in (0,1),$$

for a positive constant C_3 (independent of μ). Furthermore, for any sequence $\mu_n \to 0$, there exists a subsequence $\mu_k \to 0$ that has the following properties:

(1) *there exists a positive constant C_4 such that*

$$\frac{1}{\mu_k}(\langle \widetilde{u}^{\mu_k}\rangle - m)^2 \le \frac{C_4}{|\log \mu_k|}, \quad k = 1, 2, \dots,$$

(2) *$\widetilde{u}^{\mu_k}$ converges strongly in $L^1(\Omega)$ to a solution u^* of*

$$\begin{cases} \min\limits_{u\in\mathcal{M}} P_\Omega(\{u=1\}), \\ \mathcal{M} := \left\{ u \in \mathcal{G}\,;\, \displaystyle\int_\Omega u\,dx = m|\Omega| \right\}. \end{cases}$$

In particular, $(\widetilde{u}^{\mu_k})$ generates the Young measure $\nu_x = \delta_{u^*(x)}$ for almost every $x \in \Omega$.

Remark 2.2. From Theorem 2.2, the solutions (u_k, v_k) in Theorem 2.1 are not spatially constant for large k.

We see that a phase separation with fine structures occurs in the scaling (i) of Theorem 2.2, in the sense that we may construct a sequence of solutions that converges to a pattern with an arbitrarily large perimeter if we choose sufficiently small μ. For variational problems with two scales, see, for example, [1, 7, 13].

THEOREM 2.3. (i) *For fixed $\mu > 0$, let u be a solution of $(P)^\mu$ and $\Gamma = \partial'\{u = 1\}$. Assume that Γ is smooth in a neighborhood U of $x_0 \in \Gamma$. Then there holds*

$$\mu H = c_0 v \quad \textit{on } \Gamma \cap U,$$

where H is the mean curvature of Γ (when the normal vector points from $\{u = -1\}$ to $\{u = +1\}$) and v is the solution of

$$\begin{cases} (-\Delta + \gamma)v = u - m & \textit{in } \Omega, \\ \dfrac{\partial v}{\partial n} = 0 & \textit{on } \partial\Omega. \end{cases}$$

(ii) *For fixed $\mu > 0$, let $\widetilde{u}$ be a solution of $(\widetilde{P})^\mu$ and $\widetilde{\Gamma} = \partial'\{\widetilde{u} = 1\}$. Assume that $\widetilde{\Gamma}$ is smooth in a neighborhood U of $x_0 \in \widetilde{\Gamma}$. Then there holds*

$$\mu H = \frac{c_0}{\gamma}(\langle \widetilde{u} \rangle - m) \quad \text{on } \widetilde{\Gamma} \cap U,$$

where H is the mean curvature of $\widetilde{\Gamma}$.

Remark 2.3. Theorem 2.3(ii) implies that solutions of $(\widetilde{P})^\mu$ typically involve a partition of Ω into regions separated by surfaces of a constant mean curvature. In [5], the authors obtained a limiting free boundary problem from an Allen–Cahn equation with a nonlocal term, which arises as a limit of a reaction-diffusion system. Then we see that any smooth surface that corresponds to a stationary solution of the motion law obtained in [5] has also a constant mean curvature.

Hereafter, for the sake of notational simplicity, we will use the same letters C to denote some positive constants whose values may vary from line to line. This notational convention does not apply to such letters as $C_1, C_2, C_3, \ldots$.

3. Proof of Theorem 2.1. In this section, we prove Theorem 2.1. For simplicity, we show the claim when

$$m = 0$$

is satisfied since we can prove in other cases by the same manner. Denote by $(-D\Delta + \gamma)^{-1}$ the Green operator of $-D\Delta + \gamma$ with the homogeneous Neumann boundary condition. Define a functional

$$I^\mu_{\varepsilon,D}(u) := \int_\Omega \left\{ \frac{\varepsilon^2}{2}|\nabla u|^2 + W(u) \right\} dx + \frac{\varepsilon}{2\mu}((-D\Delta + \gamma)^{-1}u, u)$$

for $u \in H^1(\Omega)$. Let $u_{\varepsilon,D} \in H^1(\Omega)$ be a critical point of $I^\mu_{\varepsilon,D}$ and let $v_{\varepsilon,D} := (-D\Delta + \gamma)^{-1}u_{\varepsilon,D}$. Then $u_{\varepsilon,D}$ satisfies the Euler–Lagrange equation

$$\varepsilon^2 \Delta u_{\varepsilon,D} + f(u_{\varepsilon,D}) - \frac{\varepsilon}{\mu}(-D\Delta + \gamma)^{-1}u_{\varepsilon,D} = 0,$$

which means that $(u_{\varepsilon,D}, v_{\varepsilon,D})$ is a stationary solution of $(E)^\mu_{\varepsilon,D}$. Moreover, by the result of section 3 in [16], $(u_{\varepsilon,D}, v_{\varepsilon,D})$ is stable for all ε if $\tau = 0$ and for $\varepsilon < \tau^{-1}\mu\gamma^2$ if $\tau > 0$. (The condition $\tau \cdot \frac{\varepsilon}{\mu} < \gamma^2$ is sufficient for stability but need not be necessary.) Consider the minimization problem

$$(P)^\mu_{\varepsilon,D} \qquad \min_{u \in H^1(\Omega)} I^\mu_{\varepsilon,D}(u).$$

Noting that the H^1-norm is weakly lower semicontinuous, $(-D\Delta + \gamma)^{-1}$ is compact and continuous on $L^2(\Omega)$, and $I^\mu_{\varepsilon,D}$ is coercive, the problem $(P)^\mu_{\varepsilon,D}$ has a solution $u_{\varepsilon,D}$. To establish Theorem 2.1, it remains to show the behavior of $u_{\varepsilon,D}$ in the limit as $\varepsilon \to 0$ or $\varepsilon \to 0$ and $D \to \infty$.

Define a functional $a^\mu_{\varepsilon,D} : L^1(\Omega) \to [0, \infty]$ for $\varepsilon, \mu, D > 0$ as follows:

$$a^\mu_{\varepsilon,D}(u) := \begin{cases} \varepsilon^{-1} I^\mu_{\varepsilon,D}(u) & \text{for } u \in H^1(\Omega), \\ \infty & \text{otherwise.} \end{cases}$$

We obtain the following lemma.

LEMMA 3.1. *Let $D = 1$. Then*

(1) *for any $\varepsilon_k \to 0$, and $u_k \to u$ weakly in $L^2(\Omega)$ and strongly in $L^1(\Omega)$, there holds*

$$\liminf_{k\to\infty} a^{\mu}_{\varepsilon_k, D}(u_k) \geq b^{\mu}(u);$$

(2) *for any $u \in L^2(\Omega)$, there exist sequences $\varepsilon_k \to 0$ and $u_k \to u$ weakly in $L^2(\Omega)$ and strongly in $L^1(\Omega)$ such that*

$$b^{\mu}(u) \geq \limsup_{k\to\infty} a^{\mu}_{\varepsilon_k, D}(u_k),$$

where

$$b^{\mu}(u) := \begin{cases} B^{\mu}(u) = 2(c_0)^{-1} P_{\Omega}(\{u = 1\}) + (2\mu)^{-1}(Ku, u) & \text{for } u \in \mathcal{G}, \\ \infty & \text{for } u \in L^1(\Omega)\backslash\mathcal{G}. \end{cases}$$

(The definitions of B^{μ} and $\mathcal{G}$ are in Theorem 2.1(i)*, and K is defined in* (2.1)*.)*

LEMMA 3.2. *The following two properties hold:*

(1) *For any $\varepsilon_k \to 0$, $D_k \to \infty$, and $u_k \to u$ weakly in $L^2(\Omega)$ and strongly in $L^1(\Omega)$, there holds*

$$\liminf_{k\to\infty} a^{\mu}_{\varepsilon_k, D_k}(u_k) \geq \widetilde{b}^{\mu}(u);$$

(2) *for any $u \in L^2(\Omega)$, there exist sequences $\varepsilon_k \to 0$, $D_k \to \infty$, and $u_k \to u$ weakly in $L^2(\Omega)$ and strongly in $L^1(\Omega)$ such that*

$$\widetilde{b}^{\mu}(u) \geq \limsup_{k\to\infty} a^{\mu}_{\varepsilon_k, D_k}(u_k),$$

where

$$\widetilde{b}^{\mu}(u) := \begin{cases} \widetilde{B}^{\mu}(u) = 2(c_0)^{-1} P_{\Omega}(\{u = 1\}) + |\Omega|(2\mu\gamma)^{-1}\langle u\rangle^2 & \text{for } u \in \mathcal{G}, \\ \infty & \text{for } u \in L^1(\Omega)\backslash\mathcal{G}. \end{cases}$$

($\widetilde{B}^{\mu}$ is defined in Theorem 2.1(ii)*.)*

Proof of Lemmas. First recall Modica–Mortola theorem, namely, a sequence $(E^{\varepsilon})_{0<\varepsilon\leq 1}$ of functionals on $L^1(\Omega)$ such that

$$E^{\varepsilon}(u) := \begin{cases} \displaystyle\int_{\Omega} \left(\frac{\varepsilon}{2}|\nabla u|^2 + \frac{1}{\varepsilon} W(u) \right) dx & \text{for } u \in H^1(\Omega), \\ \infty & \text{for } u \in L^1(\Omega)\backslash H^1(\Omega). \end{cases} \tag{3.1}$$

Γ converges to

$$E(u) := \begin{cases} 2(c_0)^{-1} P_{\Omega}(\{u = 1\}) & \text{for } u \in \mathcal{G}, \\ \infty & \text{for } u \in L^1(\Omega)\backslash\mathcal{G}\ . \end{cases}$$

That is,

(i) for any $\varepsilon_k \to 0$, and $u_k \to u$ strongly in $L^1(\Omega)$, there holds

$$\liminf_{k\to\infty} E^{\varepsilon_k}(u_k) \geq E(u);$$

(ii) for any $u \in L^1(\Omega)$, there exist sequences $\varepsilon_k \to 0$ and $u_k \to u$ strongly in $L^1(\Omega)$ such that

$$E(u) \geq \limsup_{k\to\infty} E^{\varepsilon_k}(u_k).$$

Note that in (ii), for $u \in \mathcal{G}$, we may choose u_k such that $\|u_k\|_{L^\infty(\Omega)} = 1$. (For example, see [17].) Hence we may assume that $u_k \rightharpoonup u$ weakly in $L^2(\Omega)$.

Noting that if $u_k \rightharpoonup u$ weakly in $L^2(\Omega)$, then

$$(Ku_k, u_k) \to (Ku, u) \quad \text{as } k \to \infty.$$

Lemma 3.1 follows.

To verify Lemma 3.2, it suffices to see that for any sequences $\varepsilon_k \to 0$, $D_k \to \infty$, and $u_k \to u$ weakly in $L^2(\Omega)$ and strongly in $L^1(\Omega)$, there holds

$$(T_k u_k, u_k) \to \frac{|\Omega|\langle u\rangle^2}{\gamma} \quad \text{as } k \to \infty, \tag{3.2}$$

where T_k is the Green operator of $-D_k\Delta + \gamma$ with homogeneous Neumann boundary data. Indeed, we may consider only the case $u_k \in H^1(\Omega)$ since other cases are trivial. In the case $u \notin \mathcal{G}$, by the above Γ-convergence result,

$$\liminf_{k\to\infty} a^\mu_{\varepsilon,D}(u_k) = \infty;$$

hence the assertion follows by the penalty term. Thus we may assume that $u \in \mathcal{G} \subset L^2(\Omega)$.

To see the above property (3.2), denote $0 = \lambda_1 < \lambda_2 \leq \cdots \leq \lambda_i \leq \cdots \to \infty$ the eigenvalues of $-\Delta$ on Ω with homogeneous Neumann boundary condition and let $(\phi_i)_{i=1}^\infty$ be the corresponding eigenfunctions with $\|\phi_i\|_{L^2(\Omega)} = 1$. We may assume that $D_k \geq 1$ for all $k \geq 1$. Expanding u_k and u in terms of (ϕ_i), we have

$$u_k = \sum_{i=1}^\infty a_i^{(k)}\phi_i, \quad u = \sum_{i=1}^\infty a_i\phi_i,$$

where $a_i^{(k)} = (u_k, \phi_i)$ and $a_i = (u, \phi_i)$. Note that

$$a_1 = |\Omega|^{-\frac{1}{2}} \int_\Omega u\, dx$$

and

$$(T_k u_k, u_k) = \frac{(a_1^{(k)})^2}{\gamma} + \sum_{i\geq 2} \frac{(a_i^{(k)})^2}{D_k\lambda_i + \gamma}.$$

Here we estimate

$$\begin{aligned} 0 \leq \sum_{i\geq 2} \frac{(a_i^{(k)})^2}{D_k\lambda_i + \gamma} &\leq 2\sum_{i\geq 2} \frac{(a_i^{(k)} - a_i)^2}{\lambda_i + \gamma} + 2\sum_{i\geq 2} \frac{(a_i)^2}{D_k\lambda_i + \gamma} \\ &= 2\left(K\left(\sum_{i\geq 2} a_i^{(k)}\phi_i - \sum_{i\geq 2} a_i\phi_i\right), \sum_{i\geq 2} a_i^{(k)}\phi_i - \sum_{i\geq 2} a_i\phi_i\right) \\ &\quad + 2\sum_{i\geq 2} \frac{(a_i)^2}{D_k\lambda_i + \gamma}. \end{aligned} \tag{3.3}$$

Since

$$\sum_{i\geq 2} a_i^{(k)}\phi_i \rightharpoonup \sum_{i\geq 2} a_i\phi_i, \quad k\to\infty,$$

weakly in $L^2(\Omega)$, the first term of the right-hand side of (3.3) goes to 0 as $k\to\infty$. Moreover, since

$$\frac{(a_i)^2}{D_k\lambda_i+\gamma} \leq \frac{(a_i)^2}{\lambda_i+\gamma}, \quad \sum_{i\geq 2}\frac{(a_i)^2}{\lambda_i+\gamma} < \infty,$$

the second term of the right-hand side of (3.3) also goes to 0 as $k\to\infty$ by the Lebesgue dominated convergence theorem. Hence

$$\lim_{k\to\infty}\sum_{i\geq 2}\frac{(a_i^{(k)})^2}{D_k\lambda_i+\gamma} = 0.$$

Noting that $a_1^{(k)}\to a_1$ as $k\to\infty$, we have

$$\lim_{k\to\infty}(T_k u_k, u_k) = \frac{(a_1)^2}{\gamma} = \frac{|\Omega|\langle u\rangle^2}{\gamma},$$

as desired. The proof is complete. □

To show Theorem 2.1(i), it is sufficient to see that for any sequence $\varepsilon_n\to 0$, there exists a subsequence $\varepsilon_k\to 0$ such that $u_{\varepsilon_k,D}\to u^*$ weakly in $L^2(\Omega)$ and strongly in $L^1(\Omega)$. Indeed, by Lemma 3.1(1), there holds

$$b^\mu(u^*) \leq \liminf_{k\to\infty} a^\mu_{\varepsilon_k,D}(u_{\varepsilon_k,D}).$$

On the other hand, by Lemma 3.1(2), for any $u\in\mathcal{G}$, there exists a sequence u_k such that

$$\limsup_{k\to\infty} a^\mu_{\varepsilon_k,D}(u_k) \leq b^\mu(u).$$

Hence we have $b^\mu(u^*)\leq b^\mu(u)$, which means that u^* is a solution of $(P)^\mu$.

It remains to see the existence of a subsequence ε_k such that $u_{\varepsilon_k,D}$ converges to some u^* weakly in $L^2(\Omega)$ and strongly in $L^1(\Omega)$. Note that there holds by the assumption (f4), for some positive constants C_5, C_6,

$$\begin{aligned}\int_\Omega |u|^{s+1}\,dx &\leq \int_\Omega (C_5 + C_6 W(u))\,dx\\ &\leq C_5|\Omega| + C_6\varepsilon E^\varepsilon(u),\end{aligned}$$

where E^ε is defined in (3.1). Furthermore, by constructing a comparison function, it is easy to see that for some constant $C>0$, independent of $\varepsilon\in(0,1]$ and $D\geq 1$,

$$E^\varepsilon(u_{\varepsilon,D}) \leq a^\mu_{\varepsilon,D}(u_{\varepsilon,D}) = \frac{1}{\varepsilon} I^\mu_{\varepsilon,D}(u_{\varepsilon,D}) \leq C.$$

Hence it follows that $(u_{\varepsilon,D})_{0<\varepsilon\leq 1}$ is bounded in $L^2(\Omega)$ and that by the Modica–Mortola theorem, $(u_{\varepsilon,D})_{0<\varepsilon\leq 1}$ is relatively compact in $L^1(\Omega)$. Thus the claim follows. The proof of Theorem 2.1(i) is complete.

Similarly, we establish Theorem 2.1(ii) by using Lemma 3.2. □

4. Asymptotic behavior of solutions. In this section, we prove Theorem 2.2.

Proof of Theorem 2.2(i). For simplicity, we show the claim when there hold

$$\frac{c_0}{4} = 1 \text{ and } \gamma = 1,$$

since we can prove other cases by the same manner. We will show that our solution u^μ oscillates rapidly with the average wave length of order $\mu^{\frac{1}{3}}$ when μ goes to zero.

First we give an upper bound for the minimal energy. Define the function oscillating in a one-dimensional direction as follows. Let r be the solution of

$$\begin{array}{ll} -r''(s) = q(s) - m & \text{for } 0 < s < 1, \\ r(s+1) = r(s), r'(s+1) = r'(s) & \text{for all } s \in \mathbb{R}, \end{array}$$

where

$$q(s) := \begin{cases} +1, & s < [s] + \frac{1+m}{2}, \\ -1, & [s] + \frac{1+m}{2} \le s. \end{cases}$$

Letting $Q_\mu(x) := q(x_1/\mu^{\frac{1}{3}})$ and $R_\mu(x) := r(x_1/\mu^{\frac{1}{3}})$, there holds

$$-\Delta(\mu^{\frac{2}{3}} R_\mu) = Q_\mu - m.$$

Let $W_\mu := K(Q_\mu - m)$, that is, W_μ is the solution of

$$\begin{cases} -\Delta W_\mu + W_\mu = Q_\mu - m & \text{in } \Omega, \\ \dfrac{\partial W_\mu}{\partial n} = 0 & \text{on } \partial\Omega. \end{cases}$$

Then there holds $W_\mu = \mu^{\frac{2}{3}} R_\mu - X_\mu - Y_\mu$, where X_μ, Y_μ are the solutions of

$$\begin{cases} -\Delta X_\mu + X_\mu = 0 & \text{in } \Omega, \\ \dfrac{\partial X_\mu}{\partial n} = \dfrac{\partial}{\partial n}(\mu^{\frac{2}{3}} R_\mu) & \text{on } \partial\Omega, \end{cases}$$

and

$$\begin{cases} -\Delta Y_\mu + Y_\mu = \mu^{\frac{2}{3}} R_\mu & \text{in } \Omega, \\ \dfrac{\partial Y_\mu}{\partial n} = 0 & \text{on } \partial\Omega, \end{cases}$$

respectively. Since $P_\Omega(\{Q_\mu = 1\}) = O(\mu^{-\frac{1}{3}})$ and $\|W_\mu\|_{L^2(\Omega)} = O(\mu^{\frac{2}{3}})$ as $\mu \to 0$, one obtains the upper bound for $(P)^\mu$, that is, $B^\mu(u^\mu) \le C\mu^{-\frac{1}{3}}$, in particular,

$$(K(u^\mu - m), u^\mu - m) \le C\mu^{\frac{2}{3}}, \quad P_\Omega(\{u^\mu = 1\}) \le C\mu^{-\frac{1}{3}}. \tag{4.1}$$

To see a Young measure associated with u^μ, choose any sequence $\mu_l \to 0$. Then by the fundamental existence theorem, for a subsequence μ_k, we may assume that (u^{μ_k}) generates a Young measure $\nu = (\nu_x)_{x\in\Omega}$. Since $u^{\mu_k}(x) = 1$ or $u^{\mu_k}(x) = -1$ for almost all $x \in \Omega$, there exists a measurable function $\varphi : \Omega \to [0,1]$ such that $\nu_x = (1 - \varphi(x))\delta_{-1} + \varphi(x)\delta_1$ for almost every $x \in \Omega$. Then since

$$\int_{-\infty}^{\infty} \lambda \, d\nu_x(\lambda) = 2\varphi(x) - 1,$$

we have $u^{\mu_k} \rightharpoonup 2\varphi - 1$ weakly in $L^2(\Omega)$. Thus as $k \to \infty$,

$$(K(u^{\mu_k} - m), u^{\mu_k} - m) \to (K(2\varphi - 1 - m), 2\varphi - 1 - m) = 0$$

by (4.1). Therefore $\varphi(x) \equiv \frac{1+m}{2}$ for almost all $x \in \Omega$.

It remains to see the lower bound in (2.2). The key estimate to see this is the following interpolation inequality.

LEMMA 4.1. *There exist positive constants β_1 and δ_1 such that for all $0 < \delta \le \delta_1$ and $u \in \mathcal{G}$,*

$$\beta_1 \le \delta P_\Omega(\{u = 1\}) + \delta^{-2}(K(u - m), u - m). \tag{4.2}$$

This is obtained by standard arguments used in [6]. (A similar inequality was also proved in [12].) For the one-dimensional case, see [14]. Then we have the following.

COROLLARY 4.2. (1)

$$\min_{u \in \mathcal{G}} B^\mu(u) \ge \frac{1}{2}\beta_1 \mu^{-\frac{1}{3}}.$$

(2) *For any $u \in \mathcal{G}$ such that*

$$(K(u - m), u - m) \le \left(\frac{\delta_1}{2}\right)^2,$$

there holds

$$[P_\Omega(\{u = 1\})]^2 \cdot (K(u - m), u - m) \ge \frac{9}{64}. \tag{4.3}$$

By (4.3) and the upper bound (4.1), we obtain the estimates with a uniform constant $C > 0$,

$$(K(u^\mu - m), u^\mu - m) \ge C\mu^{\frac{2}{3}},$$
$$P_\Omega(\{u^\mu = 1\}) \ge C\mu^{-\frac{1}{3}}$$

for all $\mu \in (0, 1)$. The proof is complete. □

Proof of Theorem 2.2(ii). Here, for simplicity, we show the claim when there hold

$$\frac{c_0}{4\gamma}|\Omega| = 1 \text{ and } m = 0,$$

since we can prove other cases by the same manner. Let $\widetilde{p}^\mu := \dfrac{c_0}{2}\widetilde{B}^\mu(\widetilde{u}^\mu)$. Choose $u_0 \in L^1(\Omega)$ such that $\langle u_0 \rangle = 0$ and $W(u_0) = 0$ (i.e., $u_0(x) = -1$ or $u_0(x) = +1$) for almost all $x \in \Omega$. Then

$$\widetilde{p}^\mu \le P_\Omega(\{u_0 = 1\}) = C,$$

where C is a constant independent of μ. Moreover, by the definition of $\widetilde{p}^\mu$, the function $\mu \mapsto \widetilde{p}^\mu$ is nonincreasing and hence differentiable at almost every number $\mu \in (0, 1)$. We will find the sequence $\mu_k \to 0$ satisfying the conditions (1) and (2) in Theorem 2.2(ii). For $0 < \mu_1 < \mu_0$,

$$\begin{aligned} C \ge \widetilde{p}^{\mu_1} - \widetilde{p}^{\mu_0} &\ge \int_{\mu_1}^{\mu_0} \left| \frac{d}{d\mu}\widetilde{p}^\mu \right| d\mu \\ &\ge \operatorname*{ess\,inf}_{\mu_1 < \mu < \mu_0} \mu \left| \frac{d}{d\mu}\widetilde{p}^\mu \right| \left| \log \frac{\mu_1}{\mu_0} \right|. \end{aligned} \tag{4.4}$$

Thus we infer that

$$\mu_k \left| \frac{d}{d\mu} \widetilde{p}^{\mu} \right| \le \frac{C}{|\log \mu_k|}$$

for a suitable sequence $\mu_k \to 0$. Hence, noting Lemma 7.2 in [18], there holds

$$\mu_k^{-1} \langle \widetilde{u}^{\mu_k} \rangle^2 = \mu_k \left| \frac{\partial \widetilde{B}^{\mu}}{\partial \mu}(\widetilde{u}^{\mu_k}) \right| \le \mu_k \left| \frac{d}{d\mu} \widetilde{p}^{\mu} \right| \le \frac{C}{|\log \mu_k|}. \tag{4.5}$$

We may assume that $\widetilde{u}^{\mu_k} \to u^*$ strongly in $L^1(\Omega)$. (Note that $\widetilde{u}^{\mu_k}$ is bounded in $BV(\Omega)$ and the compactness of the embedding $BV(\Omega) \hookrightarrow L^1(\Omega)$.) Then $W(u^*(x)) = 0$ almost every $x \in \Omega$. Furthermore, by (4.5), we have

$$\langle u^* \rangle = \frac{1}{|\Omega|} \lim_{k \to \infty} \int_{\Omega} \widetilde{u}^{\mu_k} \, dx = 0,$$

that is, $u_0 \in \mathcal{M}$. It remains to show that the minimality of $P_{\Omega}(\{u = 1\})$ in $\mathcal{M}$. For any $u \in \mathcal{M}$, by the definition of $\widetilde{u}^{\mu_k}$,

$$P_{\Omega}(\{\widetilde{u}^{\mu_k} = 1\}) + \frac{1}{\mu_k} \langle \widetilde{u}^{\mu_k} \rangle^2 \le P_{\Omega}(\{u = 1\}).$$

Taking the limit $k \to \infty$, we obtain by the weak lower semicontinuity of the total variation in L^1 topology,

$$P_{\Omega}(\{u^* = 1\}) \le \liminf_{k \to \infty} P_{\Omega}(\{\widetilde{u}^{\mu_k} = 1\}) \le P_{\Omega}(\{u = 1\}).$$

The proof is complete. □

5. Stationary sharp interfaces. In this section, we prove Theorem 2.3. First we give some notation and preliminaries. Let

$$F : B \to \mathbb{R}^N, \quad B := \{z \in \mathbb{R}^{N-1}; \, |z| < 1\},$$

be a parametrized (smooth) hypersurface and let $\Gamma := \{F(z); \, z \in B\} \subset \mathbb{R}^N$ be the corresponding hypersurface. We identify functions on Γ with functions on B. For a given surface F, we use the following notation. Let (g_{ij}) represent the first fundamental form, and let $g^{ij} = (g_{ij})^{-1}$ be the coefficients of the inverse of the matrix (g_{ij}). Let

$$ds := \sqrt{|g|} \, dz,$$

where $|g| = \det(g_{ij})$.

Let $\mathbf{n} = \mathbf{n}(z)$ be the normal vector and let (ℓ_{ij}) represent the second fundamental form. Let $H := \kappa_1 + \cdots + \kappa_{N-1}$ be the mean curvature, where $\kappa_1, \ldots, \kappa_{N-1}$ are the principal curvatures.

Given a parametrized surface F, let $\overline{\alpha}$ be a "variation" of F with "velocity" $\Phi \cdot \mathbf{n}$, namely, $\overline{\alpha}(\eta)$ is a parametrized surface for each $\eta \in (-\delta, \delta)$, $\delta > 0$, such that

$$\overline{\alpha}(\eta)(z) = \alpha(\eta, z) := F(z) + \eta \Phi(z) \mathbf{n}(z), \quad (\eta, z) \in (-\delta, \delta) \times B,$$

for a given scalar function $\Phi : B \to \mathbb{R}$. Let

$$A(F) := \int_{\Gamma} ds = \int_{B} \sqrt{|g|} \, dz$$

be the area of F.

The first variation of the area is given by the following standard lemma. (We omit the proof.)

LEMMA 5.1.

$$\left.\frac{dA(\overline{\alpha}(\eta))}{d\eta}\right|_{\eta=0} = \int_\Gamma \Phi H\, ds.$$

Proof of Theorem 2.3(i). Here, for simplicity, we show the claim when $\frac{c_0}{2\mu} = 1$ is satisfied since we can prove in other cases by the same manner. For $u \in \mathcal{G}$ and a function $v \in H^1(\Omega)$, let

$$e(u,v) := P_\Omega(\{u=1\}) - \int_\Omega \left[\frac{1}{2}(|\nabla v|^2 + \gamma v^2) - (u-m)v\right] dx.$$

Remark 5.1. For any $u \in \mathcal{G}$, there holds

$$\begin{aligned}\frac{c_0}{2}B^\mu(u) &= P_\Omega(\{u=1\}) + \frac{1}{2}(K(u-m), u-m)\\ &= \max_{v\in H^1(\Omega)} e(u,v),\end{aligned}$$

where the maximum in the right-hand side is attained at v if and only if $v = K(u-m)$. Let $\overline{\beta} = \overline{\beta}(\eta)$ be a smooth path in $\mathcal{G}$ such that

$$\overline{\beta}(0) = u.$$

We call $\overline{\beta}$ a variation of u. We identify the velocity vector

$$\left.\frac{d}{d\eta}\overline{\beta}\right|_{\eta=0}$$

with a section $\Phi\mathbf{n} = \Phi(s)\mathbf{n}(s)$, $s \in \Gamma$ of the normal bundle of Γ. Given a smooth function $v, \Psi : \Omega \to \mathbb{R}$, choose a smooth map $\varrho : V \times \Omega \to \mathbb{R}$ such that

$$\varrho(0,x) = v(x), \quad \frac{\partial\varrho}{\partial\eta}(0,x) = \Psi(x).$$

We call $\overline{\varrho}(\eta) := \varrho(\eta,\cdot)$ a variation of v with velocity Ψ.

Claim 1.

$$\left.\frac{d}{d\eta}e(\overline{\beta}(\eta),\overline{\varrho}(\eta))\right|_{\eta=0} = \int_\Gamma (H-2v)\Phi\, ds + \int_\Omega [-\nabla v\cdot\nabla\Psi - \gamma\Psi + (u-m)\Psi]\, dx.$$

Proof. We calculate the first variation of e, that is,

$$\left.\frac{d}{d\eta}e(\overline{\beta}(\eta),\overline{\varrho}(\eta))\right|_{\eta=0} = \left.\frac{d}{d\eta}e(\overline{\beta}(\eta),v)\right|_{\eta=0} + \left.\frac{d}{d\eta}e(u,\overline{\varrho}(\eta))\right|_{\eta=0}. \tag{5.1}$$

By Lemma 5.1,

$$\left.\frac{d}{d\eta}P_\Omega(\overline{\beta}(\eta)=1)\right|_{\eta=0} = \int_\Gamma H\Phi\, ds.$$

Similarly,

$$\begin{aligned}\left.\frac{d}{d\eta}\int_\Omega \overline{\beta}(\eta)v\, dx\right|_{\eta=0} &= -2\int_\Gamma v\Phi\, ds,\\ \left.\frac{d}{d\eta}e(u,\overline{\varrho}(\eta))\right|_{\eta=0} &= \int_\Omega [-\nabla v\cdot\nabla\Psi - \gamma\Psi + (u-m)\Psi]\, dx.\end{aligned}$$

Hence by (5.1), the claim follows. ∎

If u is a solution of $(P)^\mu$, for any variation $\overline{\beta}(\eta), \overline{\varrho}(\eta)$, the quantity

$$\frac{d}{d\eta} e(\overline{\beta}(\eta), \overline{\varrho}(\eta))\Big|_{\eta=0}$$

must vanish; hence there holds

$$H - 2v = 0$$

on Γ. The proof is complete. □

Proof of Theorem 2.3(ii). For simplicity, we show the claim when there hold

$$\frac{c_0}{2\mu\gamma} = 1 \text{ and } m = 0,$$

since we can prove in other cases by the same manner. Then

$$\frac{c_0}{2}\widetilde{B}^\mu(u) = P_\Omega(\{u = 1\}) + \frac{1}{2|\Omega|}\left(\int_\Omega u\,dx\right)^2.$$

On the other hand,

$$\frac{d}{d\eta}\left(\int_\Omega \overline{\beta}(\eta)\,dx\right)^2\Big|_{\eta=0} = -2\int_\Omega u\,dx \int_\Gamma 2\Phi\,ds.$$

Hence if u is a solution of $(\widetilde{P})^\mu$, then by Lemma 5.1, there holds $H - 2\langle u\rangle = 0$ on Γ. The proof is complete. □

6. Remarks on the sharp interface problems. In this section, we give some remarks about the problems $(P)^\mu$ and $(\widetilde{P})^\mu$. Throughout this section, we set for simplicity

$$c_0 := 2,\ \gamma := \frac{1}{2},\ |\Omega| := 1.$$

We study the lower bound in Theorem 2.2(i). Define

$$\alpha := \inf\left\{\liminf_{k\to\infty} \mu_k^{\frac{1}{3}} B^{\mu_k}(u^{\mu_k})\,;\ \mu_k \to 0\right\} > 0.$$

Here α depends only on N, Ω, and m. By an argument similar to the one used in section 4, we have

$$\beta := \inf\left\{\liminf_{k\to\infty}\left[\mu_k^{-\frac{2}{3}}(\langle u_k\rangle - m)^2 + S(u_k)\right]\ ;\ \mu_k \to 0, u_k \in \mathcal{G}\right\} > 0,$$

where

$$S(u) := \frac{3}{2}[P_\Omega(\{u = 1\})]^{\frac{2}{3}} \cdot [(K(u - \langle u\rangle), u - \langle u\rangle)]^{\frac{1}{3}}.$$

Proposition 6.1. *There holds*

$$\alpha \geq \beta.$$

Proof. Let $\mu_k \to 0$ and $u^{\mu_k} \in \mathcal{G}$ be sequences such that

$$\lim_{k\to\infty} \mu_k^{\frac{1}{3}} B^{\mu_k}(u^{\mu_k}) = \alpha.$$

Then

$$\liminf_{k\to\infty}\left\{\mu_k^{-\frac{2}{3}}(\langle u^{\mu_k}\rangle - m)^2 + S(u^{\mu_k})\right\} \leq \alpha.$$

Hence $\beta \leq \alpha$. □

Next we give an upper bound for α using the functional defined as follows:

$$S^*(u) := \frac{3}{2}[P_\Omega(\{u=1\})]^{\frac{2}{3}} \cdot [((-\Delta)^{-1}(u-m), u-m)]^{\frac{1}{3}}$$

for $u \in \mathcal{M}$. Here, $(-\Delta)^{-1}$ denotes the Green operator of $-\Delta$ acting on $\{u \in L^2(\Omega)\,;\, \int_\Omega u\,dx = 0\}$ with homogeneous Neumann data.

As in the proof of Theorem 2.2 and Proposition 6.1, we obtain

$$\begin{aligned}\alpha^* := \inf \Bigg\{ &\liminf_{k\to\infty} \left[\mu_k^{\frac{1}{3}} P_\Omega(\{u_k=1\}) + \frac{1}{2}\mu_k^{-\frac{2}{3}}((-\Delta)^{-1}(u_k-m), u_k-m)\right]; \\ &\mu_k \to 0 \text{ and } u_k \in \mathcal{M}\Bigg\} \\ \geq \beta^* > 0,\end{aligned}$$

where

$$\beta^* := \inf_{u\in\mathcal{M}} S^*(u).$$

Remark 6.1. If

$$\alpha^* > \beta^*, \tag{6.1}$$

then β^* is attained at some $u^* \in \mathcal{M}$, that is, $S^*(u^*) = \beta^*$. Moreover, the condition (6.1) is necessary and sufficient for the BV-boundedness of all minimizing sequences for S^* in $\mathcal{M}$.

Proof. To see that (6.1) is necessary, let u_μ^* be a solution of

$$\min_{u\in\mathcal{M}} \left\{ P_\Omega(\{u=1\}) + \frac{1}{2\mu}((-\Delta)^{-1}(u-m), u-m)\right\}.$$

Choose a sequence $\mu_k \to 0$ such that

$$\alpha^* = \lim_{k\to\infty} \left\{ \mu_k^{\frac{1}{3}} P_\Omega(\{u_k=1\}) + \frac{1}{2}\mu_k^{-\frac{2}{3}}((-\Delta)^{-1}(u_k-m), u_k-m)\right\}$$

with $u_k = u_{\mu_k}^*$. If $\beta^* \geq \alpha^*$, then u_k is a minimizing sequence for S^* in $\mathcal{M}$. On the other hand, by an argument similar to the one used in the proof of Theorem 2.2, we have for a uniform constant C,

$$P_\Omega(\{u_k=1\}) \geq C\mu_k^{-\frac{1}{3}}.$$

Hence u_k is unbounded in $BV(\Omega)$.

We will see that (6.1) is sufficient. Suppose that $\alpha^* > \beta^*$. Assume by contradiction that u_k is a minimizing sequence for S^* in $\mathcal{M}$ such that

$$P_\Omega(\{u_k=1\}) \to \infty,$$

as $k \to \infty$. Let

$$\mu_k := \frac{((-\Delta)^{-1}(u_k-m), u_k-m)}{P_\Omega(\{u_k=1\})} \leq \frac{C}{[P_\Omega(\{u_k=1\})]^3} \to 0. \tag{6.2}$$

Then

$$\begin{aligned}\alpha^* &\le \lim_{k\to\infty}\left\{\mu_k^{\frac{1}{3}}P_\Omega(\{u_k=1\})+\frac{1}{2}\mu_k^{-\frac{2}{3}}((-\Delta)^{-1}(u_k-m),u_k-m)\right\}\\ &= \lim_{k\to\infty} S^*(u_k)\\ &= \beta^*,\end{aligned}$$

which is a contradiction. □

PROPOSITION 6.2. *Assume* $\Omega = Q := (0,1)^N$. *Then*

$$\alpha \le \beta^* := \inf_{u\in\mathcal{M}} S^*(u).$$

Proof. It is easily seen that $\alpha \le \alpha^*$. Hence it is sufficient to show that $\alpha^* = \beta^*$. Assume by contradiction that $\alpha^* > \beta^*$. Then by Remark 6.1, there exists a minimizer u_1 for S^* in $\mathcal{M}$. By rescaling the axially symmetric and periodic extension of u_1, we will construct a minimizing sequence u_k for S^* in $\mathcal{M}$ such that

$$P_\Omega(\{u_k=1\}) \to \infty.$$

Indeed, let u^* be the function on $\mathbb{R}^N$ which satisfies $u^*|_Q = u_1$ and is axially symmetric about the hyperplanes $\{x_i = 0\}$, $1 \le i \le N$, and invariant under the translations generated by two times of the fundamental vectors. Let

$$u_k(x) := u^*(2^{k-1}x), \quad x \in Q, \tag{6.3}$$

for $k \ge 2$. Then $u_k \in \mathcal{M}$. We claim that

$$\begin{aligned}P_\Omega(\{u_k=1\}) &= 2^{k-1}P_\Omega(\{u_1=0\}),\\ ((-\Delta)^{-1}(u_k-m),u_k-m) &= \left(\frac{1}{2}\right)^{2(k-1)}((-\Delta)^{-1}(u_1-m),u_1-m),\end{aligned}$$

and hence $S^*(u_k) = \beta^*$. To see the first equality, note that the contribution to the perimeter of $\{u^*=1\}$ on the hyperplanes $\{x_i = k\}$, $k = 1,2,\dots,n$ $(n \in \mathbb{N})$, vanishes because of the symmetry. Thus we have $P_{nQ}(\{u^*=1\}) = n^N P_Q(\{u^*=1\})$. The claim follows by a scaling argument. Therefore u_k is a minimizing sequence for S^*, which is unbounded in $BV(Q)$. This is a contradiction by Remark 6.1. □

Hereafter let $N = 2$, $\Omega = Q := (0,1)^2$. Now we have the following.

DEFINITION 6.3. $u \in \mathcal{G}$ *is called planar if* $u = u(x,y)$ *depends only on* x.

We show the following.

PROPOSITION 6.4. *There exists a constant* $m \in (0,1)$ *and a sequence* $\mu_k \to 0$ *such that every solution* u^{μ_k} *of* $(P)^{\mu_k}$ *is not planar.*

Proof. We construct comparison functions as follows. Let $m = 1 - \frac{2}{n^2}$ $(n = 2,3,\dots)$ and

$$u_1(x,y) := \begin{cases} -1, & \max\{|x|,|y|\} \le \dfrac{1}{n},\\ +1 & \text{otherwise.}\end{cases}$$

Then $\langle u_1\rangle = m$. We write u_1 as

$$u_1 = m + \sum_{\substack{i,j\ge 0\\(i,j)\ne(0,0)}} a_{i,j}\phi_{i,j},$$

where

$$\phi_{i,j} := 2\cos i\pi x \cdot \cos j\pi y, \quad i,j \geq 0, (i,j) \neq (0,0),$$

and

$$a_{i,j} := \begin{cases} -\dfrac{4}{ij\pi^2} \sin\dfrac{i\pi}{n} \cdot \sin\dfrac{j\pi}{n}, & i \geq 1, j \geq 1, \\ -\dfrac{4}{i\pi n} \sin\dfrac{i\pi}{n}, & i \geq 1, j = 0, \\ -\dfrac{4}{j\pi n} \sin\dfrac{j\pi}{n}, & i = 0, j \geq 1, \end{cases}$$

are Fourier coefficients. Noting that if $i \equiv k \pmod{n}$, then

$$\sin^2\frac{i\pi}{n} = \sin^2\frac{k\pi}{n} \leq \frac{(k\pi)^2}{n^2},$$

we estimate for $\lambda_{i,j} := \pi^2(i^2+j^2)$ $(i,j \geq 1)$,

$$\begin{aligned} \sum_{i,j\geq 1} \frac{(a_{ij})^2}{\lambda_{i,j}} &= \sum_{i,j\geq 1} \frac{16}{\pi^6} \frac{\sin^2\frac{i\pi}{n}\sin^2\frac{j\pi}{n}}{i^2j^2(i^2+j^2)} \\ &\leq \frac{16}{\pi^2 n^4} \sum_{k,l=1}^{n} \sum_{i,j\geq 0} \frac{k^2l^2}{(ni+k)^2(nj+l)^2[(ni+k)^2+(nj+l)^2]} \\ &= \frac{16}{\pi^2 n^4} \sum_{k,l=1}^{n} \left[\sum_{i,j\geq 1} \frac{k^2l^2}{(ni+k)^2(nj+l)^2[(ni+k)^2+(nj+l)^2]} \right. \\ &\quad + \frac{1}{k^2+l^2} + \sum_{j=1}^{\infty} \frac{l^2}{(nj+l)^2[k^2+(nj+l)^2]} \\ &\quad \left. + \sum_{i=1}^{\infty} \frac{k^2}{(ni+k)^2[(ni+k)^2+l^2]} \right] \\ &\leq C\frac{(\log n)^2}{n^4}, \end{aligned}$$

where C is a constant independent of n. Estimating other terms similarly, we get

$$((-\Delta)^{-1}(u_1 - m), u_1 - m) \leq \alpha_0 \frac{(\log n)^2}{n^4}.$$

Hence noting that $P_Q(\{u_1 = 1\}) = \frac{2}{n}$, we see that for any $\delta \in (0,1)$, there exists a constant α_1 such that

$$(\alpha \leq \beta^* \leq) S^*(u_1) \leq \frac{\alpha_1}{n^{2-\delta}}$$

for large n. Define u_k as (6.3) in the proof of Proposition 6.2, and define μ_k as (6.2).

We estimate

$$\begin{aligned}\limsup_{k\to\infty} \mu_k^{\frac{1}{3}} B^{\mu_k}(u^{\mu_k}) &\le \lim_{k\to\infty} \left\{ \mu_k^{\frac{1}{3}} P_\Omega(\{u_k = 1\}) + \frac{1}{2}\mu_k^{-\frac{2}{3}}((-\Delta)^{-1}(u_k - m), u_k - m) \right\} \\ &= \lim_{k\to\infty} S^*(u_k) \\ &\le \frac{\alpha_1}{n^{2-\delta}}. \qquad (6.4)\end{aligned}$$

Completion of the proof of Proposition 6.4. To the contrary, assume that each $(P)^{\mu_k}$ has a planar solution u^{μ_k}. Then to get a lower bound, we apply the following.

LEMMA 6.5. *Let $u \in \mathcal{G}$ be planar and let*

$$m' := \int_Q u\,dxdy, \quad L := P_Q(\{u = 1\}).$$

Then

$$S(u) \ge \frac{3}{2}\left(\frac{L}{L+1}\right)^{\frac{2}{3}} \left(\frac{C_0}{12}\right)^{\frac{1}{3}} \min\{1 - m', 1 + m'\}^{\frac{2}{3}},$$

where $C_0 = \frac{2\pi^2}{2\pi^2+1}$. □

Letting $m_k := \langle u^{\mu_k} \rangle$ and $L_k := P_Q(\{u^{\mu_k} = 1\})$, we have

$$\begin{aligned} C &\ge \mu_k^{\frac{1}{3}} B^{\mu_k}(u^{\mu_k}) \\ &\ge \mu_k^{-\frac{2}{3}}(m_k - m)^2 + S(u^{\mu_k}) \\ &\ge \mu_k^{-\frac{2}{3}}(m_k - m)^2 + \frac{3}{2}\left(\frac{L_k}{L_k+1}\right)^{\frac{2}{3}} \left(\frac{C_0}{12}\right)^{\frac{1}{3}} (1 - m_k)^{\frac{2}{3}}, \end{aligned}$$

which implies that $\lim_{k\to\infty} m_k = m$, and hence $L_k \ge 1$ for large k. Thus taking the limit $k \to \infty$, we have

$$\begin{aligned}\liminf_{k\to\infty} \mu_k^{\frac{1}{3}} B^{\mu_k}(u^{\mu_k}) &\ge \frac{3}{2}\left(\frac{1}{2}\right)^{\frac{2}{3}} \left(\frac{C_0}{12}\right)^{\frac{1}{3}} (1 - m)^{\frac{2}{3}} \\ &=: \alpha_2 n^{-\frac{4}{3}},\end{aligned}$$

which contradicts the upper bound (6.4) for large n. Proposition 6.4 has been proved. □

Proof of Lemma 6.5. Without loss of generality, we may assume that $m' \in [0, 1]$.

Since the second eigenvalue of $-\frac{d^2}{dx^2}$ acting on $H^2(0,1)$ with homogeneous Neumann data is π^2, we have

$$(K(u - m'), u - m') \ge \frac{\pi^2}{\pi^2 + \frac{1}{2}}((-\Delta)^{-1}(u - m'), u - m').$$

Writing $u(x, y) = g(x)$, let $0 < x_1 < x_2 < \cdots < x_L < 1$ be the discontinuity points of g in the interval $(0, 1)$. Here $L = P_\Omega(\{u = 1\})$. Then we estimate

$$S(u) \ge \frac{3}{2} L^{\frac{2}{3}} (C_0)^{\frac{1}{3}} [((-\Delta)^{-1}(u - m'), u - m')]^{\frac{1}{3}}. \qquad (6.5)$$

Letting $\psi \in C^1[0,1]$ be the (weak) solution of

$$\begin{cases} -\psi'' = g - m', & 0 < x < 1, \\ \psi'(0) = \psi'(1) = 0, \end{cases}$$

we have

$$((-\Delta)^{-1}(u-m'), u-m') = \int_0^1 \psi(g-m')\,dx = \int_0^1 (\psi'(x))^2\,dx.$$

To complete the proof of Lemma 6.5, we show the following.

LEMMA 6.6. *Let $u : [0,1] \to \mathbb{R}$ be a piecewise linear function such that $u'(x)$ is constant on each interval (x_i, x_{i+1}) and takes c_i, where $0 = x_0 < x_1 < \cdots < x_L < x_{L+1} = 1$ and $c_i \in \mathbb{R}$. Then*

$$\int_0^1 u^2\,dx \geq \frac{1}{12}\left(\sum_{i=0}^{L} \frac{1}{|c_i|}\right)^{-2}.$$

Proof. Since $u'(x) = c_i$ for $x_i < x < x_{i+1}$, it follows that

$$\int_{x_i}^{x_{i+1}} u^2\,dx = \int_{x_i}^{x_{i+1}} \left(c_i\left(x - \frac{x_i + x_{i+1}}{2}\right) + \frac{u(x_i) + u(x_{i+1})}{2}\right)^2 dx \geq \frac{c_i^2}{12}(x_{i+1} - x_i)^3.$$

Summing over i, we get

$$\begin{aligned} \int_0^1 u^2\,dx &\geq \frac{1}{12}\sum_{i=0}^{L} \frac{1}{|c_i|}(|c_i|(x_{i+1} - x_i))^3 \\ &\geq \frac{1}{12}\left(\sum_{i=0}^{L} \frac{1}{|c_i|}\right)^{-2}. \end{aligned}$$

Here in the second inequality we used the Jensen's inequality. □

Completion of the proof of Lemma 6.5. Applying Lemma 6.6 to ψ', we have

$$((-\Delta)^{-1}(u-m'), u-m') \geq \frac{(1-m')^2}{12}(L+1)^{-2}. \tag{6.6}$$

Combining (6.5) and (6.6), the proof is complete. □

Finally we remark on the problem $(\widetilde{P})^\mu$.

Remark 6.2. We think that typical interfaces for solutions of $(\widetilde{P})^\mu$ should be lines or circles when $N = 2$. We believe that, for m sufficiently close to 1, and μ small, an interface approximated by a circle of a small radius, centered near the points on the boundary, which have the maximum mean curvature, should arise as in Cahn–Hilliard theory. We remark that phase separation with circular interfaces are observed in experiments of block copolymer (see [15]).

Example. Let $m = 0.8$ and $\Omega = (0,1)^2$. If u is planar, then

$$\widetilde{B}^\mu(u) \geq \min\left\{\frac{0.04}{\mu}, 1\right\}.$$

On the other hand, if $u \in \mathcal{G}$ has the circular interface such that $u(x,y) = -1$ for $x^2 + y^2 < \frac{0.4}{\pi}$ and $u = +1$, otherwise,

$$\widetilde{B}^\mu(u) \leq \frac{\sqrt{0.4\pi}}{2} \sim 0.56.$$

Hence solutions for $(\widetilde{P})^\mu$ are not planar.

Acknowledgments. I would like to thank Professor G. Weiss for showing me the private communication [12], Professor D. Hilhorst for introducing me the paper [5], and the referee for carefully reading the manuscript.

REFERENCES

[1] G. Alberti and S. Müller, *A new approach to variational problems with multiple scales*, Comm. Pure Appl. Math., 54 (2001), pp. 761–825.
[2] D. G. Aronson and H. F. Weinberger, *Nonlinear diffusion in population genetics, combustion, and nerve pulse propagation*, in Partial Differential Equations and Related Topics, Lecture Notes in Math. 446, Springer, New York, 1975, pp. 5–49.
[3] A. Bonami, D. Hilhorst, and E. Logak, *Modified motion by mean curvature: Local existence and uniqueness and qualitative properties*, Differential Integral Equations, 13 (2000), pp. 1371–1392.
[4] X. Chen, *Generation and propagation of interfaces in reaction-diffusion systems*, Trans. Amer. Math. Soc., 334 (1992), pp. 877–913.
[5] X. Chen, D. Hilhorst, and E. Logak, *Asymptotic behavior of solutions of an Allen-Cahn equation with a nonlocal term*, Nonlinear Anal., 28 (1997), pp. 1283–1298.
[6] R. Choksi, *Scaling laws in microphase separation of diblock copolymers*, J. Nonlinear Sci., 11 (2001), pp. 223–236.
[7] R. Choksi, R. Kohn, and F. Otto, *Domain branching in uniaxial ferromagnets: A scaling law for the minimum energy*, Comm. Math. Phys., 201 (1999), pp. 61–79.
[8] P. C. Fife, *Dynamics of Internal Layers and Diffusive Interfaces*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 53, SIAM, Philadelphia, 1988.
[9] P. C. Fife and L. Hsiao, *The generation and propagation of internal layers*, Nonlinear Anal., 12 (1988), pp. 19–41.
[10] E. Logak, *Singular limit of reaction-diffusion systems and modified motion by mean curvature*, Proc. Roy. Soc. Edinburgh Sect. A, 132 (2002), pp. 951–973.
[11] L. Modica, *The gradient theory of phase transitions and the minimal interface criterion*, Arch. Ration. Mech. Anal., 98 (1987), pp. 123–142.
[12] S. Müller and G. Weiss, *private communication*, 1996.
[13] Y. Nishiura and I. Ohnishi, *Some mathematical aspects of the micro-phase separation in diblock copolymers*, Phys. D, 84 (1995), pp. 31–39.
[14] I. Ohnishi, Y. Nishiura, M. Imai, and Y. Matsushita, *Analytical solution describing the phase separation driven by a free energy functional containing a long-range interaction term*, Chaos, 9 (1999), pp. 329–341.
[15] T. Ohta and K. Kawasaki, *Equilibrium morphology of block copolymer melts*, Macromolecules, 19 (1986), pp. 2621–2632.
[16] Y. Oshita, *On stable stationary solutions and mesoscopic patterns for FitzHugh-Nagumo equations in higher dimensions*, J. Differential Equations, 188 (2003), pp. 110–134.
[17] P. Sternberg, *The effect of a singular perturbation on nonconvex variational problems*, Arch. Ration. Mech. Anal., 101 (1988), pp. 209–260.
[18] M. Struwe, *Variational Methods; Applications to Nonlinear Partial Differential Equations and Hamiltonian Systems*, Springer, New York, 1996.

SIAM J. MATH. ANAL.
Vol. 36, No. 2, pp. 498–522

VISCOUS SHOCK WAVE TO A GAS-SOLID FREE BOUNDARY PROBLEM FOR COMPRESSIBLE GAS*

FEIMIN HUANG‡, AKITAKA MATSUMURA†, AND XIAODING SHI§

Abstract. We continue to study the large time behavior of the solution to a gas-solid free boundary problem for a one-dimensional model system of compressible viscous gas proposed by us in [*SIAM J. Math. Anal.*, 34 (2003), pp. 1331–1355], where the travelling wave and the rarefaction wave were investigated. In this paper we prove the asymptotic stability of the superposition of a travelling wave and a viscous shock wave under some smallness conditions. The asymptotic behavior of the free boundary is also obtained. The proof is given by the elementary energy estimate.

Key words. viscous shock wave, gas-solid free boundary, compressible gas

AMS subject classification. 35L65

DOI. 10.1137/S003614100240943X

1. Introduction. In this paper, we continue to study the large time behavior of the solution to a gas-solid free boundary problem for a one-dimensional model system of compressible viscous gas proposed by us in [4]. The free boundary value problem reads in the *Eulerian* coordinates as

$$
\begin{cases}
\tilde{\rho}_t + (\tilde{\rho}\tilde{u})_{\tilde{x}} = 0 & \text{in } \tilde{x} > X(t),\\
(\tilde{\rho}\tilde{u})_t + (\tilde{\rho}\tilde{u}^2 + \tilde{p})_{\tilde{x}} = \mu\tilde{u}_{\tilde{x}\tilde{x}} & \text{in } \tilde{x} > X(t),\\
\tilde{\rho}(X(t),t) = \rho_b,\\
\mu\tilde{u}_{\tilde{x}}(X(t),t) = \dfrac{\rho_b\bar{\rho}}{\bar{\rho}-\rho_b}\tilde{u}^2(X(t),t),\\
\dfrac{dX(t)}{dt} = \dfrac{\rho_b}{\rho_b-\bar{\rho}}\tilde{u}(X(t),t), & X(0) = 0,\\
(\tilde{\rho},\tilde{u})|_{(+\infty,t)} = (\rho_+,u_+),\\
(\tilde{\rho},\tilde{u})|_{t=0} = (\rho_0,u_0),
\end{cases}
\tag{1.1}
$$

where $\tilde{\rho}(\tilde{x},t)$ denotes the density of gas, $\tilde{u}(\tilde{x},t)$ denotes the velocity, and $\tilde{p} = \tilde{\rho}^\gamma$ $1 \le \gamma \le 3$ denotes the pressure; where the viscosity coefficient μ and the density $\bar{\rho}$ of the solid are given positive constants; and where

$$
\rho_b < \bar{\rho}, \qquad \frac{dX(t)}{dt} < 0. \tag{1.2}
$$

As shown in [4], the part $\tilde{x} > X(t)$ is filled by the gas with the density $\tilde{\rho}(\tilde{x},t)$ and velocity $\tilde{u}(\tilde{x},t)$ satisfying the conservation of the mass and momentum, the part $\tilde{x} < X(t)$ is filled by the solid with the constant density $\bar{\rho}$ and zero velocity, and the phase transition from the solid to the gas occurs on the free boundary $\tilde{x} = X(t)$. It

*Received by the editors June 11, 2002; accepted for publication (in revised form) October 3, 2003; published electronically July 29, 2004.

http://www.siam.org/journals/sima/36-2/40943.html

†Department of Mathematics, Graduate School of Science, Osaka University, Osaka 560-0045, Japan (fhuang@mail.amt.ac.cn, akitaka@ist.osaka-u.ac.jp, shixd@mail.buct.edu.cn). The first author was supported in part by the JSPS Research Fellowship for foreign researchers and Grant-in-Aid P-00269 for JSPS from the Ministry of Education, Science, Sports and Culture of Japan.

‡Institute of Applied Mathematics, AMSS, Academia Sinica, Beijing 100080, China.

§Department of Mathematics, Graduate School of Science, Beijing University of Technology and Chemical, Beijing 100029, China.

is interesting to compare our free boundary problem (1.1), (1.2) with the previous ones. In all previous works (e.g., Kazhikhov [8], Nagasawa [16]) they assume $\frac{dX(t)}{dt} = \tilde{u}(X(t),t)$. In this case, if we introduce the Lagrangian mass coordinates, we can reformulate the problem to that with the fixed boundary. On the other hand, in our case we cannot do it, which gives us a major difficulty. For the related free boundary problem, see [4, 6, 8, 16] and references therein.

Now let us turn to our free boundary problem (1.1), (1.2). In [4] we proved the free boundary problem (1.1), (1.2) admits a travelling wave under some conditions. We then proved the asymptotic stability of the superposition of a travelling wave and a rarefaction wave under some smallness conditions. However, the viscous shock wave case was left open, and the asymptotic behavior of the free boundary $\tilde{x} = X(t)$ is not obtained yet.

In this paper we investigate the case when the asymptotic state is given by the combination of a travelling wave and a viscous shock wave. We further show the asymptotic behavior of the free boundary $\tilde{x} = X(t)$. The main novelty of this paper is to determine the phase shift of the viscous shock wave. As we know, it is difficult to locate the phase shift even for the viscous scalar conservation laws with boundary (see [9, 10]). For 2×2 compressible Navier–Stokes equations with the fixed boundary, Matsumura and Mei [12] developed a new technique to calculate the shift. In their case, only conservation of mass was used to determine the shift. Because the problem (1.1), (1.2) is a free boundary problem, our case is much more difficult than [12] in many aspects. Since the velocity $\tilde{u}(\tilde{x},t)$ on the free boundary is unknown, conservation of momentum has to be used here. Our main difficulty comes from the free boundary $\tilde{x} = X(t)$. To treat the free boundary problem more easily, we study it in the Lagrangian coordinates instead of Eulerian coordinates. The free boundary in the Lagrangian coordinates (x,t) becomes $x = x(t) = \bar{\rho}X(t)$. When the initial data is closed to the superposition of a travelling wave with speed $\bar{s} < 0$ and a 2-viscous shock wave, we expect the free boundary $x = x(t)$ to tend to $x = \bar{s}t +$ constant as t tends to infinity. To overcome the main difficulty from the free boundary, we manipulate both conservation laws, and then the phase shift α is explicitly determined. Furthermore, the asymptotic behavior of the free boundary is consequently given after the large time behavior of the solution is obtained. Namely, $x(t) - \bar{s}t$ converges to a constant as time tends to infinity. We note that our results naturally include the case that the asymptotic state is given by a single travelling wave. This case was also considered in [4]; however, the asymptotic behavior of the free boundary $x = x(t)$ was not yet obtained there. In this sense, we also improve some results of [4]. We now formulate our main result.

As in [4], we transform (1.1) into the problem in the Lagrangian coordinates:

$$
\begin{cases}
v_t - u_x = 0, & x > x(t),\ t > 0,\\
u_t + p(v)_x = \mu\left(\dfrac{u_x}{v}\right)_x, & x > x(t),\ t > 0,\\
v(x(t),t) = v_b,\\
\mu u_x(x(t),t) = \dfrac{v_b}{v_b - \bar{v}}u^2(x(t),t),\\
\dfrac{dx(t)}{dt} = \dfrac{1}{\bar{v} - v_b}u(x(t),t), & x(0) = 0,\\
(v,u)|_{(+\infty,t)} = (v_+,u_+) = \left(\dfrac{1}{\rho_+},u_+\right),\\
(v,u)|_{t=0} = (v_0,u_0), & v_0(0) = v_b,
\end{cases}
\tag{1.3}
$$

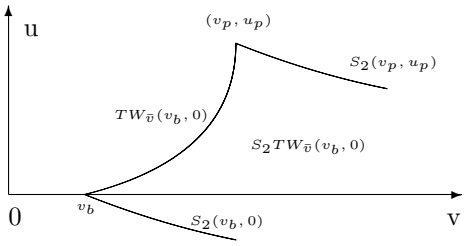

FIG. 1.1.

where $v = \frac{1}{\rho}$, $p(v) = v^{-\gamma}$, $1 \le \gamma \le 3$, and $x(t) = \bar\rho X(t)$. By the argument of [4], a new travelling wave solution (simply denoted by TW-solution) exists for the free boundary problem (1.1), (1.2). For any $0 < \bar v < v_b < +\infty$, the TW-curve with the parameter $\bar v$ through the point $(v_b, 0)$ is defined by

(1.4)
$$TW_{\bar v}(v_b, 0) = \{(v,u); u = (v-\bar v)^{\frac12}(p(v_b)-p(v))^{\frac12},\ u^2 < (v-\bar v)^2|p'(v)|,\ v > v_b\}.$$

This shows that for any $(v_*, u_*) \in TW_{\bar v}(v_b, 0)$ there exists a travelling wave solution $(V_T, U_T)(\xi)$, $\xi = x - \bar s t$, satisfying

$$\begin{cases} -\bar s V_T' - U_T' = 0, \\ -\bar s U_T' + p(V_T)' = \mu\left(\dfrac{U_T'}{V_T}\right)', \end{cases} \tag{1.5}$$

with

$$\begin{cases} V_T(0) = v_b, \qquad U_T(0) = u_b, \qquad \mu U_T'(0) = v_b(v_b - \bar v)\bar s^2, \\ V_T(+\infty) = v_*, \qquad U_T(+\infty) = u_*, \end{cases} \tag{1.6}$$

where

$$\bar s = \frac{u_b}{\bar v - v_b}, \qquad u_b = \frac{\bar v - v_b}{\bar v - v_*}u_*. \tag{1.7}$$

That is, the TW-solution $(V_T, U_T)(\xi)$ connects (v_b, u_b) and (v_*, u_*). On the other hand, it is known that the 2-shock curve $S_2(v_b, 0)$ through the point $(v_b, 0)$ is

$$S_2(v_b, 0) = \{(v,u); u = -s_b(v - v_b),\ v > v_b\}, \qquad s_b = \sqrt{\frac{p(v_b)-p(v)}{v - v_b}}, \tag{1.8}$$

and $S_2(v_p, u_p)$ through the point (v_p, u_p) is

$$S_2(v_p, u_p) = \{(v,u); u = u_p - s_p(v - v_p),\ v > v_p\}, \qquad s_p = \sqrt{\frac{p(v_p)-p(v)}{v - v_p}}, \tag{1.9}$$

where

$$u_p = (v_p - \bar v)^{\frac12}(p(v_b)-p(v_p))^{\frac12}, \qquad u_p^2 = (v_p - \bar v)^2|p'(v_p)|. \tag{1.10}$$

We define $S_2TW_{\bar v}(v_b,0)$ as the domain surrounded by curves $TW_{\bar v}(v_b,0)$, $S_2(v_b,0)$, and $S_2(v_p,u_p)$ (see Figure 1.1).

Now let us turn to the asymptotic behavior of the solution. When the right end state (v_+,u_+) of (1.3) belongs to the region $S_2TW_{\bar v}(v_b,0)$ in the phase plane, the solution of (1.3) is expected to tend to the superposition of a TW-solution and a 2-viscous shock wave as t tends to infinity. Our aim is to justify the above conjecture. Our result is, roughly speaking, as follows.

If $(v_+,u_+)\in S_2TW_{\bar v}(v_b,0)$, then there exists $(v_*,u_*)\in TW_{\bar v}(v_b,0)$ such that $(v_+,u_+)\in S_2(v_*,u_*)$, and the superposition of the TW-solution connecting (v_b,u_b) with (v_*,u_*) and the 2-viscous shock wave connecting (v_*,u_*) with (v_+,u_+) is stable, provided that $|v_*-v_b|$ is small. That is, the TW-solution is necessarily weak; however, the 2-viscous shock wave is not necessarily weak.

Finally, we note that when the free boundary $\tilde x = X(t)$ is fixed as a straight line, the problem (1.1), (1.2) becomes the so-called inflow problem (see [4]). We refer the reader to [3, 11, 15, 17] for the inflow problem.

Our plan for this paper is as follows. In section 2, we give some known results on the travelling wave and the viscous shock wave; in section 3, we study the phase shift of the viscous shock wave; in section 4, we state our main result; in section 5, we reformulate the original problem to a new initial-boundary value problem; in section 6, we show the local existence of the solution; in section 7, we establish the a priori estimates and prove our main result.

Notation. Throughout this paper, several positive generic constants are denoted by c, C without confusion. For function spaces, $L^p(\Omega)$, $1\le p\le\infty$, denotes a usual Lebesgue space on $\Omega\subset R=(-\infty,\infty)$ with its norm

$$\|f\|_{L^p(\Omega)}=\left(\int_\Omega |f(x)|^p dx\right)^{\frac1p},\quad 1\le p<\infty,\qquad \|f\|_{L^\infty(\Omega)}=\sup_\Omega |f(x)|. \tag{1.11}$$

$H^l(\Omega)$ denotes the lth order Sobolev space with its norm

$$\|f\|_l=\left(\sum_{j=0}^{l}\|\partial_x^j f\|^2\right)^{\frac12}\quad \text{when } \|\cdot\|:=\|\cdot\|_{L^2(\Omega)}. \tag{1.12}$$

The domain Ω will often be abbreviated without confusion.

2. Preliminaries. In this section we recall some properties of the travelling wave solution and the viscous shock wave. From [4], we have the following result of the travelling wave solution.

LEMMA 2.1. *For any given $\bar v$, v_b, and v_* with $0<\bar v<v_b<v_*$, let u_* be the number such that $(v_*,u_*)\in TW_{\bar v}(v_b,0)$. Then there exists a unique solution $(V_T,U_T)(\xi)$, $\xi=x-\bar s t$, to (1.5) and (1.6) satisfying $0<v_b<V_T(\xi)<v_*$, $V_T'>0$, where $\bar s$ and u_b are given by (1.7). Furthermore, fix $\bar v$ and v_b, and let $v_*-v_b=\delta$; then there exist positive constants $\delta_0>0$ and $c_0>0$ such that, for any $\delta\le\delta_0$,*

$$|V_T(\xi)-v_*|=O(\delta)e^{-\frac{c_0}{\sqrt\delta}\xi},\qquad |U_T(\xi)-u_*|=O(\delta^{\frac32})e^{-\frac{c_0}{\sqrt\delta}\xi}, \tag{2.1}$$

and

$$(u_*,\bar s,u_b)=O(\delta^{\frac12}),\qquad V_T'=O(\delta^{\frac12})e^{-\frac{c_0}{\sqrt\delta}\xi},\qquad V_T''=O(1)e^{-\frac{c_0}{\sqrt\delta}\xi}. \tag{2.2}$$

On the other hand, it is well known that for any point (v_*, u_*) with $v_* > 0$, if $(v_+, u_+) \in S_2(v_*, u_*)$, then there exists a unique viscous shock profile $(v, u) = (V_s, U_s)(\eta = x - st)$, $s = \sqrt{\frac{p(v_*)-p(v_+)}{v_+-v_*}} > 0$, satisfying $(V_s, U_s)(-\infty) = (v_*, u_*)$, $(V_s, U_s)(+\infty) = (v_+, u_+)$ up to a shift. Namely, (V_s, U_s) satisfies

$$\begin{cases} -sV_s' - U_s' = 0, \\ -sU_s' + p(V_s)' = \mu\left(\dfrac{U_s'}{V_s}\right)', \\ (V_s, U_s)(-\infty) = (v_*, u_*), \\ (V_s, U_s)(+\infty) = (v_+, u_+), \end{cases} \tag{2.3}$$

which yields

$$\frac{s\mu V_s'}{V_s} = -s^2 V_s - p(V_s) - b \equiv: h(V_s), \tag{2.4}$$

where $b = -s^2 v_* - p(v_*) = -s^2 v_+ - p(v_+)$. Thus, we have the following lemma.

LEMMA 2.2. *Let (v_*, u_*) be a point in the phase plane with $v_* > 0$. Suppose that $(v_+, u_+) \in S_2(v_*, u_*)$; then there exists a unique shock profile $(V_s, U_s)(\eta = x - st)$, $s = \sqrt{\frac{p(v_*)-p(v_+)}{v_+-v_*}} > 0$, up to a shift, which connects (v_*, u_*) and (v_+, u_+) and satisfies* (2.3) *and*

$$\begin{aligned} &0 < v_* < V_s(\eta) < v_+, u_+ < U_s(\eta) < u_*, h(V_s) > 0, V_s' = \frac{V_s h(V_s)}{s\mu} > 0, \\ &(|V_s(\eta) - v_*|, |U_s(\eta) - u_*|) = O(1)|v_+ - v_*|e^{-c_-|\eta|} \quad \text{as } \eta \to -\infty, \\ &(|V_s(\eta) - v_+|, |U_s(\eta) - u_+|) = O(1)|v_+ - v_*|e^{-c_+|\eta|} \quad \text{as } \eta \to +\infty, \end{aligned} \tag{2.5}$$

where $c_- = \frac{v_*|p'(v_*)+s^2|}{\mu s} > 0$, $c_+ = \frac{v_+|p'(v_+)+s^2|}{\mu s} > 0$.

3. Phase shift. This section is devoted to the phase shift of the viscous shock profile. When $(v_+, u_+) \in S_2 TW_{\bar v}(v_b, 0)$, there is $(v_*, u_*) \in TW_{\bar v}(v_b, 0)$ such that $(v_+, u_+) \in S_2(v_*, u_*)$, and the asymptotic behavior of the solution to (1.3) is expected to be the superposition of a TW-solution $(V_T, U_T)(\xi)$, $\xi = x - \bar s t$, connecting (v_b, u_b) with (v_*, u_*) and a 2-viscous shock wave $(V_s, U_s)(\eta)$, $\eta = x - st$, connecting (v_*, u_*) with (v_+, u_+), where

$$\bar s = \frac{u_*}{\bar v - v_*}, \qquad u_b = \frac{\bar v - v_b}{\bar v - v_*}u_*, \qquad s = \sqrt{\frac{p(v_*) - p(v_+)}{v_+ - v_*}}. \tag{3.1}$$

We consider the situation where the initial data $(v_0(x), u_0(x))$ are given in a neighborhood of $(V_T(x) + V_s(x - \beta) - v_*, U_T(x) + U_s(x - \beta) - u_*)$ for some large constant $\beta > 0$. That is, we ask the viscous shock wave to be suitably far from the boundary initially. We now try to locate the phase shift α such that the asymptotic state of the solution (v, u) to (1.3) is given by $(V_T(x - x(t)) + V_s(x - st + \alpha - \beta) - v_*, U_T(x - x(t)) + U_s(x - st + \alpha - \beta) - u_*)$.

To investigate the phase shift α, we consider a coordinate transformation

$$t = t, \quad y = x - x(t), \tag{3.2}$$

in which we can make the free boundary problem (1.3) easier to handle. Let $x(t) = \bar{s}t + \gamma(t)$. Then we rewrite (1.3) as

$$
\begin{cases}
v_t - (\bar{s} + \gamma'(t))v_y - u_y = 0, & y > 0,\ t > 0,\\
u_t - (\bar{s} + \gamma'(t))u_y + p(v)_y = \mu\left(\dfrac{u_y}{v}\right)_y, & y > 0,\ t > 0,\\
v(0,t) = v_b,\\
\mu u_y(0,t) = \dfrac{v_b}{v_b - \bar{v}}u^2(0,t),\\
\dfrac{d\gamma(t)}{dt} = \dfrac{1}{\bar{v} - v_b}u(0,t) - \bar{s}, & \gamma(0) = 0,\\
(v,u)|_{(+\infty,t)} = (v_+, u_+),\\
(v,u)(y,0) = (v_0,u_0)(y), & v_0(0) = v_b.
\end{cases}
\tag{3.3}
$$

On the other hand, Lemma 2.1 yields the TW-solution $(V_T, U_T)(y)$ connecting (v_b, u_b), and (v_*, u_*) satisfies

$$
\begin{cases}
-\bar{s}V_{Ty} - U_{Ty} = 0, & y > 0,\\
-\bar{s}U_{Ty} + p(V_T)_y = \mu\left(\dfrac{U_{Ty}}{V_T}\right)_y, & y > 0,\\
V_T(0) = v_b, \quad U_T(0) = u_b,\\
\mu U_{Ty}(0) = v_b(v_b - \bar{v})\bar{s}^2,\\
(V_T, U_T)|_{(+\infty)} = (v_*, u_*).
\end{cases}
\tag{3.4}
$$

In the new coordinate system (y,t), the asymptotic shock wave should take the form

$$
(V_s, U_s)(y,t) = (V_s, U_s)(y + x(t) - st + \alpha - \beta).
\tag{3.5}
$$

Note that $x(t)$ is an unknown function, and it is very difficult to determine the phase shift α if we directly use the form (3.5). Since the free boundary $x(t)$ is expected to tend to $\bar{s}t + \text{const}$ as t tends to infinity, we use the shock profile

$$
(\bar{V}_s, \bar{U}_s)(y,t) = (V_s, U_s)(y - (s - \bar{s})t + \alpha - \beta)
\tag{3.6}
$$

instead of the profile (3.5) and then estimate the L^1 distance between (V_s, U_s) and $(\bar{V}_s, \bar{U}_s)$. By Lemma 2.2, the shock profile (3.6) satisfies

$$
\begin{cases}
\bar{V}_{st} - \bar{s}\bar{V}_{sy} - \bar{U}_{sy} = 0, & y > 0,\ t > 0,\\
\bar{U}_{st} - \bar{s}\bar{U}_{sy} + p(\bar{V}_s)_y = \mu\left(\dfrac{\bar{U}_{sy}}{\bar{V}_s}\right)_y, & y > 0,\ t > 0,\\
(\bar{V}_s, \bar{U}_s)(-\infty,t) = (v_*, u_*),\\
(\bar{V}_s, \bar{U}_s)(+\infty,t) = (v_+, u_+).
\end{cases}
\tag{3.7}
$$

Let

$$
\begin{aligned}
w(y,t) &= v(y,t) - V_T(y) - V_s(y,t) + v_*,\\
z(y,t) &= u(y,t) - U_T(y) - U_s(y,t) + u_*,
\end{aligned}
\tag{3.8}
$$

and

$$
\begin{aligned}
\bar{w}(y,t) &= v(y,t) - V_T(y) - \bar{V}_s(y,t) + v_*,\\
\bar{z}(y,t) &= u(y,t) - U_T(y) - \bar{U}_s(y,t) + u_*.
\end{aligned}
\tag{3.9}
$$

Then we have from (3.3)–(3.4), (3.7), and (3.9)

$$\begin{cases} \bar{w}_t - \bar{s}\bar{w}_y - \bar{z}_y = \gamma'(t)v_y, & y>0,\ t>0,\\ \bar{z}_t - \bar{s}\bar{z}_y + P_y - Q_y = \gamma'(t)u_y, & y>0,\ t>0, \end{cases} \tag{3.10}$$

where

$$\begin{aligned} P(y,t) &= p(v(y,t)) - p(V_T(y)) - p(\bar{V}_s(y,t)) + p(v_*),\\ Q(y,t) &= \mu\left(\frac{u_y(y,t)}{v(y,t)} - \frac{U_{Ty}(y)}{V_T(y)} - \frac{\bar{U}_{sy}(y,t)}{\bar{V}_s(y,t)}\right). \end{aligned} \tag{3.11}$$

Integrating $(3.10)_1$ over R_+ yields

$$\begin{aligned} \frac{d}{dt}\int_0^\infty \bar{w}(y,t)dy &= -\bar{s}\bar{w}(0,t) - \bar{z}(0,t) + \gamma'(t)(v_+ - v_b)\\ &= -\bar{s}(v_* - \bar{V}_s(0,t)) - (u(0,t) - u_b)\\ &\quad + (\bar{U}_s(0,t) - u_*) + \gamma'(t)(v_+ - v_b)\\ &= (s-\bar{s})(v_* - \bar{V}_s(0,t)) + \gamma'(t)(v_+ - \bar{v}), \end{aligned} \tag{3.12}$$

where we have used the fact that

$$\bar{U}_s(0,t) - u_* = s(v_* - \bar{V}_s(0,t)), \qquad u(0,t) - u_b = \gamma'(t)(\bar{v} - v_b). \tag{3.13}$$

In the same fashion, integrating $(3.10)_2$ over R_+, by the fact that

$$-s(\bar{U}_s(0,t) - u_*) + p(\bar{V}_s(0,t)) - p(v_*) = \mu\frac{\bar{U}_{sy}(0,t)}{\bar{V}_s(0,t)}, \tag{3.14}$$

yields

$$\begin{aligned} \frac{d}{dt}\int_0^\infty \bar{z}(y,t)dy &= -\bar{s}\bar{z}(0,t) + P(0,t) - Q(0,t) + \gamma'(t)(u_+ - u(0,t))\\ &= -u_b\gamma'(t) + (s-\bar{s})(u_* - \bar{U}_s(0,t))\\ &\quad + \gamma'(t)(u(0,t) + u_b) + \gamma'(t)(u_+ - u(0,t))\\ &= u_+\gamma'(t) + (s-\bar{s})(u_* - \bar{U}_s(0,t)). \end{aligned} \tag{3.15}$$

Here we are very lucky that the term $\gamma'(t)u(0,t)$ is cancelled in (3.15). Integrating (3.12) and (3.15) over $(0,t)$ gives

$$\int_0^\infty \bar{w}(y,t)dy = \gamma(t)(v_+ - \bar{v}) + (s-\bar{s})\int_0^t v_* - \bar{V}_s(0,t)dt + \int_0^\infty \bar{w}(y,0)dy, \tag{3.16}$$

$$\int_0^\infty \bar{z}(y,t)dy = \gamma(t)u_+ + (s-\bar{s})\int_0^t u_* - \bar{U}_s(0,t)dt + \int_0^\infty \bar{z}(y,0)dy. \tag{3.17}$$

On the other hand, we compute

$$\begin{aligned} &\int_0^\infty \bar{V}_s(y,t) - V_s(y,t)dy\\ &\quad = -\int_0^\infty\int_0^1 V_{sy}(y + \bar{s}t - st + \alpha - \beta + \theta\gamma(t))d\theta dy \times \gamma(t)\\ &\quad = -\int_0^1 [v_+ - V_s(\bar{s}t - st + \alpha - \beta + \theta\gamma(t))]d\theta \times \gamma(t)\\ &\quad = (v_* - v_+)\gamma(t) + \sigma(t) \end{aligned} \tag{3.18}$$

and

$$\int_0^\infty \bar{U}_s(y,t) - U_s(y,t)dy = (u_* - u_+)\gamma(t) - s\sigma(t), \tag{3.19}$$

where

$$\sigma(t) = \sigma(\gamma,t) = \int_0^1 [V_s(\bar{s}t - st + \alpha - \beta + \theta\gamma(t)) - v_*]d\theta \times \gamma(t). \tag{3.20}$$

It is noted that $\bar{s}t + \theta\gamma(t) < 0$ holds for any $0 \le \theta \le 1$ due to $\bar{s} < 0$ and $x(t) < 0$. Thus, if $\alpha - \beta < 0$, then $\sigma(t)$ has the following estimate:

$$|\sigma(t)| \le C|\gamma(t)|e^{-c_-|-st+\alpha-\beta|}. \tag{3.21}$$

From (3.16)–(3.19), we have

$$\begin{aligned}\int_0^\infty w(y,t)dy &= \int_0^\infty \bar{w}(y,t)dy + \int_0^\infty \bar{V}_s(y,t) - V_s(y,t)dy\\ &= \gamma(t)(v_* - \bar{v}) + (s - \bar{s})\int_0^t v_* - \bar{V}_s(0,t)dt\\ &\quad + \int_0^\infty w(y,0)dy + \sigma(t),\end{aligned} \tag{3.22}$$

$$\begin{aligned}\int_0^\infty z(y,t)dy &= \int_0^\infty \bar{z}(y,t)dy + \int_0^\infty \bar{U}_s(y,t) - U_s(y,t)dy\\ &= \gamma(t)u_* + (s - \bar{s})\int_0^t u_* - \bar{U}_s(0,t)dt\\ &\quad + \int_0^\infty z(y,0)dy - s\sigma(t).\end{aligned} \tag{3.23}$$

Combining (3.22) and (3.23) implies

$$\begin{aligned}&u_*\int_0^\infty w(y,t)dy - (v_* - \bar{v})\int_0^\infty z(y,t)dy\\ &\quad = a(s - \bar{s})\int_0^t (v_* - \bar{V}_s(0,t))dt\\ &\qquad + \int_0^\infty [u_* w(y,0) - (v_* - \bar{v})z(y,0)]dy + a\sigma(t),\end{aligned} \tag{3.24}$$

where $a = u_* + s(v_* - \bar{v}) > 0$. Expecting $[u_*\int_0^\infty w(y,t)dy - (v_* - \bar{v})\int_0^\infty z(y,t)dy]|_{t=+\infty} = 0$, $\sigma(t)|_{t=+\infty} = 0$ yields

$$\begin{aligned}I(\alpha) &:= a(s - \bar{s})\int_0^\infty (v_* - V_s(\bar{s}t - st + \alpha - \beta))dt\\ &\quad + u_*\int_0^\infty [v_0(y) - V_T(y) - V_s(y + \alpha - \beta) + v_*]dy\\ &\quad - (v_* - \bar{v})\int_0^\infty [u_0(y) - U_T(y) - U_s(y + \alpha - \beta) + u_*]dy\\ &= 0.\end{aligned} \tag{3.25}$$

Since

$$\begin{aligned} I'(\alpha) &= a(v_* - V_s(\alpha-\beta)) - a(v_+ - V_s(\alpha-\beta)) \\ &= a(v_* - v_+) < 0, \end{aligned} \tag{3.26}$$

the equality $0 = I(\alpha) = I(0) + a(v_* - v_+)\alpha$ determines α by

$$\begin{aligned} \alpha &= -\frac{1}{a(v_*-v_+)} I(0) \\ &= -\frac{1}{a(v_*-v_+)} \Big\{ a(s-\bar{s}) \int_0^\infty (v_* - V_s(\bar{s}t - st - \beta))dt \\ &\qquad + u_* \int_0^\infty [v_0(y) - V_T(y) - V_s(y-\beta) + v_*]dy \\ &\qquad - (v_* - \bar{v}) \int_0^\infty [u_0(y) - U_T(y) - U_s(y-\beta) + u_*]dy \Big\}. \end{aligned} \tag{3.27}$$

Substituting (3.25) into (3.24) gives

$$\begin{aligned} & u_* \int_0^\infty w(y,t)dy - (v_* - \bar{v}) \int_0^\infty z(y,t)dy \\ &\quad = -a(s-\bar{s}) \int_t^\infty (v_* - V_s(\bar{s}t - st + \alpha - \beta))dt + a\sigma(t) \\ &\quad =: A(t) + a\sigma(t). \end{aligned} \tag{3.28}$$

4. Main result. Let

$$\begin{aligned} V(y,t;\alpha,\beta) &= V_T(y) + V_s(y + x(t) - st + \alpha - \beta) - v_*, \\ U(y,t;\alpha,\beta) &= U_T(y) + U_s(y + x(t) - st + \alpha - \beta) - u_*, \end{aligned} \tag{4.1}$$

where $\alpha = \alpha(\beta)$ is given by (3.27). We suppose that for some $\beta > 0$,

$$v_0(y) - V(y,0;0,\beta) \in H^1 \cap L^1, \qquad u_0(y) - U(y,0;0,\beta) \in H^1 \cap L^1. \tag{4.2}$$

Set

$$(\Phi_0, \Psi_0)(y) = -\int_y^\infty (v_0(y) - V(y,0;0,\beta), u_0(y) - U(y,0;0,\beta))dy. \tag{4.3}$$

Assume that

$$(\Phi_0, \Psi_0) \in L^2. \tag{4.4}$$

We now give our main result.

THEOREM 4.1. *Suppose that* $1 \le \gamma \le 3$ *and* $(v_+, u_+) \in S_2 TW_{\bar{v}}(v_b, 0)$. *Then there exists* (v_*, u_*) *such that* $(v_*, u_*) \in TW_{\bar{v}}(v_b, 0)$ *and* $(v_+, u_+) \in S_2(v_*, u_*)$. *Assume that* (4.2) *and* (4.4) *hold and*

$$(\gamma - 1)^2 (v_+ - v_*) < 2\gamma v_*. \tag{4.5}$$

Then there exists a positive constant δ_0 *such that for any given* $0 < v_* - v_b = \delta < \delta_0$, *if*

$$\sqrt{\beta} \|\Phi_0, \Psi_0\|_2 + e^{-\frac{2}{5}c_-\beta} < \delta^2, \tag{4.6}$$

then (3.3) *has a unique global solution* $((v,u)(y,t),\gamma(t))$ *satisfying*

$$v(y,t)-V(y,t;\alpha,\beta)\in C^0([0,\infty),H^1)\cap L^2(0,\infty;H^1), \tag{4.7}$$

$$u(y,t)-U(y,t;\alpha,\beta)\in C^0([0,\infty),H^1)\cap L^2(0,\infty;H^2), \tag{4.8}$$

$$\gamma(t)\in C^1[0,\infty), \tag{4.9}$$

and

$$\sup_{y\in\Re_+}|(v,u)(y,t)-(V,U)(y,t;\alpha,\beta)|\longrightarrow 0 \quad \text{as } t\to+\infty, \tag{4.10}$$

where $\alpha=\alpha(\beta)$ *is determined by* (3.27). *Furthermore,* $\gamma(t)$ *converges to a constant* Γ *as* $t\to\infty$, *where*

$$\begin{aligned}\Gamma=\frac{1}{\bar v-v_*}\Big\{&(s-\bar s)\int_0^\infty[v_*-V_s(-(s-\bar s)t+\alpha-\beta)]dt\\&+\int_0^\infty[v_0(y)-V_T(y)-V_s(y+\alpha-\beta)+v_*]dy\Big\}.\end{aligned} \tag{4.11}$$

Remark 4.2. In Theorem 4.1, the strength of the viscous shock wave is not necessarily weak. For general pressure p, the same result could be easily proved by the same argument if the strength of the viscous shock wave is suitably small.

5. Reformulated system. Let

$$\phi(y,t)=-\int_y^\infty w(y,t)dy=-\int_y^\infty[v(y,t)-V(y,t;\alpha,\beta)]dy,$$
$$\psi(y,t)=-\int_y^\infty z(y,t)dy=-\int_y^\infty[u(y,t)-U(y,t;\alpha,\beta)]dy.$$

We put the perturbation by

$$\begin{aligned}v(y,t)&=\phi_y(y,t)+V(y,t;\alpha,\beta),\\u(y,t)&=\psi_y(y,t)+U(y,t;\alpha,\beta).\end{aligned} \tag{5.1}$$

Note that the viscous shock profile $(V_s,U_s)(y,t)$ of (3.5) satisfies, from Lemma 2.2,

$$\begin{cases}V_{st}-(\bar s+\gamma')V_{sy}-U_{sy}=0, & y>0,\ t>0,\\ U_{st}-(\bar s+\gamma')U_{sy}+p(V_s)_y=\mu\left(\dfrac{U_{sy}}{V_s}\right)_y, & y>0,\ t>0,\\ (V_s,U_s)(-\infty,t)=(v_*,u_*),\\ (V_s,U_s)|_{(+\infty,t)}=(v_+,u_+).\end{cases} \tag{5.2}$$

Substituting (3.4), (5.1)–(5.2) into (3.3) and integrating the system on $[y,+\infty)$, we have

$$\begin{cases}\phi_t-(\bar s+\gamma')\phi_y-\psi_y=(V_T(y)-v_*)\gamma',\\ \psi_t-(\bar s+\gamma')\psi_y+p(V+\phi_y)-p(V_T)-p(V_s)+p(v_*)\\ \quad=\mu\left(\dfrac{U_y+\psi_{yy}}{V+\phi_y}-\dfrac{U_{Ty}}{V_T}-\dfrac{U_{sy}}{V_s}\right)+(U_T(y)-u_*)\gamma'.\end{cases} \tag{5.3}$$

We now investigate the initial and boundary conditions of the reformulated system (5.3). By the definition, the initial data satisfies

$$\begin{aligned}\phi(\xi,0) &= -\int_y^{+\infty}[v_0(y)-V(y,0;\alpha,\beta)]dy\\ &= \Phi_0(y)+\int_y^{\infty}\int_0^{\alpha}V_s'(y+\theta-\beta)d\theta dy \qquad (5.4)\\ &= \Phi_0(y)+\int_0^{\alpha}[v_+-V_s(y+\theta-\beta)]d\theta =: \phi_0(y),\end{aligned}$$

$$\psi(y,0)=\Psi_0(y)+\int_0^{\alpha}[u_+-U_s(y+\theta-\beta)]d\theta =: \psi_0(y). \qquad (5.5)$$

Furthermore, we have the following property of $\phi_0(y)$ and $\psi_0(y)$.

PROPOSITION 5.1. *Under assumptions* (4.2) *and* (4.4)*, there exists a positive constant* $C_0>0$ *such that the initial perturbations* $(\phi_0,\psi_0)\in H^2$*, and it satisfies*

$$\|(\phi_0,\psi_0)\|_2 \le C_0(\sqrt{\beta}\|(\Phi_0,\Psi_0)\|_2+e^{-\frac{1}{2}c_-\beta}). \qquad (5.6)$$

Proof. From (3.27), we have

$$|\alpha| \le C(|\Phi_0|+|\Psi_0|+e^{-c_-|st+\beta|}) \le C(\|\Phi_0\|_2+\|\Psi_0\|_2+e^{-c_-\beta}). \qquad (5.7)$$

Thus by the same argument of [3, Proposition 3.1], Proposition 5.1 is proved. ∎

We now show the boundary conditions. Let $B(\gamma,t)=\bar{s}t-st+\alpha-\beta+\gamma(t)$ and $\sigma(\gamma,t)=\sigma(t)$. By (3.3), (3.22), and (3.23), we have

$$\gamma'(t) = \frac{1}{\bar{v}-v_b}(\psi_y(0,t)+U_s(B(\gamma,t))-u_*), \qquad (5.8)$$

$$\begin{aligned}\phi(0,t) &= -\gamma(t)(v_*-\bar{v})-(s-\bar{s})\int_0^t[v_*-\bar{V}_s(0,t)]dt+\phi_0(0)-\sigma(t)\\ &=: C(\gamma,t)+\phi_0(0)-\sigma(t),\end{aligned} \qquad (5.9)$$

and

$$\begin{aligned}\psi_t(0,t) &= \frac{U_s(B(\gamma,t))}{v_b-\bar{v}}\psi_y(0,t)+\left[(s-\bar{s})+\frac{U_s(B(\gamma,t))}{v_b-\bar{v}}\right](U_s(B(\gamma,t))-u_*)\\ &=: D(\gamma,t)\psi_y(0,t)+E(\gamma,t).\end{aligned} \qquad (5.10)$$

We note that

$$\phi_y(0,t)=v_*-V_s(B(\gamma,t)) \qquad (5.11)$$

holds as a compatibility condition when we substitute (5.8)–(5.9) into $(5.3)_1$. Thus

our reformulated system is

$$
\begin{cases}
\phi_t - (\bar{s}+\gamma')\phi_y - \psi_y = (V_T(y)-v_*)\gamma', \\
\psi_t - (\bar{s}+\gamma')\psi_y + p(V+\phi_y) - p(V_T) - p(V_s) + p(v_*) \\
\quad = \mu\left(\dfrac{U_y+\psi_{yy}}{V+\phi_y} - \dfrac{U_{Ty}}{V_T} - \dfrac{U_{sy}}{V_s}\right) + (U_T(y)-u_*)\gamma', \\
(\phi,\psi)(y,0) = (\phi_0,\psi_0)(y), \\
\phi(0,t) = C(\gamma,t) + \phi_0(0) - \sigma(t), \\
\psi(0,t) = \psi_0(0) + \displaystyle\int_0^t D(\gamma,t)\psi_y(0,t)dt + \int_0^t E(\gamma,t)dt, \\
\gamma'(t) = \dfrac{1}{\bar{v}-v_b}(\psi_y(0,t) + U_s(B(\gamma,t)) - u_*), \quad \gamma(0) = 0.
\end{cases} \tag{5.12}
$$

6. Local existence. For any interval $I \subset \Re_+$, we define the solution space $X(I)$ by

$$
\begin{aligned}
X(I) = \Big\{ & (\phi,\psi) \in C^0(I;H^2);\ \phi_y \in L^2(I;H^1), \\
& \psi_y \in L^2(I;H^2),\ \sup_{t\in I} \|(\phi,\psi)(t)\|_2 \le \varepsilon_1 \Big\},
\end{aligned} \tag{6.1}
$$

where $\varepsilon_1 = \frac{1}{2}v_b$. Let

$$
N(t) = \sup_{0\le\tau\le t}(\|\phi(\tau)\|_2 + \|\psi(\tau)\|_2), \qquad N_0 = \|\phi_0\|_2 + \|\psi_0\|_2. \tag{6.2}
$$

By the Sobolev embedding theorem, for $(\phi,\psi) \in X([0,T])$, one obtains

$$
(V+\phi_y)(y,t) \ge v_b - \|\phi_y\|_1 \ge \frac{1}{2}v_b, \qquad (y,t) \in \Re_+ \times [0,T], \tag{6.3}
$$

which ensures that the system (5.12) is uniformly nonsingular on $[0,T]$. We have the following proposition.

PROPOSITION 6.1 (local existence). *For any $\tau \ge 0$, consider the problem*

$$
\begin{cases}
\phi_t - (\bar{s}+\gamma')\phi_y - \psi_y = (V_T(y)-v_*)\gamma', \\
\psi_t - (\bar{s}+\gamma')\psi_y + p(V+\phi_y) - p(V_T) - p(V_s) + p(v_*) \\
\quad = \mu\left(\dfrac{U_y+\psi_{yy}}{V+\phi_y} - \dfrac{U_{Ty}}{V_T} - \dfrac{U_{sy}}{V_s}\right) + (U_T(y)-u_*)\gamma', \\
(\phi,\psi)|_{t=\tau} = (\phi_\tau,\psi_\tau)(y) \in H^2, \\
\phi(0,t) = C(\gamma,t) + \phi_0(0) - \sigma(\gamma,t), \quad t \ge \tau, \\
\psi(0,t) = \psi_0(0) + \displaystyle\int_0^t D(\gamma,t)\psi_y(0,t)dt + \int_0^t E(\gamma,t)dt, \quad t \ge \tau, \\
\gamma'(t) = \dfrac{1}{\bar{v}-v_b}(\psi_y(0,t) + U_s(B(\gamma,t)) - u_*), \quad \gamma(\tau) = \gamma_\tau.
\end{cases} \tag{6.4}
$$

Then there exist positive constants $\delta_0 > 0$ and $C_1 > 0$ independent of τ such that, for any $0 < \delta \le \delta_0$, $\varepsilon \in (0, \frac{\varepsilon_2}{C_0}]$, $\varepsilon_2 = O(\delta) \ll \varepsilon_1$, and β satisfying $e^{-c_-\beta} < \varepsilon_2$, there exists a positive constant T_0 depending on ε_2 but not on τ such that, if $\|(\phi_\tau,\psi_\tau)\|_2 \le \varepsilon$,

$|\gamma(\tau)| \le |\bar{s}\tau|$, *then problem* (6.4) *has a unique solution* $(\phi,\psi) \in X([\tau,\tau+T_0])$, $\gamma(t) \in C^1[\tau,\tau+T_0]$ *satisfying* $\|(\phi,\psi)(t)\|_2 \le C_1\varepsilon$ *and* $|\gamma(t)| \le |\bar{s}|t$ *for* $t \in [\tau,\tau+T_0]$.

Proof. Without loss of generality, let $\tau = 0$. By the characteristic method, ϕ has the explicit form

$$\begin{aligned}\phi(y,t) = \phi_0(\bar{x}_0) &+ \int_0^t \psi_y(\bar{x}_0 - x(\tau),\tau)d\tau \\ &+ \int_0^t \gamma'(\tau)(V_T(\bar{x}_0 - x(\tau)) - v_*)d\tau, \quad \text{if } y \ge -x(t),\end{aligned} \tag{6.5}$$

and

$$\begin{aligned}\phi(y,t) = C(\gamma(\bar{t}_0),\bar{t}_0) + \phi_0(0) - \sigma(\gamma(\bar{t}_0),\bar{t}_0) + \int_{\bar{t}_0}^t \psi_y(x(\bar{t}_0) - x(\tau),\tau)d\tau \\ + \int_{\bar{t}_0}^t \gamma'(\tau)(V_T(x(\bar{t}_0) - x(\tau)) - v_*)d\tau \quad \text{if } 0 \le y \le -x(t),\end{aligned} \tag{6.6}$$

where $\bar{x}_0 = y + x(t)$ and $\bar{t}_0 = x^{-1}(y + x(t))$. We note that the inverse function of $x(t)$ exists because $|\gamma'(t)| \le C(\|\psi(t)\|_2 + e^{-c-\beta}) \le 2C\varepsilon_2$ is small.

On the other hand, $(6.4)_2$ is regarded as the initial-boundary value problem for the parabolic equation of ψ:

$$\begin{cases} \psi_t - \dfrac{\mu}{V+\phi_y}\psi_{yy} = g := g(\phi_y,\gamma,\gamma',\psi_y), \\ \psi(0,t) = \psi_0(0) + \displaystyle\int_0^t D(\gamma,t)\psi_y(0,t)dt + \int_0^t E(\gamma,t)dt, \\ \psi|_{t=0} = \psi_0, \end{cases} \tag{6.7}$$

where

$$\begin{aligned}g(\phi_y,\gamma,\gamma',\psi_y) = (\bar{s}+\gamma')\psi_y &- (p(V+\phi_y) - p(V_T) - p(V_s) + p(v_*)) \\ &+ \mu\left(\frac{U_y}{V+\phi_y} - \frac{U_{Ty}}{V_T} - \frac{U_{sy}}{V_s}\right) + (U_T(y) - u_*)\gamma'\end{aligned}$$

and

$$\gamma(t) = \int_0^t \frac{1}{\bar{v} - v_b}(\psi_y(0,\tau) + U_s(B(\gamma,t)) - u_*)d\tau. \tag{6.8}$$

We now approximate $(\phi_0,\psi_0) \in H^2$ by $(\phi_{0k},\psi_{0k}) \in H^3$ such that

$$(\phi_{0k},\psi_{0k}) \to (\phi_0,\psi_0), \quad \text{strongly in } H^2, \tag{6.9}$$

as $k \to \infty$ and $\|(\phi_{0k},\psi_{0k})\|_2 \le \frac{3}{2}\varepsilon$ holds for any k.

We will use the iteration method to prove Proposition 6.1. We define the sequence $\{(\phi_k^{(n)}(y,t),\psi_k^{(n)}(y,t),\gamma_k^{(n)}(t))\}$ for each k so that

$$(\phi_k^{(0)},\psi_k^{(0)})(y,t) = (\phi_{0k},\psi_{0k})(y), \tag{6.10}$$

and $\gamma_k^{(0)}(t)$ is the solution of the following ODE:

$$\gamma'(t) = \frac{1}{\bar{v} - v_b}(\psi_{0ky}(0) + U_s(B(\gamma,t)) - u_*), \qquad \gamma(0) = 0, \tag{6.11}$$

and for a given $((\phi_k^{(n-1)}, \psi_k^{(n-1)})(y,t), \gamma_k^{(n-1)}(t))$, $\psi_k^{(n)}$ is a solution to

$$
\begin{cases}
\psi_{kt}^{(n)} - \dfrac{\mu \psi_{kyy}^{(n)}}{V_k^{(n-1)} + \phi_{ky}^{(n-1)}} = g^{(n-1)} = g(\phi_{ky}^{(n-1)}, \gamma_k^{(n-1)}, \gamma_{kt}^{(n-1)}, \psi_{ky}^{(n-1)}), \\
\psi_k^{(n)}(0,t) = \psi_{0k}(0) + \displaystyle\int_0^t D(\gamma_k^{(n-1)}, t)\psi_{ky}^{(n)}(0,t)dt + \int_0^t E(\gamma_k^{(n-1)}, t)dt \\
\qquad\qquad - \displaystyle\int_0^t [h(\gamma_k^{(n-1)}, \psi_{ky}^{(n)}) - h(\gamma_k^{(n-1)}, \psi_{ky}^{(n-1)})]|_{y=0} dt, \\
\psi_k^{(n)}|_{t=0} = \psi_{0k},
\end{cases} \tag{6.12}
$$

$\gamma_k^{(n)}(t)$ is a solution to

$$
\gamma_{kt}^{(n)}(t) = \frac{1}{\bar{v} - v_b}(\psi_{ky}^{(n)}(0,t) + U_s(B(\gamma_k^{(n)}, t)) - u_*), \qquad \gamma_k^{(n)}(0) = 0, \tag{6.13}
$$

and $\phi_k^{(n)}(y,t)$ is a solution to

$$
\begin{cases}
\phi_{kt}^{(n)} - (\bar{s} + \gamma_{kt}^{(n)})\phi_{ky}^{(n)} - \psi_{ky}^{(n)} = (V_T(y) - v_*)\gamma_{kt}^{(n)}, \\
\phi_k^{(n)}(0,t) = C(\gamma_k^{(n)}, t) + \phi_0(0) - \sigma(\gamma_k^{(n)}, t), \\
\phi_k^{(n)}(y,0) = \phi_{0k};
\end{cases} \tag{6.14}
$$

i.e.,

$$
\phi_k^{(n)}(y,t) = \begin{cases}
C(\gamma_k^{(n)}(\bar{t}_k^{(n)}), \bar{t}_k^{(n)}) + \phi_{0k}(0) - \sigma(\gamma_k^{(n)}(\bar{t}_k^{(n)}), \bar{t}_k^{(n)}) \\
\quad + \displaystyle\int_{\bar{t}_k^{(n)}}^t \psi_{ky}^{(n)}(x_k^{(n)}(\bar{t}_k^{(n)}) - x_k^{(n)}(\tau), \tau)d\tau \\
\quad + \displaystyle\int_{\bar{t}_k^{(n)}}^t [\gamma_k^{(n)}(\tau)]'[V_T(x_k^{(n)}(\bar{t}_k^{(n)}) - x_k^{(n)}(\tau)) - v_*]d\tau \\
\text{if } 0 \le y \le -x_k^{(n)}(t), \\
\phi_{0k}(\bar{x}_k^{(n)}) + \displaystyle\int_0^t \psi_{ky}^{(n)}(\bar{x}_k^{(n)} - x_k^{(n)}(\tau), \tau)d\tau \\
\quad + \displaystyle\int_0^t [\gamma_k^{(n)}(\tau)]'[V(\bar{x}_k^{(n)} - x_k^{(n)}(\tau)) - v_*]d\tau \\
\text{if } y \ge -x_k^{(n)}(t),
\end{cases} \tag{6.15}
$$

where

$$
\begin{gathered}
x_k^{(n)}(t) = \bar{s}t + \gamma_k^{(n)}(t), \\
\bar{t}_k^{(n)} = (x_k^{(n)})^{-1}(y + x_k^{(n)}(t)), \qquad \bar{x}_k^{(n)} = y + x_k^{(n)}(t), \\
h(\gamma, \psi_y) = \bar{s}\psi_y + \frac{\psi_y + U_s(B(\gamma,t)) - u_*}{\bar{v} - v_b}(\psi_y + u_b - u_*),
\end{gathered} \tag{6.16}
$$

and $V_k^{(0)} = v_b - \phi_{ky}^{(0)}$, $V_k^{(n)} = V_T + V_s(y + B(\gamma_k^{(n)}, t)) - v_*$, $n \ge 1$. Since (6.14) contains only the index n, by the same argument of (5.11), we have $\phi_{ky}^{(n)}(0,t) =$

$v_* - V_s(B(\gamma_k^{(n)}, t))$, which infers

$$(V_k^{(n)} + \phi_{ky}^{(n)})(0,t) = v_b, \quad n \geq 0. \tag{6.17}$$

Since ε_2 is small, by the principle of contraction mapping, it is easy to prove there exist $C_1 > 2$ and positive time $t_0(\varepsilon_2) \ll 1$ such that, if $g^{(n-1)} \in C(0, t_0; H^2)$ and $\psi_{0k} \in H^3$, there exists a unique-local solution $\psi_k^{(n)}$ to (6.12) satisfying

$$\psi_k^{(n)} \in C(0, t_0; H^3) \cap C^1(0, t_0; H^1) \cap L^2(0, t_0; H^4)$$

and $\sup_{y \in R_+ \times (0,t_0)} |\psi_k^{(n)}(y,t)| \leq C_1\varepsilon$.

Thus, if $(\|\phi_k^{(n-1)}\|_2, \|\psi_k^{(n-1)}\|_2) \leq C_1\varepsilon$, multiplying (6.12) by $\psi_k^{(n)}$ and integrating it over R_+, we have

$$\begin{aligned} &\|\psi_k^{(n)}\|_t^2 + \frac{\mu}{v_+}\|\psi_{ky}^{(n)}\|^2 \\ &\qquad \leq C(\varepsilon)(1 + \|\psi_k^{(n)}\|^2) + \frac{1}{4}\|\psi_{kyy}^{(n)}\|^2. \end{aligned} \tag{6.18}$$

Multiplying (6.12) by $-\psi_{kyy}^{(n)}$ and integrating it over $\mathbb{R}_+$, one has

$$\|\psi_{ky}^{(n)}\|_t^2 + \frac{\mu}{v_+}\|\psi_{kyy}^{(n)}(t)\|^2 \leq C(\varepsilon)(1 + \|\psi_{ky}^{(n)}\|^2). \tag{6.19}$$

Combining (6.18) and (6.19) gives, if T_0 is chosen suitably small,

$$\|\psi_k^{(n)}\|_1 \leq C_1\varepsilon \quad \text{if } t < T_0. \tag{6.20}$$

Differentiating (6.12) with respect to y, multiplying by $-\psi_{kyyy}^{(n)}$, and integrating over R_+, we have

$$\|\psi_{kyy}^{(n)}\|_t^2 + \frac{\mu}{v_+}\|\psi_{kyyy}^{(n)}(t)\|^2 + 2\psi_{kyt}^{(n)}\psi_{kyy}^{(n)}\Big|_{y=0} \leq C(\varepsilon). \tag{6.21}$$

Substituting (6.16), (6.17) into (6.12) yields, on the boundary $y = 0$,

$$\psi_{kyy}^{(n)} = e_1[\psi_{ky}^{(n)}]^2 + e_2\psi_{ky}^{(n)} + e_3, \tag{6.22}$$

where the coefficients e_i, $i = 1, 2, 3$, depend only on $\gamma_k^{(n-1)}$ and t. Thus the integration of (6.21) over $(0,t)$ and using (6.22) give

$$\|\psi_{kyy}^{(n)}\|^2 \leq C_1\varepsilon. \tag{6.23}$$

On the other hand, a direct estimate on (6.15) yields

$$\|\phi_k^{(n)}(t)\|_2 \leq C_1\varepsilon. \tag{6.24}$$

Therefore, $(\|\phi_k^{(n)}\|_2, \|\psi_k^{(n)}\|_2) \leq C_1\varepsilon$. From (6.13), it is easy to see that $\gamma_k^{(n)}(t) \in C^1([0, T_0])$ and $|\gamma_k^{(n)}(t)| \leq C_1\varepsilon(t - \tau) + \gamma_\tau \leq |\bar{s}|t$. By the classical method of the contraction mapping principle, $(\phi_k^{(n)}, \psi_k^{(n)})$ is a Cauchy sequence in $C(0, t_0; H^3)$. Thus we have a solution $(\phi_k(y,t), \psi_k(y,t))$, $\gamma_k(t)$ by letting n tend to infinity. In the same way, letting $k \to \infty$, we obtain the desired unique-local solution $(\phi(y,t), \psi(y,t), \gamma(t))$ to (6.4). We omit the details. ∎

7. A priori estimates. This section is devoted to the a priori estimates. Throughout this section, we use c, C to denote the positive constants which are independent of T, β, and α. We first rewrite the system (5.12) as

$$
\begin{cases}
\phi_t - (\bar{s}+\gamma')\phi_y - \psi_y = (V_T(y)-v_*)\gamma', \\
\psi_t - (\bar{s}+\gamma')\psi_y - f(V)\phi_y - \dfrac{\mu}{V}\psi_{yy} = F + G + (U_T(y)-u_*)\gamma', \\
(\phi,\psi)(y,0) = (\phi_0,\psi_0)(y), \\
\phi(0,t) = C(\gamma,t) + \phi_0(0) - \sigma(t), \\
\psi(0,t) = \psi_0(0) + \displaystyle\int_0^t D(\gamma,t)\psi_y(0,t)dt + \int_0^t E(\gamma,t)dt, \\
\gamma'(t) = \dfrac{1}{\bar{v}-v_b}(\psi_y(0,t) + U_s(0,t) - u_*), \quad \gamma(0) = 0,
\end{cases} \tag{7.1}
$$

where

$$
f(V) = f(V_s,V_T) = -p'(V) + \frac{h(V_s)}{V} + \frac{g(V_T)}{V}, \tag{7.2}
$$

$$
h(V_s) = \frac{\mu s V_s'}{V_s}, \qquad g(V_T) = \frac{\mu\bar{s}V_T'}{V_T}, \tag{7.3}
$$

$$
\begin{aligned}
F = &-\{p(V+\phi_y) - p(V) - p'(V)\phi_y\} \\
&+ (h(V_s)\phi_y + g(V_T)\phi_y)\left(\frac{1}{V+\phi_y} - \frac{1}{V}\right) + \mu\psi_{yy}\left(\frac{1}{V+\phi_y} - \frac{1}{V}\right),
\end{aligned} \tag{7.4}
$$

$$
G = -\{p(V) - p(V_s) - p(V_T) + p(v_*)\} + h(V_s)\frac{V_T - v_*}{V+\phi_y} + g(V_T)\frac{V_s - v_*}{V+\phi_y}. \tag{7.5}
$$

It is easy to see that

$$
|F| = O(|\phi_y|^2 + |\phi_y|\cdot|\psi_{yy}|). \tag{7.6}
$$

Proposition 7.1 (a priori estimates). *There exist positive constants δ_0 and $C_2 > 0$ such that if $0 < v_* - v_b = \delta < \delta_0$ and $(\phi,\psi) \in X([0,T])$, $\gamma(t) \in C^1[0,T]$ is a solution of (7.1) for some positive T satisfying $|\gamma(t)| \le |\bar{s}|t$ and $(N(T), e^{-\frac{1}{5}c_-\beta}) = O(\delta)$, then (ϕ,ψ) satisfies the a priori estimates*

$$
\|(\phi,\psi)(t)\|_2^2 + \int_0^t \{\|\phi_y\|_1^2 + \|\psi_y\|_2^2\}d\tau \le C_2(\delta^{-2}\|(\phi_0,\psi_0)\|_2^2 + \delta^2), \tag{7.7}
$$

$$
\int_0^t \left|\frac{d}{dt}\|\phi_y\|^2\right| + \left|\frac{d}{dt}\|\psi_y\|^2\right| d\tau \le C_2(\delta^{-2}\|(\phi_0,\psi_0)\|_2^2 + \delta^2). \tag{7.8}
$$

Before proving Proposition 7.1, we first give some lemmas.

Lemma 7.2. *For $0 \le t \le T$, the following holds:*

$$
\begin{gathered}
\psi(0,t) = -\bar{s}\phi(0,t) + \frac{A(t) + s\sigma(t)}{v_* - \bar{v}}, \\
\phi_y(0,t) = v_* - V_s(0,t), \\
\psi_{yy}(0,t) = A_1\psi_y(0,t)^2 + A_2(t)\psi_y(0,t) + A_3(t),
\end{gathered} \tag{7.9}
$$

$$
(|A(t)|, |A'(t)|, |\sigma(t)|, |\sigma'(t)|) \le Ce^{-c_-(st+\beta)}, \quad \bar{s} < 0, \quad |\bar{s}| = O(\delta^{\frac{1}{2}}),
$$

$$
\begin{aligned}
&\phi(0,t)\psi(0,t) \geq -\frac{3}{4}\bar{s}\phi(0,t) - Ce^{-c_-(st+\beta)},\\
&\psi(0,t)^2 \leq C\delta\phi^2(0,t) + Ce^{-c_-(st+\beta)},\\
&\delta^{\frac{1}{2}}\gamma'^2 \leq \varepsilon_3\|\psi_y\|^2 + C(\varepsilon_3)\delta\|\psi_{yy}\|^2 + Ce^{-c_-(st+\beta)},\\
&\phi_y(0,t)^2 \leq Ce^{-c_-(st+\beta)}, \quad |\psi(0,t)\phi_y(0,t)| \leq C(\psi(0,t)^2 + \phi_y(0,t)^2),\\
&|\psi_y(0,t)\psi(0,t)| \leq \varepsilon_3(\delta^{\frac{1}{2}}\phi(0,t)^2 + \|\psi_y\|^2)\\
&\qquad\qquad + C(\varepsilon_3)(\delta\|\psi_{yy}\|^2 + \delta^{-\frac{1}{2}}e^{-c_-(st+\beta)}),\\
&\psi_y(0,t)\psi_t(0,t) \geq \frac{u_*}{2(v_b-\bar{v})}\psi_y(0,t)^2 - Ce^{-c_-(st+\beta)} - C\delta^2\gamma'^2,
\end{aligned}
\tag{7.10}
$$

where ε_3 is an arbitrary positive constant and

$$
\begin{aligned}
&A_1 = \frac{v_b}{\mu(v_b-\bar{v})}, \qquad A_2(t) = \frac{2(u_b + U_s(0,t) - u_*)v_b}{\mu(v_b-\bar{v})},\\
&A_3(t) = \frac{v_b(u_b + U_s(0,t) - u_*)^2 - v_b u_b^2}{\mu(v_b-\bar{v})} - U_{sy}(0,t).
\end{aligned}
$$

Proof. In view of the boundary conditions of (7.1), we have (3.28) because α is determined by (3.27). By (3.28), we get

$$
u_*\phi(0,t) - (v_* - \bar{v})\psi(0,t) = -A(t) - s\sigma(t), \tag{7.11}
$$

which is equivalent to the first term of (7.9). It is noted that our boundary conditions of the reformulated system (7.1) are derived from the original boundary conditions of (3.3). To the contrary, it is easy to verify that (3.3) holds from (7.1). By the boundary conditions of (3.3), the other terms of (7.9) are easily proved.

On the other hand, the first term of (7.10) is easy from Lemmas 2.1–2.2, (3.20), and (3.28) because α is small due to Proposition 5.1. For the fourth term of (7.10), we compute by (7.1) and Lemma 2.2

$$
\begin{aligned}
\delta^{\frac{1}{2}}\gamma'^2 &\leq \delta^{\frac{1}{2}}(\psi_y^2(0,t) + Ce^{-c_-(st+\beta)})\\
&\leq \varepsilon_3\|\psi_y\|^2 + C(\varepsilon_3)\delta\|\psi_{yy}\|^2 + Ce^{-c_-(st+\beta)}.
\end{aligned}
$$

In a similar way, we can also get the other terms of (7.10). We omit the proofs here. ∎

Let $x' = \bar{s} + \gamma'$; then $0 > x' \geq -\frac{1}{2}\bar{s}$ and $|x'| = O(\delta^{\frac{1}{2}})$. We now establish the a priori estimates.

LEMMA 7.3. *There exists a positive constant δ_0 which depends only on $\bar{v}$, v_b, and v_+. For any given $0 < v_* - v_b = \delta < \delta_0$, the following holds:*

$$
\begin{aligned}
&\|(\phi,\psi)(t)\|^2 + \int_0^t \|\psi_y\|^2 d\tau + \int_0^t\int_0^\infty V_{sy}\psi^2 dydt + \int_0^t \delta^{\frac{1}{2}}\phi^2(o,t)dt\\
&\leq C\left\{\|(\phi_0,\psi_0)\|^2 + \delta^4 + \delta^2\int_0^t \|\phi_y\|^2 d\tau\right.\\
&\qquad \left. + \delta\int_0^t \|\psi_{yy}\|^2 dt + N(T)\int_0^t [\|\phi_y\|^2 + \|\psi_{yy}\|^2]d\tau\right\}.
\end{aligned}
\tag{7.12}
$$

Proof. Multiplying $(7.1)_1$ by ϕ and $(7.1)_2$ by $f(V)^{-1}\psi$, then we have

$$\begin{aligned}&\left(\frac{1}{2}\phi^2\right)_t-\left(\frac{x'}{2}\phi^2\right)_y-(\phi\psi)_y+\left(\frac{1}{2f(V)}\psi^2\right)_t-\left(\frac{1}{2f(V)}\right)_t\psi^2-\frac{x'}{2}\left(\frac{\psi^2}{f(V)}\right)_y\\&+\frac{x'}{2}\psi^2\left(\frac{1}{f(V)}\right)_y-\left(\frac{\mu}{Vf(V)}\psi_y\psi\right)_y+\frac{\mu}{Vf(V)}\psi_y^2+\left(\frac{\mu}{Vf(V)}\right)_y\psi_y\psi\\&=(F+G)\psi f(V)^{-1}+\gamma'(V_T(y)-v_*)\phi+\gamma'(U_T(y)-u_*)\psi f(V)^{-1}.\end{aligned}\tag{7.13}$$

By the definition of $f(V)$ and Lemma 2.1, one has

$$\begin{aligned}&\left|\left(\frac{\mu}{Vf(V)}\right)_y\psi_y\psi\right|\\&\le\left|\frac{\mu|K'(V_s)|+C\delta}{(Vf(V))^2}V_{sy}\psi_y\psi\right|+C(|V_T'|+|\bar{s}V_T''|)|\psi_y\psi|\\&\le\bar{\varepsilon}\frac{\mu}{Vf(V)}\psi_y^2+\frac{\mu K'(V_s)^2+C\delta}{4\bar{\varepsilon}K(V_s)^3}V_{sy}^2\psi^2+C(|V_T'|^2+\bar{s}^2|V_T''|^2)\psi^2,\end{aligned}\tag{7.14}$$

for any $\bar{\varepsilon}>0$ which will be determined later, where

$$K(V_s)=-p'(V_s)V_s+h(V_s).$$

Substituting this inequality into (7.13) yields

$$\begin{aligned}&\left\{\frac{1}{2}\phi^2+\frac{1}{2f(V)}\psi^2\right\}_t+(Z(V_s)-C\delta)V_{sy}\psi^2+(1-\bar{\varepsilon})\frac{\mu}{Vf(V)}\psi_y^2\\&\quad-\left\{\frac{x'}{2}\phi^2+\phi\psi+\frac{\mu}{Vf(V)}\psi_y\psi+\frac{x'}{2f(V)}\psi^2\right\}_y\\&\le(F+G)\psi f(V)^{-1}+\gamma'(V_T(y)-v_*)\phi\\&\quad+\gamma'(U_T(y)-u_*)\psi f(V)^{-1}+C(\delta^{\frac{1}{2}}V_T'+\delta|V_T''|)\psi^2,\end{aligned}\tag{7.15}$$

where

$$Z(V_s)=\frac{s}{2}\frac{K(V_s)-V_sK'(V_s)}{K(V_s)^2}-\frac{\mu K'(V_s)^2V_{sy}}{4\bar{\varepsilon}K(V_s)^3}.\tag{7.16}$$

Let $1\le\gamma\le3$; if we choose

$$\max\left\{\frac{(\gamma-1)^2(v_+-v_*)}{2\gamma v_*},\frac{1}{2}\right\}\le\bar{\varepsilon}<1,\tag{7.17}$$

then by the argument of [3, 12] we have

$$Z(V_s)\ge C_3>0.\tag{7.18}$$

Thus integration of (7.15) over $[0,+\infty)\times[0,t]$ yields, if $\delta<\frac{1}{2}C_3$,

$$\int_0^{+\infty}\left\{\frac{\phi^2}{2}+\frac{\psi^2}{2f(V)}\right\}dy+\int_0^t\int_0^{+\infty}\frac{1}{2}Z(V_s)V_{sy}\psi^2dyd\tau$$

$$
\begin{aligned}
&+\int_0^t\int_0^{+\infty}\frac{\mu\psi_y^2}{2Vf(V)}dyd\tau+\int_0^t\left[\frac{x'}{2}\phi^2+\phi\psi\right](0,t)dt\\
&\le C\int_0^{+\infty}\{\phi_0^2+\psi_0^2\}dy+\left|\int_0^t\left[\frac{\mu\psi_y\psi}{Vf(V)}+\frac{x'}{2f(V)}\psi^2\right](0,t)d\tau\right|\\
&+C\int_0^t\int_0^{+\infty}(|F|+|G|)|\psi|dyd\tau+C\delta^{\frac12}\int_0^t\int_0^{\infty}(V_T'+\delta^{\frac12}|V_T''|)\psi^2dyd\tau\\
&+\int_0^t\int_0^{+\infty}\gamma'(V_T(y)-v_*)\phi+\gamma'(U_T(y)-u_*)\psi f(V)^{-1}dyd\tau.
\end{aligned}\tag{7.19}
$$

We estimate the right-hand sides of (7.19). We compute

$$
\begin{aligned}
&\int_0^\infty\int_0^\infty|G|dydt\\
&\le\int_0^\infty\int_0^\infty|V_s(y+x(t)-st+\alpha-\beta)-v_*||V_T(y)-v_*|dydt\\
&\le C\int_0^\infty\int_0^{st+\beta-\alpha-x(t)}\delta e^{-c_0\frac{y}{\sqrt\delta}}e^{-c_-|y+x(t)-st+\alpha-\beta|}dydt\\
&\quad+C\int_0^\infty\int_{st+\beta-\alpha-x(t)}^\infty\delta e^{-c_0\frac{y}{\sqrt\delta}}dydt\le C\delta e^{-c_-\beta}\le C\delta^6,
\end{aligned}\tag{7.20}
$$

$$
\begin{aligned}
&\int_0^\infty(V_T'+\delta^{\frac12}|V_T''|)\psi^2dy\le C\int_0^\infty(V_T'+\delta^{\frac12}|V_T''|)(\psi^2(0,t)+y\|\psi_y\|^2)dy\\
&\le C\delta\psi^2(0,t)+C\delta^{\frac32}\|\psi_y\|^2,
\end{aligned}\tag{7.21}
$$

and

$$
\begin{aligned}
&\int_0^\infty|\gamma'(V_T(y)-v_*)\phi|dy\\
&\le C\int_0^\infty|\gamma'(V_T(y)-v_*)|(|\phi(0,t)|+y^{\frac12}\|\phi_y\|)dy\\
&\le C\delta^{\frac32}(\gamma'^2+\phi^2(0,t))+C\delta^2\|\phi_y\|^2.
\end{aligned}\tag{7.22}
$$

Similar to (7.22), we also have

$$
\int_0^\infty|\gamma'(U_T(y)-u_*)\psi f(V)^{-1}|dy\le C\delta^2(\|\psi_y\|^2+\gamma'^2+\psi^2(0,t)).\tag{7.23}
$$

We now estimate the terms from the boundary. We calculate from (7.10)

$$
\frac{x'}{2}\phi^2+\phi\psi\ge-\frac14\bar s\phi^2(0,t)-Ce^{-c_-(st+\beta)}\ge c_1\delta^{\frac12}\phi^2(0,t)-Ce^{-c_-(st+\beta)},\tag{7.24}
$$

$$
\begin{aligned}
&\left|\frac{\mu}{Vf(V)}\psi_y(0,t)\psi(0,t)\right|\\
&\le\varepsilon_3(\delta^{\frac12}\phi(0,t)^2+\|\psi_y\|^2)+C(\varepsilon_3)(\delta\|\psi_{yy}\|^2+\delta^{-\frac12}e^{-c_-(st+\beta)}).
\end{aligned}\tag{7.25}
$$

Since $e^{-c_-\beta}=O(\delta^5)$, we have

$$
\delta^{-\frac12}e^{-c_-\beta}\le C\delta^4.\tag{7.26}
$$

Thus there exists a positive constant $\delta_0 > 0$. For any $\delta < \delta_0$, if we choose $\varepsilon_3 \le \frac{1}{2}c_1$, then the estimate (7.12) holds from (7.6), (7.10), and (7.19)–(7.26). Hence Lemma 7.3 is proved. □

LEMMA 7.4. *It follows that*

$$\begin{aligned}\|\phi_y\|^2 + \int_0^t \|\phi_y\|^2 d\tau \le C\delta \int_0^t \|\psi_{yy}\|^2 dt \\ + C\left\{\|(\phi_0,\psi_0)\|_1^2 + \delta^4 + N(T)\int_0^t [\|\phi_y\|^2 + \|\psi_y\|_1^2]d\tau\right\}.\end{aligned} \tag{7.27}$$

Proof. From the system (7.1), we have

$$\begin{aligned}\frac{\mu}{V}\phi_{yt} - x'\frac{\mu}{V}\phi_{yy} + f(V)\phi_y + x'\psi_y \\ = \psi_t - F - G + \frac{\mu}{V}\gamma' V_T' - \gamma'(U_T(y) - u_*).\end{aligned} \tag{7.28}$$

Multiplying (7.28) by ϕ_y yields

$$\begin{aligned}\left(\frac{\mu}{2V}\phi_y^2\right)_t - \frac{(s-x')\mu}{2V^2}V_{sy}\phi_y^2 - \left\{x'\frac{\mu}{2V}\phi_y^2\right\}_y - \frac{\mu x'}{2V^2}V_y\phi_y^2 + f(V)\phi_y^2 + x'\psi_y\phi_y \\ = \psi_t\phi_y - (F+G)\phi_y + \frac{\mu}{V}\gamma' V_T'\phi_y - \gamma'(U_T(y) - u_*)\phi_y.\end{aligned} \tag{7.29}$$

The system $(7.1)_1$ gives

$$\begin{aligned}\psi_t\phi_y = (\psi\phi_y)_t - \psi\phi_{yt} = (\psi\phi_y)_t - \psi(x'\phi_{yy} + \psi_{yy} + \gamma' V_T') \\ = (\psi\phi_y)_t - (x'\psi\phi_y)_y - (\psi\psi_y)_y - \gamma'\psi V_T' + x'\psi_y\phi_y + \psi_y^2,\end{aligned} \tag{7.30}$$

and Lemma 2.1 and the Cauchy inequality yield

$$|\gamma'(U_T(y) - u_*)\phi_y| \le C\delta\phi_y^2 + C\delta^2\gamma'^2 e^{-\frac{c_0}{\sqrt{\delta}}y}. \tag{7.31}$$

Substituting (7.30)–(7.31) into (7.29), we get

$$\begin{aligned}\left(\frac{\mu}{2V}\phi_y^2 - \psi\phi_y\right)_t + \left(f(V) - \frac{h(V_s)}{2V} - C\delta^{\frac{1}{2}}\right)\phi_y^2 + \left\{\psi\psi_y + x'\psi\phi_y - \frac{\mu x'}{2V}\phi_y^2\right\}_y \\ \le \psi_y^2 - (F+G)\phi_y + \frac{\mu}{V}\gamma' V_T'\phi_y - \gamma' V_T'\psi + C\delta^2\gamma'^2 e^{-\frac{c_0}{\sqrt{\delta}}y}.\end{aligned} \tag{7.32}$$

Note that (7.2) gives

$$f(V) - \frac{h(V_s)}{2V} > -p'(V) \ge -p'(v_+), \tag{7.33}$$

and the boundary terms of (7.32) are investigated in Lemma 7.2. Thus integrating (7.32) over $[0,+\infty)\times[0,t]$ and using (7.6), (7.20), Lemmas 7.2–7.3, and the inequalities

$$\int_0^{+\infty} |\psi\phi_y| dy \le \frac{\mu}{4v_+}\|\phi_y(t)\|^2 + \frac{v_+}{\mu}\|\psi\|^2, \tag{7.34}$$

$$\int_0^{\infty} \left|\frac{\mu}{V}\gamma' V_T'\phi_y\right| dy \le C\delta^{\frac{1}{2}}\|\phi_y\|^2 + C\delta\gamma'^2, \tag{7.35}$$

$$\int_0^{\infty} |\gamma' V_T'\psi| dy \le C\delta(\gamma'^2 + \psi(0,t)^2) + C\delta^{\frac{3}{2}}\|\psi_y\|^2, \tag{7.36}$$

we get (7.27). Thus Lemma 7.4 is proved. □

LEMMA 7.5. *It follows that*

$$\begin{aligned}\|\psi_y(t)\|^2 + \int_0^t \|\psi_{yy}\|^2 d\tau + \delta^{\frac{1}{2}} \int_0^t \psi_y(0,t)^2 dt \\ \le C\left\{ \|(\phi_0,\psi_0)\|_1^2 + \delta^4 + N(T)\int_0^t \left[\|\phi_y(\tau)\|^2 + \|\psi_y(\tau)\|_1^2\right] d\tau \right\}.\end{aligned} \tag{7.37}$$

Proof. Multiplying $(7.1)_2$ by $-\psi_{yy}$, one obtains

$$\begin{aligned}(-\psi_y\psi_t)_y + \left(\frac{\psi_y^2}{2}\right)_t + \left(\frac{x'}{2}\psi_y^2\right)_y + f(V)\phi_y\psi_{yy} + \frac{\mu}{V}\psi_{yy}^2 \\ = -(F+G)\psi_{yy} - \gamma'(U_T(y)-u_*)\psi_{yy}.\end{aligned} \tag{7.38}$$

The Cauchy inequality yields

$$|f(V)\phi_y\psi_{yy}| \le \frac{\mu}{4v_+}\psi_{yy}^2 + C\phi_y^2, \tag{7.39}$$

and (7.6) and the Cauchy inequality yield

$$|F\psi_{yy}| \le C(|\phi_y|^2 + |\phi_y|\cdot|\psi_{yy}|)|\psi_{yy}| \le C|\phi_y|(|\phi_y|^2 + |\psi_{yy}|^2), \tag{7.40}$$

$$|G\psi_{yy}| \le \frac{\mu}{4v_+}\psi_{yy}^2 + C|G|, \tag{7.41}$$

$$|\gamma'(U_T(y)-u_*)\psi_{yy}| \le \frac{\mu}{4v_+}\psi_{yy}^2 + C\delta^3{\gamma'}^3 e^{-\frac{c_0}{\sqrt{\delta}}y}. \tag{7.42}$$

Substituting (7.39)–(7.42) into (7.38), we have

$$\begin{aligned}\frac{1}{2}(\psi_y^2)_t + \left\{\frac{x'}{2}\psi_y^2 - \psi_y\psi_t\right\}_y + \frac{\mu}{4v_+}\psi_{yy}^2 \\ \le C\phi_y^2 + C|\phi_y|(|\phi_y|^2 + |\psi_{yy}|^2) + C\delta^3{\gamma'}^3 e^{-\frac{c_0}{\sqrt{\delta}}y}.\end{aligned} \tag{7.43}$$

The boundary terms of (7.43) are estimated in Lemma 7.2. Thus integrating (7.43) over $[0,+\infty)\times[0,t]$ and using Lemmas 7.2 and 7.4 and the fact that $x' < 0$, we get the estimate (7.37). □

From Lemmas 7.3–7.5, we get the following inequality.

LEMMA 7.6. *It follows that*

$$\|(\phi,\psi)(t)\|_1^2 + \int_0^t \left\{\|\phi_y\|^2 + \|\psi_y\|_1^2\right\} d\tau \le C\{\|(\phi_0,\psi_0)\|_1^2 + \delta^4\}. \tag{7.44}$$

LEMMA 7.7. *It follows that*

$$\begin{aligned}\|\phi_{yy}(t)\|^2 + \int_0^t \|\phi_{yy}\|^2 d\tau \le C\delta^{-2}\{\|(\phi_0,\psi_0)\|_2^2 + \delta^4\} \\ + C\left\{\int_0^t \|F_y\|^2 d\tau + \delta\int_0^t \|\psi_{yyy}\|^2 d\tau\right\}.\end{aligned} \tag{7.45}$$

Proof. Differentiating (7.28) with respect to y, one gets

$$\begin{aligned}&\frac{\mu}{V}\phi_{yyt}-\frac{\mu V_y}{V^2}\phi_{yt}-x'\frac{\mu}{V}\phi_{yyy}+x'\frac{\mu V_y}{V^2}\phi_{yy}+f(V)\phi_{yy}+f(V)_y\phi_y+x'\psi_{yy}\\&\quad=\psi_{yt}-F_y-G_y+\frac{\mu}{V}\gamma' V_T''-\frac{\mu V_y}{V^2}\gamma' V_T'-\gamma' U_T'.\end{aligned}\tag{7.46}$$

Multiplying (7.46) by ϕ_{yy}, we have

$$\begin{aligned}&\left(\frac{\mu}{2V}\phi_{yy}^2\right)_t+\left(f(V)-\frac{h(V_s)}{2V}-C\delta^{\frac12}\right)\phi_{yy}^2-\left(\frac{x'\mu}{2V}\phi_{yy}^2\right)_y\\&\quad-\frac{\mu V_y}{V^2}\phi_{yt}\phi_{yy}+f(V)_y\phi_{yy}\phi_y+x'\psi_{yy}\phi_{yy}\\&\quad\le\psi_{yt}\phi_{yy}-(F_y+G_y)\phi_{yy}+\left(\frac{\mu}{V}\gamma' V_T''-\frac{\mu V_y}{V^2}\gamma' V_T'-\gamma' U_T'\right)\phi_{yy}.\end{aligned}\tag{7.47}$$

Using Lemmas 2.1 and 7.2 and the Cauchy inequality, we obtain

$$\begin{aligned}\left|\frac{\mu V_y}{V^2}\phi_{yt}\phi_{yy}\right|&=\left|\frac{\mu V_y}{V^2}(x'\phi_{yy}+\psi_{yy}+\gamma' V_T')\phi_{yy}\right|\\&\le\frac18|p'(v_+)|\phi_{yy}^2+C(\psi_{yy}^2+\delta^{\frac12}\gamma'^2V_T'),\end{aligned}\tag{7.48}$$

$$|f(V)_y\phi_{yy}\phi_y|\le\frac18|p'(v_+)|\phi_{yy}^2+C\phi_y^2,\tag{7.49}$$

$$|F_y\phi_{yy}|\le\frac18|p'(v_+)|\phi_{yy}^2+C|F_y|^2,\tag{7.50}$$

$$\begin{aligned}\psi_{yt}\phi_{yy}&=(\psi_y\phi_{yy})_t-\psi_y\phi_{yyt}=(\psi_y\phi_{yy})_t-(\psi_y\phi_{yt})_y+\phi_{yt}\psi_{yy}\\&=(\psi_y\phi_{yy})_t-(\psi_y\phi_{yt})_y+x'\phi_{yy}\psi_{yy}+\psi_{yy}^2+\gamma' V_T'\psi_{yy},\end{aligned}\tag{7.51}$$

$$\left|\left(\frac{\mu}{V}\gamma' V_T''-\frac{\mu V_y}{V^2}\gamma' V_T'-\gamma' U_T'\right)\phi_{yy}\right|\le\frac18|p'(v_+)|\phi_{yy}^2+C\gamma'^2e^{-\frac{c_0}{\sqrt\delta}y}.\tag{7.52}$$

Substituting (7.48)–(7.52) into (7.47), we have

$$\begin{aligned}&\left(\frac{\mu}{2V}\phi_{yy}^2-\psi_y\phi_{yy}\right)_t+\frac14|p'(v_+)|\phi_{yy}^2+\left(\psi_y\phi_{yt}-\frac{x'\mu}{2V}\phi_{yy}^2\right)_y\\&\quad\le2\psi_{yy}^2+C\gamma'^2e^{-\frac{c_0}{\sqrt\delta}y}+C|G_y|^2.\end{aligned}\tag{7.53}$$

Similar to (7.20), we have

$$|G_y|\le H+C\delta\phi_{yy},\tag{7.54}$$

where H satisfies

$$\int_0^t\int_0^\infty|H|dydt\le C\delta^{\frac12}e^{-c_-\beta}\le C\delta^5.\tag{7.55}$$

Since

$$\phi_{yy}|_{y=0}=\frac{1}{x'}(\phi_{yt}-\psi_{yy}-\gamma' V_T')|_{y=0},$$

Lemma 7.2 and the Cauchy inequality yield

$$\left|\frac{x'\mu}{2V}\phi_{yy}(0,t)^2\right| \le C(\delta^{\frac{1}{2}}\gamma'^2 + \delta^4 e^{-c_- st} + \delta^{-2}\|\psi_{yy}\|^2 + \delta\|\psi_{yyy}\|^2), \tag{7.56}$$

where we have used the fact that $\phi_{yt} = -V_s'(0,t)$. Note that $|\psi_y\phi_{yt}|(0,t)$ is controlled by $\delta\psi_y(0,t)^2$ and $\delta^{-1}\phi_{yt}(0,t)^2$. Thus integrating (7.53) over $[0,+\infty)\times[0,t]$ and making use of Lemmas 7.2 and 7.6 and (7.54)–(7.56), we obtain the inequality (7.45). Lemma 7.7 is proved. ∎

LEMMA 7.8. *It follows that*

$$\begin{aligned}&\|\psi_{yy}(t)\|^2 + \int_0^t \|\psi_{yyy}\|^2 d\tau \\ &\quad\le C\delta^{-2}\{\|(\phi_0,\psi_0)\|_2^2 + \delta^4\} + C\int_0^t \|F_y(\tau)\|^2 d\tau.\end{aligned} \tag{7.57}$$

Proof. Differentiating $(7.1)_2$ with respect to y and multiplying the derivative by $-\psi_{yyy}$, we have

$$\begin{aligned}&\left(\frac{1}{2}\psi_{yy}^2\right)_t + \left(\frac{x'}{2}\psi_{yy}^2 - \psi_{yt}\psi_{yy}\right)_y + f(V)\phi_{yy}\psi_{yyy} + f(V)_y\phi_y\psi_{yyy} \\ &\quad + \frac{\mu}{V}\psi_{yyy}^2 - \frac{\mu}{V^2}V_y\psi_{yy}\psi_{yyy} = -(F_y + G_y + \gamma' U_T')\psi_{yyy}.\end{aligned} \tag{7.58}$$

The Cauchy inequality and Lemma 2.1 yield

$$|f(V)\phi_{yy}\psi_{yyy}| \le C|\phi_{yy}|^2 + \frac{\mu}{8v_+}|\psi_{yyy}|^2, \tag{7.59}$$

$$|f(V)_y\phi_y\psi_{yyy}| \le C|\phi_y|^2 + \frac{\mu}{8v_+}|\psi_{yyy}|^2, \tag{7.60}$$

$$(|F_y| + |G_y|)|\psi_{yyy}| \le C\left(|F_y|^2 + G_y^2 + \frac{\mu}{8v_+}|\psi_{yyy}|^2\right), \tag{7.61}$$

$$|\gamma' U_T'\psi_{yyy}| \le C\delta^3\gamma'^2 e^{-\frac{c_0}{\sqrt{\delta}}y} + \frac{\mu}{8v_+}|\psi_{yyy}|^2. \tag{7.62}$$

On the other hand, Lemma 7.2 yields, on the boundary $y = 0$,

$$\begin{aligned}\psi_{yt}\psi_{yy} &= \psi_{yt}(A_1\psi_y^2 + A_2(t)\psi_y + A_3(t)) \\ &= \left(\frac{A_1}{3}\psi_y^3 + \frac{A_2(t)}{2}\psi_y^2 + A_3(t)\psi_y\right)_t - \frac{A_2'(t)}{2}\psi_y^2 - A_3'(t)\psi_y.\end{aligned} \tag{7.63}$$

Here $A_2(t) > 0$ because β is large and $(|A_2'(t)|, |A_3'(t)|) \le Ce^{-c_-(st+\beta)}$.

Substituting (7.54), (7.55), and (7.59)–(7.63) into (7.58), integrating it over $[0,+\infty)\times[0,t]$, and making use of Lemmas 7.6 and 7.7 and the fact that $x' < 0$, we obtain the inequality (7.57). Lemma 7.8 is proved. ∎

Proof of Proposition 7.1. By using the Sobolev embedding theorem and (7.4), we have

$$\begin{aligned}\|F_y\|^2 &\le C\int_0^{+\infty}(\phi_y^4 + \phi_y^2\phi_{yy}^2 + \psi_{yy}^2\phi_{yy}^2 + \psi_{yyy}^2\phi_y^2 + \phi_y^2\psi_{yy}^2)dy \\ &\le C\sup_{y\in\Re_+}\left[\phi_y^2\int_0^{+\infty}(\phi_y^2 + \phi_{yy}^2 + \psi_{yyy}^2 + \psi_{yy}^2)dy + \psi_{yy}^2\int_0^{+\infty}\phi_{yy}^2\,dy\right] \\ &\le CN(T)(\|\phi_y\|_1^2 + \|\psi_y\|_2^2),\end{aligned} \tag{7.64}$$

which, together with Lemmas 7.3–7.8, yields the inequality (7.7). To prove the inequality (7.8), we differentiate the system $(7.1)_1$ with respect to y, multiply it by ϕ_y, and integrate the resulting equality with respect to y to get

$$\frac{d}{dt}\|\phi_y(t)\|^2 = 2\int_0^{+\infty} \psi_{yy}\phi_y dy + 2x'\int_0^{+\infty} \phi_y\phi_{yy}dy + 2\int_0^{\infty} \gamma' V_T'\phi_y dy. \tag{7.65}$$

Integrating (7.65) over $[0,+\infty)$, we get (7.8) for ϕ due to (7.7). In the same way, we can also prove (7.8) for ψ. Proposition 7.1 is proved. □

THEOREM 7.9. *Suppose that the assumptions of Theorem* 4.1 *hold; then the initial-boundary value problem* (7.1) *has a unique global solution* $(\phi,\psi)\in X([0,+\infty))$, $\gamma(t)\in C^1[0,\infty)$ *satisfying the inequalities* (7.7) *and* (7.8) *for any* $t\geq 0$. *Moreover, the solution is asymptotically stable:*

$$\lim_{t\to\infty}\sup_{y\in\Re_+}|(\phi_y,\psi_y)(y,t)| = 0, \qquad \lim_{t\to\infty}\gamma(t) = \Gamma,$$

where Γ *is determined by* (4.11).

Proof. The assumption (4.6) gives $\|\phi_0,\psi_0\|_2 < C_4\delta^2 \leq \varepsilon_4 = \sqrt{C_2(1+C_4^2)}\delta$ due to Proposition 5.1. From Proposition 6.1, there exists a positive time $T_0 = T_0(\varepsilon_4) > 0$ such that a unique local solution $(\phi,\psi)(y,t)$, $\gamma(t)$ of (7.1) exists in $[0,T_0]$ satisfying $(\|\phi\|_2,\|\psi\|_2) \leq C_1C_4\delta^2 \leq \sqrt{C_2(1+C_4^2)}\delta$ and $|\gamma(t)| \leq |\bar{s}|t$. We now assume that for some positive integer $n>1$ there exists a unique local solution $(\phi,\psi)(y,t)$, $\gamma(t)$ of (7.1) in $[0,(n-1)T_0]$ satisfying $(\|\phi\|_2,\|\psi\|_2)(t)\leq\sqrt{C_2(1+C_4^2)}\delta$ and $|\gamma(t)|\leq|\bar{s}|t$. From Proposition 6.1, we know the solution (ϕ,ψ), $\gamma(t)$ can be extended to the interval $[(n-1)T_0,nT_0]$. Furthermore, Proposition 7.1 yields $(\|\phi\|_2,\|\psi\|_2)(t)\leq\sqrt{C_2(1+C_4^2)}\delta = \varepsilon_4$, and $|\gamma(t)|\leq|\bar{s}|t$ still hold when $t\in[(n-1)T_0,nT_0]$. Repeating the above argument, we get the existence of a unique global solution $(\phi,\psi)\in X([0,+\infty))$, $\gamma(t)\in C^1[0,\infty)$ satisfying the inequalities (7.7) and (7.8) for any $t\geq 0$. By (7.7), $\|(\phi_y,\psi_y)(t)\|_1$ tends to zero as time tends to infinity. By the Sobolev embedding theorem, we obtain

$$\sup_{y\in\Re_+}|(\phi_y,\psi_y)(y,t)|\to 0 \quad \text{as } t\to+\infty. \tag{7.66}$$

On the other hand, the boundary condition (5.9) yields

$$\gamma(t)(v_*-\bar{v}) = -\phi(0,t) - (s-\bar{s})\int_0^t [v_* - \bar{V}_s(0,t)]dt + \phi_0(0) - \sigma(t). \tag{7.67}$$

It is noted that $|\phi(0,t)|\leq\|\phi\|_1$ and $\sigma(t)$ tend to zero when $t\to\infty$ due to (7.7) and Lemma 7.2. From (7.67), we have

$$\begin{aligned}\lim_{t\to\infty}\gamma(t) = \Gamma = \frac{1}{\bar{v}-v_*}\Big\{&(s-\bar{s})\int_0^{\infty}[v_* - V_s(\bar{s}t - st + \alpha - \beta)]dt \\ &+ \int_0^{\infty}[v_0(y) - V_T(y) - V_s(y+\alpha-\beta) + v_*]dy\Big\}.\end{aligned} \tag{7.68}$$

Proof of Theorem 4.1. From Theorem 7.9, Theorem 4.1 is obtained at once. □

REFERENCES

[1] E. Fermi, *Thermodynamics*, Dover, New York, 1956.

[2] W. Grenier, L. Neise, and H. Stocker, *Thermodynamics and Statistical Mechanics*, Springer-Verlag, New York, 1995.

[3] F. M. Huang, A. Matsumura, and X. D. Shi, *Viscous shock wave and boundary layer solution to an inflow problem for compressible viscous gas*, Comm. Math. Phys., 239 (2003), pp. 261–285.

[4] F. M. Huang, A. Matsumura, and X. D. Shi, *A gas-solid free boundary problem for a compressible viscous gas*, SIAM J. Math. Anal., 34 (2003), pp. 1331–1355.

[5] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice–Hall, Englewood Cliffs, NJ, 1964.

[6] I. A. Kaliev and A. V. Kazhikhov, *Well-posedness of a gas-solid phase transition problem*, J. Math. Fluid Mech., 1 (1999), pp. 282–308.

[7] S. Kawashima and Y. Nikkuni, *Stability of stationary solutions to the half-space problem for the discrete Boltzmann equation with multiple collisions*, Kyushu J. Math., 54 (2000), pp. 233–255.

[8] A. Kazhikhov, *On the theory of boundary value problems for equations of the one-dimensional time dependent motion of a viscous heat-conducting gas,* Comm. Math. Phys., 82 (1981/82), pp. 37–62.

[9] T. P. Liu and K. Nishihara, *Asymptotic behavior for scalar viscous conservation laws with boundary effect*, J. Differential Equations, 133 (1997), pp. 296–320.

[10] T. P. Liu and S. H. Yu, *Propagation of a stationary shock layer in the presence of a boundary*, Arch. Ration. Mech. Anal., 139 (1997), pp. 57–82.

[11] A. Matsumura, *Inflow and outflow problems in the half space for a one-dimensional isentropic model system of compressible viscous gas*, Methods Appl. Anal., 8 (2001), pp. 645–666.

[12] A. Matsumura and M. Mei, *Convergence to travelling fronts of solutions of the p-system with viscosity in the presence of a boundary*, Arch. Ration. Mech. Anal., 146 (1999), pp. 1–22.

[13] A. Matsumura and K. Nishihara, *On the stability of traveling wave solutions of a one-dimensional model system for compressible viscous gas*, Japan J. Appl. Math., 2 (1985), pp. 17–25.

[14] A. Matsumura and K. Nishihara, *Global asymptotics toward the rarefaction wave for solutions of viscous p-system with boundary effect*, Quart. Appl. Math., 58 (2000), pp. 69–83.

[15] A. Matsumura and K. Nishihara, *Large time behaviors of solutions to an inflow problem in the half space for a one-dimensional system of compressible viscous gas*, Comm. Math. Phys., 222 (2001), pp. 449–474.

[16] T. Nagasawa, *On the one-dimensional free boundary problem for the heat-conductive compressible viscous gas*, in Recent Topics in Nonlinear PDE IV, North-Holland Math. Stud. 160, North-Holland, Amsterdam, 1989, pp. 83–99.

[17] X. D. Shi, *On the stability of rarefaction wave solutions for viscous p-system with boundary effect,* Acta Math. Appl. Sin. Engl. Ser., 19 (2003), pp. 341–352.

SIAM J. MATH. ANAL.
Vol. 36, No. 2, pp. 523–551

PLANE-LIKE MINIMAL SURFACES IN PERIODIC MEDIA WITH EXCLUSIONS*

MONICA TORRES†

Abstract. We consider minimal surfaces in a medium with exclusions (voids). This extends the results given in [*Comm. Pure Appl. Math.*, 54 (2001), pp. 1403–1441] to the case of a degenerate metric such that the area of a surface of codimension 1 is measured by neglecting the parts inside the exclusions. We prove that, given any plane in the medium, there is at least one minimal surface that always stays at a bounded distance from the plane. We also explore the connections of this problem with the theory of homogenization of Hamilton–Jacobi equations.

Key words. minimal surfaces, sets of finite perimeter, homogenization, periodic media

AMS subject classification. 35R99

DOI. 10.1137/S0036141001399970

1. Introduction.

The recent results in [14] consider a generalization of the problem of minimal surfaces in periodic media and show that, given a metric with periodic coefficients, there exists a number M so that one can find a minimizer in any strip of width M. The width M is independent of the orientation of the strip. Moreover, the minimizers constructed in [14] have the property that, when folded to the fundamental domain, they are laminations. For a discussion on the history of the problem of constructing minimizers that are asymptotic to a plane we refer the reader to [14] and the references therein.

The goal of this paper is to extend the results of [14] to a situation where the medium has exclusions, i.e., regions for which the metric vanishes. We also discuss the behavior of the minimizers near the exclusions, which is an issue not considered in [14]. Since similar situations of media with exclusions appear naturally in the theory of homogenization, we consider in this paper the relation of the minimizers with the theory of homogenization, and we develop several explicit calculations.

We recall that minimal surfaces can be studied using geometric measure theory (see, e.g., [26, 34]) in which the surfaces are interpreted as currents, i.e., dual to forms. Then the laminations can be interpreted as homologically minimizing currents (see, for instance, [6, 5, 4]). One can also study minimal surfaces by considering the surfaces as boundaries of sets in which the perimeter is defined in a weak sense (see, e.g., [27]).

In this paper we will follow the approach of locally finite perimeter sets, which is the one followed in [14]. For the problem considered in this paper, this approach is more advantageous because the fundamental domain is a manifold with boundary, and the theory of homologically minimizing currents in manifolds with boundary is not readily available to our knowledge. We refer the reader to [27, 25, 2] for a comprehensive survey on the theory of sets of finite perimeter.

The setting of the problem is as follows: the space $\mathbb{R}^n$ is considered as the lattice of cubes $[0,1]^n + \mathbb{Z}^n$ where each cube has an internal exclusion. If I denotes the

*Received by the editors December 20, 2001; accepted for publication (in revised form) November 14, 2003; published electronically July 29, 2004. This work was supported by a Research Assistanship from Luis A. Caffarelli.

http://www.siam.org/journals/sima/36-2/39997.html

†Department of Mathematics, Northwestern University, 2033 Sheridan Rd., Evanston, IL 60208-2730 (torres@math.northwestern.edu).

exclusion contained in $Y = [0,1]^n$, we assume the following:

1. I is compact, connected, and has Lipschitz boundary.
2. The distance between I and the boundary of Y, which we shall denote by α, is strictly positive.
3. Any other exclusion is of the form $I + z$ for some $z \in \mathbb{Z}^n$; i.e., the exclusions are periodic.

Once we have set up the domain for our problem, we proceed to explain our definition of *minimal surface*, which is made precise in section 2. If Σ is a surface in $\mathbb{R}^n$ of codimension one, we consider the following procedure for measuring the area of Σ: the portions that are inside the exclusions do not contribute to the area, and outside the exclusions the area is measured in the standard way. We say that Σ is a *minimal surface* if Σ minimizes area outside the exclusions. This means, loosely speaking, that any compact perturbation to Σ increases its area outside the exclusions.

We can now introduce the main result of this paper, which reads as follows: Under the assumptions 1, 2, and 3 given above, there exists a universal constant C (that depends only on n an α) such that, for every $(n-1)$-dimensional hyperplane Π, we can find a minimal surface Σ satisfying $d(\Pi, \Sigma) \leq C$.

The minimizers constructed in this paper are regular away from the boundaries of the exclusions. This follows directly from standard interior regularity theory for minimal surfaces (see Remark A.2). For the case when the exclusions have C^2 boundaries, the regularity of the minimizers near the boundary of the exclusions is a consequence of [29], where techniques of geometric measure theory are used to prove optimal regularity for codimension one minimal surfaces with a free boundary.

An important property of the surfaces constructed in this paper is that they meet the exclusions orthogonally. This means, loosely speaking, that the intersection of the minimizers with the exclusions looks like two perpendicular hyperplanes (in a small neighborhood). This orthogonality result can be deduced (once we have the regularity of the minimizers up to the boundary of the exclusions) by studying the first variation of the area. An analysis of the Euler–Lagrange equation is done in [31], where numerical and theoretical analysis for minimal surfaces involving two media is performed. We discuss the orthogonality property in section 6, and we explain how it can be obtained from [31]. For a proof of this orthogonality property, in the context of geometric measure theory, we refer the reader to [29].

The existence of plane-like minimizers implies that, in spite of having a heterogeneous media, the minimizer looks like a plane (homogeneous media) when seen from a far distance. This suggests connections with the theory of homogenization of PDEs, which studies the asymptotic behavior of a family of PDEs that oscillate with small period of size $\epsilon > 0$. The last section of this paper begins to explore the connection with the theory of homogenization of Hamilton–Jacobi equations. Hamilton–Jacobi equations arise in optimal control, differential games, geometric optics, calculus of variations, etc., and their solutions are understood in the *viscosity sense*. We refer the reader to [8, 23, 7] and the references therein for the definitions and basic properties of viscosity solutions that we will use in this paper.

The study of asymptotics of solutions of Hamilton–Jacobi equations is a fundamental question, as well as their applications to mathematical sciences. The homogenization of Hamilton–Jacobi equations has been extensively studied (see, for instance, [32, 21, 22, 15, 9]). The homogenized equation is also a Hamilton–Jacobi equation, and the corresponding Hamiltonian, usually denoted by $\overline{H}$, is called the *effective Hamiltonian*. It is a difficult but interesting task to find explicit formulas for $\overline{H}$. The

references [22, 19, 20, 17, 16, 24] contain results in this direction. In this paper, we introduce a particular example, and we perform several explicit computations in search of its corresponding effective Hamiltonian. The homogenization of Hamilton–Jacobi equations in perforated domains was treated in [30], where both the Neumann-type and the Dirichlet boundary value problems were considered. A generalization of [30] has been studied in [1].

The organization of the paper is as follows.

Section 2 contains the proof of the existence of minimizers.

Section 3 uses some subadditivity properties of sets of finite perimeter to define an *infimal minimizer* which is contained in all the other minimizers and satisfies several monotonicity properties. The results presented in section 3 are contained in [14], but for clarity of the exposition we present again the proofs with more detail.

Section 4 deals with the proof of a geometric property that is specific to the infimal minimizer. This property is analogous to the so-called Birkhoff property in Aubry–Mather theory.

Section 5 contains the proof of the main theorem, which relies on the fact that minimizers must satisfy some *density estimates*. The geometric property proven in section 4, together with the density estimates, allows us to prove that the infimal minimizer is contained in a band whose width is independent of the direction of the plane.

Section 6 discusses the behavior of the minimizers near the boundaries of the exclusions.

Section 7 explores the connection with the theory of homogenization of Hamilton–Jacobi equations and contains several explicit computations.

We present at the end an appendix that includes the main definitions concerning sets of finite perimeter, as well as several remarks regarding some conventions and notation that we are using throughout the paper.

2. Existence of minimizers. We proceed now to prove the existence of minimizers. We refer the reader to the appendix for the definition and main properties of sets of finite perimeter. As explained before, our setting in this paper is $\mathbb{R}^n$ with exclusions (voids) that satisfy the three properties stated in the introduction.

We denote I as the exclusion contained in $[0,1]^n$. We let $\mathcal{I}$ denote the union of all exclusions and O its complement; i.e.,

$$\mathcal{I} = \bigcup_{k\in\mathbb{Z}^n} (I+k), \tag{1}$$

$$O = \mathbb{R}^n \backslash \mathcal{I}. \tag{2}$$

We let $\omega \in \mathbb{R}^n$, and we consider first the case when $\omega \in \mathbb{Q}^n$. Given $\tilde{M} \in \mathbb{R}$, we define

$$\Gamma_{\omega,\tilde{M}} = \left\{ x \in \mathbb{R}^n : x \cdot \frac{\omega}{|\omega|} \leq \tilde{M} \right\}, \tag{3}$$

where $\frac{\omega}{|\omega|}$ is the outward unit normal to $\partial\Gamma_{\omega,\tilde{M}}$. We denote T_k as the translation operator by $k \in \mathbb{Z}^n$; that is, $T_k(x) = x+k$, $x \in \mathbb{R}^n$. Given $N \in \mathbb{N}^+$ and $M > 0$, we define

$$\begin{aligned} A_{S_1,S_2} = \{&E : E \text{ is a set of finite perimeter,} \\ &S_1 \subset E \subset S_2, T_{Nk}E = E \ \forall k \in \mathbb{Z}^n, \omega \cdot k = 0\}, \end{aligned} \tag{4}$$

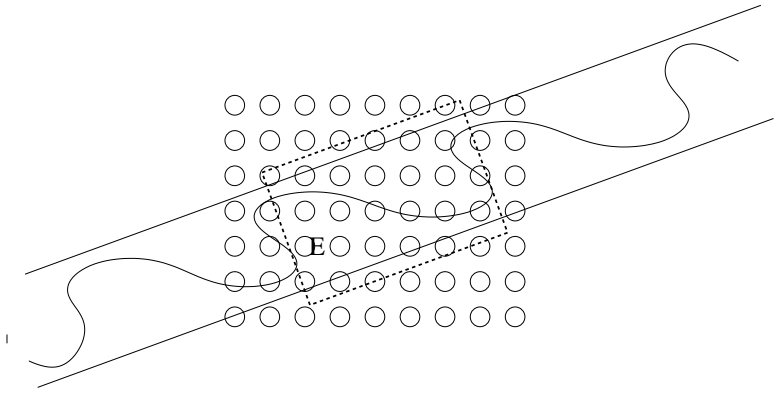

FIG. 1. *Diagram showing parallel plane restrictions and the period for minimization.*

where $S_1 = \Gamma_{\omega,0}$ and $S_2 = \Gamma_{\omega,M}$. We will refer to the sets $\Pi_1 \equiv \{x \in \mathbb{R}^n : x \cdot \omega = 0\}$ and $\Pi_2 \equiv \{x \in \mathbb{R}^n : x \cdot \frac{\omega}{|\omega|} = M\}$ as the parallel plane restrictions. Throughout this paper, we consider (without loss of generality) sets of finite perimeter that satisfy Remark A.1.

Since ω is rational, the sets in A_{S_1,S_2} can be identified with sets in the manifold

$$\Gamma_{\omega,M}/\approx, \tag{5}$$

where $\approx$ is the equivalence relation defined by

$$x \approx y \Longleftrightarrow x = y + Nk \quad \text{for some} \quad k \in \mathbb{Z}^n, \omega \cdot k = 0. \tag{6}$$

The space defined in (5) is $[-\infty, M] \times \mathbb{T}^{n-1}$. Moreover, we can identify the period of the class A_{S_1,S_2} as $[-\epsilon, M+\epsilon] \times \mathbb{T}^{n-1}$ for a fixed $\epsilon > 0$ (see Figure 1). We define

$$\Omega = ([-\epsilon, M+\epsilon] \times \mathbb{T}^{n-1}) \setminus \mathcal{I}. \tag{7}$$

For each set $E \in A_{S_1,S_2}$, we consider

$$J(E) = \int_\Omega |D\varphi_E|, \tag{8}$$

where the measure $|D\varphi_E|$ is introduced in Definition A.4. We let $\beta = \inf_{E \in A_{S_1,S_2}} J(E)$ and $\{E_j\}$ be a sequence such that $J(E_j) \to \beta$. This implies that the sequence $\{\int_\Omega |D\varphi_{E_j}|\}$ is uniformly bounded. Since the exclusions have at least Lipschitz boundary, it follows from Theorem A.2 that $BV(\Omega)$ is relatively compact in $L^1(\Omega)$. Therefore, there exists a convergent subsequence, which we denote again by $\{E_j\}$, in $L^1(\Omega)$. We let $E_0 \in L^1(\Omega)$ be the limit. Using Proposition A.1 we obtain

$$\int_\Omega |D\varphi_{E_0}| \leq \liminf \int_\Omega |D\varphi_{E_j}|.$$

Thus,

$$J(E_0) = \inf_{E \in A_{S_1,S_2}} J(E).$$

We make the following definitions.

DEFINITION 2.1. *Any $E \in A_{S_1,S_2}$ that satisfies $J(E) = J(E_0)$ shall be called a minimizer corresponding to the class A_{S_1,S_2}, or simply a minimizer, when it is not necessary to specify the class.*

DEFINITION 2.2. *We say that the minimizer E is an unconstrained minimizer if there exists a universal constant $\tilde{M} > 0$ such that, for all $M \geq \tilde{M}$ and all $\epsilon \geq 0$, E is a minimizer corresponding to the class $A_{\Gamma_{\omega,-\epsilon},\Gamma_{\omega,M}}$.*

DEFINITION 2.3. *We say that the minimizer E is a class A minimizer if, for any open ball B_R,*

$$\int_{B_R \cap O} |D\varphi_E| = \inf \left\{ \int_{B_R \cap O} |D\varphi_F| : F \text{ is a set of finite perimeter, } \operatorname{spt}(\varphi_F - \varphi_E) \subset B_R \right\}.$$

DEFINITION 2.4. *We say that $\Sigma \subset \mathbb{R}^n$ is* a minimal surface *if $\Sigma = \partial E$, where E is a class A minimizer.*

Remark 2.1. We shall prove later (Proposition 5.2) that if the distance between the two restrictions Π_1 and Π_2 is large enough (independently of the slope of the restrictions), then there exists at least one unconstrained class A minimizer. That is, if the distance between Π_1 and Π_2 is large enough, then the restrictions do not interfere in the minimization, which means that they do not prevent the minimizers from doing "better."

The following lemma tells us that, without loss of generality, we can assume that minimizers are closed sets.

LEMMA 2.1. *If E is a minimizer corresponding to the class A_{S_1,S_2}, then there exists a closed set $\tilde{E}$, which is also a minimizer for the class A_{S_1,S_2}.*

Proof. Define $\tilde{E} = E \cup \partial E$ (see Definition A.8). We have that $\tilde{E}$ is closed. We need to prove that $\tilde{E}$ and E differ (outside the exclusions) on a set of $\mathcal{L}^n$-measure zero. Since the restrictions Π_1 and Π_2 have $\mathcal{L}^n$-measure zero, we need only to consider the set $\mathcal{K} \equiv \partial E \cap O \cap B_{\Pi_1,\Pi_2}$, where B_{Π_1,Π_2} is the open slab enclosed by Π_1 and Π_2. Since E minimizes area outside the exclusions, it follows from Lemma A.5 that if $x \in \mathcal{K}$ has density γ_x, then $0 < \gamma_x < 1$ (see Definition A.6 for the definition of density of a point), which implies that such x is not a Lebesgue point for φ_E. Therefore, from Definition A.6 we obtain that $\mathcal{L}^n(\mathcal{K}) = 0$. We can now prove that $\tilde{E}$ is a minimizer, which is a consequence of the fact that the sets E and $\tilde{E}$ differ (outside the exclusions) on a set of $\mathcal{L}^n$-measure zero. In fact, if $V \subset O$ is any open set, we have

$$\begin{aligned}
\int_V |D\varphi_E| &= \sup \left\{ \int_V \varphi_E \operatorname{div} g : g \in C_0^1(V;\mathbb{R}^n), \quad |g(x)| \leq 1, \quad \text{for} \quad x \in V \right\} \\
&= \sup \left\{ \int_V \varphi_{\tilde{E}} \operatorname{div} g : g \in C_0^1(V;\mathbb{R}^n), \quad |g(x)| \leq 1, \quad \text{for} \quad x \in V \right\} \\
&= \int_V |D\varphi_{\tilde{E}}|,
\end{aligned}$$

which proves that both measures coincide outside the exclusions. ∎

Remark 2.2. From now on, we shall assume that minimizers are closed sets.

We now proceed to prove that a minimizer (minus the exclusions) is connected for the case when the exclusions are simply connected sets and have at least C^1

boundaries. We remark that we do not need the connectivity of the minimizers in any of the proofs in this paper, but we present the result since it is interesting by itself.

LEMMA 2.2. *Let E be a minimizer corresponding to the class A_{S_1,S_2}. Assume that the exclusions are simply connected and have at least C^1 boundaries; then $E \cap O$ is connected.*

Proof. We let $\tilde{E} = E \cap O$. We prove that $\tilde{E}_{int}$ is connected. We proceed by contradiction and assume that

$$\tilde{E}_{int} = A \cup B, \tag{9}$$

where A,B are two disjoints open sets. Since $\Gamma_{\omega,0} \cap O$ is connected, it must be contained in either A or B. We assume that $\Gamma_{\omega,0} \cap O \subset A$, and we let $F = \mathbb{R}^n \setminus \tilde{E}$. Since E minimizes area outside the exclusions it follows that the points in ∂F have uniform density; i.e., there exists a universal constant C such that

$$|F \cap B(x,r)| \geq Cr^n, \quad x \in \partial F, \quad r \leq r_0, \tag{10}$$

for some small enough universal constant r_0. We prove this claim in Lemma A.6. We now proceed to prove that (10) implies that we can approximate $\tilde{E}_{int}$ from inside with smooth sets. We recall (see [2]) that sets of finite perimeter in $\mathbb{R}^n$ can be approximated in measure by open sets with smooth boundaries in such a way that we also have convergence of perimeters to perimeters. It is not, in general, possible to approximate a set of finite perimeter E by C^∞ sets contained inside E, nor it is possible from outside (see [27, p. 24] for a counterexample). However, in our case, we prove in Lemma A.7 that we can find sequences of sets $\{A_t\}$, $\{B_t\}$ with smooth boundaries satisfying

$$A_t \subset\subset A, \quad B_t \subset\subset B \tag{11}$$

and

$$\mathrm{Per}(A \cup B) = \lim_{t\to 0} \mathrm{Per}(A_t \cup B_t), \quad A_t \to A \quad B_t \to B \quad \text{in measure.} \tag{12}$$

From (11), (12), and the lower semicontinuity property given in Proposition A.1 we obtain

$$\begin{aligned} \mathrm{Per}(A \cup B) &= \lim_{t\to 0} \mathrm{Per}(A_t \cup B_t) \\ &= \lim_{t\to 0} \mathrm{Per}(A_t) + \lim_{t\to 0} \mathrm{Per}(B_t) \\ &\geq \mathrm{Per}(A) + \mathrm{Per}(B). \end{aligned}$$

This is a contradiction since we can eliminate B and obtain a set with less perimeter. ∎

3. Infimal minimizer. The minimizer we have just constructed may not be unique. However, we can prove the existence of an infimal minimizer, that is, a minimizer that is contained in any other minimizer. The results presented in this section are contained in [14], but, for clarity of the exposition, we present here the proofs with more detail.

In this section, Ω denotes the set defined in (7).

THEOREM 3.1. *There exists $E_* \in A_{S_1,S_2}$ such that, if E is any other minimizer, then $E_* \subset E$. We refer to E_* as the infimal minimizer.*

Proof. We denote $\mathcal{B}$ as the set of all minimizers. We have that $\mathcal{B} \subset L^1(\Omega)$. If $E_1, E_2 \in \mathcal{B}$, by Theorem A.4 we have

$$\text{Per}(E_1 \cap E_2, \Omega) + \text{Per}(E_1 \cup E_2, \Omega) \leq \text{Per}(E_1, \Omega) + \text{Per}(E_2, \Omega). \tag{13}$$

Since $E_1 \cup E_2$ is an admissible set we have $\text{Per}(E_1 \cup E_2, \Omega) \geq \text{Per}(E_1, \Omega)$. Since $\text{Per}(E_1, \Omega) = \text{Per}(E_2, \Omega)$ and using inequality (13), it follows that

$$\text{Per}(E_1 \cap E_2, \Omega) \leq \text{Per}(E_1, \Omega),$$

which implies that $E_1 \cap E_2$ is also a minimizer. Since we can uniformly bound the perimeters of minimizers in Ω, it follows from Proposition A.1 and Theorem A.2 that $\mathcal{B}$ is a compact subset of $L^1(\Omega)$. Since $L^1(\Omega)$ is separable, $\mathcal{B}$ is also separable. We let $\{E_j\}$ denote a dense subset of $\mathcal{B}$, and we define

$$\tilde{E_n} = \bigcap_{j=1}^{n} E_j.$$

Since $\tilde{E_n}$ is a minimizer and $\tilde{E}_{n+1} \subset \tilde{E_n}$ with $|\tilde{E_1} \cap \Omega| < \infty$, it follows that $|\tilde{E_n} \cap \Omega| \to |\bigcap_{n=1}^{\infty} \tilde{E_n} \cap \Omega|$, and therefore $\tilde{E_n} \to \bigcap_{n=1}^{\infty} \tilde{E_n}$ in $L^1(\Omega)$. We define

$$E_* = \bigcap_{n=1}^{\infty} \tilde{E_n}.$$

By Proposition A.1

$$\text{Per}(E_*, \Omega) \leq \liminf \text{Per}(\tilde{E_n}, \Omega),$$

which implies that E_* is a minimizer.

If E denotes any other minimizer we claim that $|(E_* \backslash E) \cap \Omega| = 0$. We proceed by contradiction and assume this is not true; i.e., $|(E_* \backslash E) \cap \Omega| > \delta > 0$. Since $\{E_j\}$ is a dense subset of $\mathcal{B}$, we can find E_k such that $|(E_k \backslash E) \cap \Omega| < \frac{\epsilon}{2}, \epsilon < \delta$. We choose N large enough such that $\tilde{E_N} \subset E_k$ and $|(E_* \backslash \tilde{E_N}) \cap \Omega| < \frac{\epsilon}{2}$. We have

$$\begin{aligned} |(E_* \backslash E) \cap \Omega| &\leq |(E_* \backslash \tilde{E_N}) \cap \Omega| + |(\tilde{E_N} \backslash E) \cap \Omega| \\ &\leq |(E_* \backslash \tilde{E_N}) \cap \Omega| + |(E_k \backslash E) \cap \Omega| \\ &\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon < \delta, \end{aligned}$$

which is a contradiction. Since E and E_* are both minimizers and are closed, if follows from Remark A.1 that $E_* \subset E$. ∎

COROLLARY 3.1. *The infimal minimizer is unique.*

We let $M_1 < 0$ and $M_2 > 0$ be such that $T_2 := \Gamma_{\omega, M_2} \subset S_2$. We have $T_1 := \Gamma_{\omega, M_1} \subset S_1$ and $T_1 \subset T_2$. The following proposition shall be used later to establish properties of the infimal minimizer.

PROPOSITION 3.1. *If E is a minimizer corresponding to the class A_{S_1, S_2} and L a minimizer corresponding to the class A_{T_1, T_2}, then*

(a) *$E \cap L$ is a minimizer corresponding to the class A_{T_1, T_2};*

(b) *$E \cup L$ is a minimizer corresponding to the class A_{S_1, S_2};*

(c) $E_{*, T_1, T_2} \subset E_{*, S_1, S_2}$.

Proof. We note that $E \cup L \in A_{S1,S2}$ and $E \cap L \in A_{T_1,T_2}$. Using Theorem A.4 and since $\mathrm{Per}(E,\Omega) \le \mathrm{Per}(E \cup L,\Omega)$, it follows that

$$\begin{aligned}\mathrm{Per}(E \cap L,\Omega) + \mathrm{Per}(E,\Omega) &\le \mathrm{Per}(E \cap L,\Omega) + \mathrm{Per}(E \cup L,\Omega)\\ &\le \mathrm{Per}(E,\Omega) + \mathrm{Per}(L,\Omega),\end{aligned}$$

which implies that $\mathrm{Per}(E \cap L,\Omega) \le \mathrm{Per}(L,\Omega)$; i.e., $E \cap L$ is a minimizer in the class A_{T_1,T_2}. In the same way we prove (b). In order to prove (c) we note that, by (a), $E_{*,T_1,T_2} \cap E_{*,S_1,S_2}$ is a minimizer corresponding to the class A_{T_1,T_2}, and hence

$$\begin{aligned}&E_{*,T_1,T_2} \subset (E_{*,T_1,T_2} \cap E_{*,S_1,S_2})\\ \Rightarrow\ &E_{*,T_1,T_2} \subset E_{*,S_1,S_2}. \qquad \square\end{aligned}$$

4. Birkhoff property. We denote E as the infimal minimizer corresponding to the class A_{S_1,S_2}. We recall that T_k denotes the translation operator by $k \in \mathbb{Z}^n$; that is, $T_k(x) = x + k$, $x \in \mathbb{R}^n$. The infimal minimizer satisfies an important geometric property (quite analogous to the property called Birkhoff in Aubry–Mather theory), which is proven in [14].

LEMMA 4.1. *If $k \in \mathbb{Z}^n$, we have the following:*

(a) *If $k \cdot \omega \le 0$, then $T_k E \subset E$.*

(b) *If $k \cdot \omega \ge 0$, then $E \subset T_k E$.*

Proof. (a) We let $T_1 = T_k(S_1)$ and $T_2 = T_k(S_2)$, where as before $S_1 = \{x \in \mathbb{R}^n : x \cdot \omega \le 0\}$ and $S_2 = \{x \in \mathbb{R}^n : x \cdot \omega \le M\}$. If $k \cdot \omega \le 0$ we have that $T_1 \subset S_1$, $T_2 \subset S_2$, and $T_1 \subset T_2$. We note that $T_k E$ is the infimal minimizer in A_{T_1,T_2}. By Proposition 3.1(c) we have $T_k E \subset E$.

(b) If $k \cdot \omega \ge 0$, we have that $S_1 \subset T_1$, $S_2 \subset T_2$, and $T_1 \subset T_2$. Since $T_k E$ is the infimal minimizer in A_{T_1,T_2}, by Proposition 3.1(c) it follows that $E \subset T_k E$. $\square$

We make the following important observation.

Remark 4.1. From (a) and (b) above, we have that if $k \cdot \omega = 0$, then $T_k E = E$. This implies that even though in the minimization of (8) the size of the period of the candidate sets is given by the number N (recall the definition (4)), the infimal minimizer E has indeed a periodicity that depends only on the slope of the restrictions.

DEFINITION 4.1. *Given any two hyperplanes Π and $\tilde{\Pi}$ parallel to the restrictions, we denote $B_{\Pi,\tilde{\Pi}}$ as the open* slab *enclosed by Π and $\tilde{\Pi}$.*

The following two results are needed in order to handle the exclusions. They play the analogous role that the lower estimates in [14] play for the case without exclusions.

LEMMA 4.2. *If $\mathcal{C} \subset B_{\Pi_1,\Pi_2}$ is a cube of edge length $l \ge 5$ with sides parallel to the coordinate axis and integer vertices, we have the following:*

(a) *If $\mathcal{C} \subset (\mathbb{R}^n \backslash E)$, then there exists $0 < M_a < M$ such that $E \subset \Gamma_{w,M_a}$.*

(b) *If $\mathcal{C} \subset E$, then there exist $0 < M_b < M$ such that $\Gamma_{w,M_b} \subset E$.*

Proof. (a) We denote $\tilde{\Pi}$ as the hyperplane parallel to the restrictions Π_1 and Π_2 in such a way that the intersection $\tilde{\Pi} \cap \mathcal{C}$ consists only of the edge of $\mathcal{C}$ that is closer to the lower restriction Π_1. The equation of $\tilde{\Pi}$ is $x \cdot \frac{\omega}{|\omega|} = \tilde{M}$ for some $0 < \tilde{M} < M$. We define

$$\mathcal{D} = \bigcup_{\omega \cdot k \ge 0} T_k C. \tag{14}$$

If $\omega \cdot k \ge 0$, we claim that $T_k C \subset \mathbb{R}^n \setminus E$. In fact, if this is not true, there exist $x \in C$, $y \in E$ such that $T_k(x) = y$. Then $T_{-k}(y) = x$, which is a contradiction since

Lemma 4.1 implies $T_{-k}E \subset E$. We conclude that the set $\mathcal{D} \subset \mathbb{R}^n \setminus E$. We note that $\mathcal{D}$ contains the set $\{x \cdot \frac{\omega}{|\omega|} \geq \tilde{M} + \sqrt{n}\}$. If we define $M_a = \tilde{M} + \sqrt{n}$, we obtain that $E \subset \Gamma_{w,M_a}$.

(b) We denote $\tilde{\Pi}$ as the hyperplane parallel to the restrictions Π_1 and Π_2 in such a way that the intersection $\tilde{\Pi} \cap \mathcal{C}$ consists only of the edge of $\mathcal{C}$ that is closer to the upper restriction Π_2. The equation of $\tilde{\Pi}$ is $x \cdot \frac{\omega}{|\omega|} = \tilde{M}$ for some $0 < \tilde{M} < M$. We define

$$\mathcal{G} = \bigcup_{\omega \cdot k \leq 0} T_k C. \tag{15}$$

If $\omega \cdot k \leq 0$, it follows from Lemma 4.1 that $T_k E \subset E$, and therefore $T_k C \subset E$. We conclude that $\mathcal{G} \subset E$. We note that $\mathcal{G}$ contains the set $\{x \cdot \frac{\omega}{|\omega|} \leq \tilde{M} - \sqrt{n}\}$. If we define $M_b = \tilde{M} - \sqrt{n}$, we obtain $\Gamma_{w,M_b} \subset E$. ∎

We use the previous lemma to prove the following proposition.

PROPOSITION 4.1. *If $\mathcal{C} \subset B_{\Pi_1,\Pi_2}$ is a cube of edge length $l \geq 5$, with sides parallel to the coordinate axis and integer vertices, then we cannot have $\mathcal{C} \subset E$.*

Proof. We proceed by contradiction. We let M_b be the number given by Lemma 4.2(b), and we define $\Pi_b = \{x \in \mathbb{R}^n : x \cdot \frac{\omega}{|\omega|} = M_b\}$. By subtracting a small number $\epsilon > 0$ to M_b, if necessary, we can assume that $|\omega| M_b \in \mathbb{Q}$. Since $l \geq 5$, there exists $p \in \mathbb{Z}^n$ such that $p \in C \cap \Gamma_{w,M_b}$. We define $M_c = p \cdot \frac{\omega}{|\omega|}$, and we take $k \in \{x \cdot \frac{\omega}{|\omega|} = M_b - M_c\} \cap \mathbb{Z}^n$ (which can be chosen because $|\omega|(M_b - M_c) \in \mathbb{Q}$). Since $M_b - M_c = k \cdot \frac{\omega}{|\omega|}$ we have

$$\begin{aligned} T_{-k}(\Pi_b) &= \left\{ x - k : x \cdot \frac{\omega}{|\omega|} = M_b \right\} \\ &= \left\{ y : y \cdot \frac{\omega}{|\omega|} = M_b - k \cdot \frac{\omega}{|\omega|} \right\} \\ &= \left\{ y : y \cdot \frac{\omega}{|\omega|} = M_c \right\} := \Pi_c. \end{aligned}$$

The plane Π_c divides E in two parts, say E_1 and E_2, where $\Pi_1 \subset E_1$ and $\Pi_b \subset E_2$. We consider now the set $E_1 \cup T_{-k}(E_2 \backslash B_{\Pi_b,\Pi_c})$. Clearly, this set is also a minimizer contained (and not equal) in E. This contradicts the fact that E is the infimal minimizer, that is, a minimizer that is contained in any other minimizer. ∎

5. Proof of the main theorem. We proceed in this section to prove the main theorem. We recall that we are considering $\mathbb{R}^n$ as the lattice $[0,1]^n + \mathbb{Z}^n$ with periodic exclusions; i.e., each cube $[0,1]^n + k$ with $k \in \mathbb{Z}^n$ has an internal exclusion. If I denotes the exclusion contained $Y = [0,1]^n$, we assume the following:

1. I is compact, connected, and has Lipschitz boundary.
2. The distance between I and the boundary of Y, which we denote by α, is strictly positive.
3. Any other exclusion is of the form $I + z$ for some $z \in \mathbb{Z}^n$; i.e., the exclusions are periodic.

Remark 5.1. From now on, given the restrictions S_1 and S_2, *we work with the unique infimal minimizer E corresponding to the class A_{S_1,S_2}.*

Remark 5.2. In order to clarify exposition we use the same C to denote different universal constants.

We now state the main theorem.

THEOREM 5.1. *Assume that the exclusions satisfy* $1, 2$, *and* 3 *above. Then there exists a universal constant* C *(that depends only on* n *and* α*) such that, for every* $(n-1)$*-dimensional hyperplane* Π*, we can find a minimal surface* Σ *satisfying* $d(\Pi, \Sigma) \leq C$.

We recall from Definition 2.4 that a surface Σ is a minimal surface if it is the boundary of a class A minimizer (recall Definition 2.3), which means that any compact perturbation to Σ will increase its area outside the exclusions. The tool used to prove Theorem 5.1 is essentially a covering argument. This argument is similar to the one used in [14] to obtain the theorem for the case without exclusions. However, in our case we need to make several adjustments in order to extend the theorem to the case with exclusions. Lemmas 5.1, 5.3, and 5.4 are needed to handle the presence of exclusions. Using these lemmas we prove Propositions 5.1 and 5.2. Then Theorem 5.1 follows, for the case ω rational, from Proposition 5.3. Finally, we consider the case ω irrational at the end of this section.

LEMMA 5.1. *We let* E *denote the infimal minimizer corresponding to the class* A_{S_1,S_2}*, and we let* $x \in \partial E$*. If* Q_q *is a closed cube of edge length* q *(or a closed ball of radius* q*) containing* x *and such that* $Q_q \cap \Pi_1 = \emptyset$ *and* $Q_q \cap \Pi_2 = \emptyset$*, then*

$$\mathrm{Per}(E, Q_q^0 \cap O) \leq C q^{n-1},$$

where Q_q^0 *denotes the interior of the set* Q_q.

Proof. We can consider the set E as a candidate in the class with a period large enough (choosing N large enough in the definition (4)) in such a way that Q_q is completely contained inside the period $[0, M] \times \mathbb{T}^{n-1}$. Using Remark 4.1, it follows that the set E is a minimizer for the new class. Proceeding as in Lemmas A.1 and A.2 we can prove that, for almost every $0 < s < q$,

$$\mathrm{Per}(E \backslash Q_s, Q_q^0 \cap O) = \mathrm{Per}(E, (Q_q^0 \backslash Q_s) \cap O) + \mathcal{H}_{n-1}(\partial Q_s \cap E \cap O). \tag{16}$$

(In fact, we can use Lemma A.1 with $f(x) = \varphi_E$, $A = (Q_q^0 \backslash Q_s) \cap O$, and $\Omega = Q_q^0 \cap O$.) Since E is a minimizer we have

$$\mathrm{Per}(E, Q_q^0 \cap O) \leq \mathrm{Per}(E \backslash Q_s, Q_q^0 \cap O). \tag{17}$$

From (16) and (17) we obtain that, for almost every $0 < s < q$,

$$\begin{aligned} \mathrm{Per}(E, Q_q^0 \cap O) &\leq \mathrm{Per}(E, (Q_q^0 \backslash Q_s) \cap O) + \mathcal{H}_{n-1}(\partial Q_s \cap E \cap O) \\ &\leq \mathrm{Per}(E, (Q_q^0 \backslash Q_s) \cap O) + C s^{n-1}. \end{aligned} \tag{18}$$

We now choose a sequence $\{s_j\} \to q$ such that (18) holds for each s_j. If we let $j \to \infty$, we conclude that $\int_{Q_q^0 \cap O} |D\varphi_E| \leq C q^{n-1}$. We note that we can use $C = 2n$ if Q_q is a cube and $C = n w_n$ (where w_n is the volume of the n-dimensional unit ball) if Q_q is a ball. □

LEMMA 5.2. *We let* E *denote the infimal minimizer corresponding to the class* A_{S_1,S_2}*, and we let* $y \in \partial E$*. We assume that there exists* $\tilde{r} > 0$ *that satisfies* $B(y, \tilde{r}) \cap \Pi_1 = \emptyset$*,* $B(y, \tilde{r}) \cap \Pi_2 = \emptyset$*, and* $B(y, \tilde{r}) \subset O$*. Then there exists a universal constant* $C > 0$ *such that, for all* $r \leq \tilde{r}$*,*

$$\int_{B(y,r)} |D\varphi_E| \geq C r^{n-1}. \tag{19}$$

Proof. Since E minimizes area outside the exclusions we have, for all $r \leq \tilde{r}$,

$$\int_{B(y,r)} |D\varphi_E| \leq \mathcal{H}_{n-1}(E \cap \partial B(y,r)). \tag{20}$$

We define $V(r) = |E \cap B(y,r)|$, $r \leq \tilde{r}$. Using the isoperimetric inequality given in Lemma A.3 we have that

$$|E \cap B(y,r)| \leq C[\mathrm{Per}(E \cap B(y,r))]^{\frac{n}{n-1}}. \tag{21}$$

From Lemma A.2 and using (20) and (21) it follows that, for almost every $r \leq \tilde{r}$,

$$\begin{aligned} |E \cap B(y,r)| &\leq C[\mathrm{Per}(E \cap B(y,r), \mathbb{R}^n)]^{\frac{n}{n-1}} \\ &= C[\mathrm{Per}(E, B(y,r)) + \mathcal{H}_{n-1}(E \cap \partial B(y,r))]^{\frac{n}{n-1}} \\ &\leq C[\mathcal{H}_{n-1}(E \cap \partial B(y,r))]^{\frac{n}{n-1}}. \end{aligned}$$

Due to Remark A.1 it follows that $V(r) > 0$ for all $r \leq \tilde{r}$. Since $V'(r) = \mathcal{H}_{n-1}(E \cap \partial B(y,r))$ we have, for almost every $r \leq \tilde{r}$,

$$V(r) \leq CV'(r)^{\frac{n}{n-1}}. \tag{22}$$

If we divide (22) by $V(r)$, we obtain $C \leq V(r)^{\frac{1-n}{n}} V'(r) = (V(r)^{\frac{1}{n}})'$. If we integrate, we obtain $V(r)^{\frac{1}{n}} \geq Cr$; i.e.,$V(r) \geq Cr^n$ for all $r \leq \tilde{r}$. In the same way we can prove that $|(\mathbb{R}^n \backslash E) \cap B(y,r)| \geq Cr^n$,$r \leq \tilde{r}$. The isoperimetric inequality stated in Lemma A.4 gives us

$$\begin{aligned} \min\{|(\mathbb{R}^n \backslash E) \cap B(y,r)|, |E \cap B(y,r)|\} &\leq C \left(\int_{B(y,r)} |D\varphi_E| \right)^{\frac{n}{n-1}} \\ &\Rightarrow \\ Cr^n &\leq \left(\int_{B(y,r)} |D\varphi_E| \right)^{\frac{n}{n-1}}. \end{aligned}$$

We conclude that

$$\int_{B(y,r)} |D\varphi_E| \geq Cr^{n-1}.$$

This completes the proof of the lemma. □

LEMMA 5.3. *We let E denote the infimal minimizer for the class A_{S_1,S_2}, and we take $x \in \partial E \cap O$. We assume that $x \in Y$, where $Y = [0,1]^n + k$ for some $k \in \mathbb{Z}^n$, and we denote I as the exclusion contained in Y. We assume also that Y does not intersect the parallel plane restrictions Π_1 and Π_2. Then $\partial E \cap \partial Y_\alpha \neq \emptyset$, where $Y_\alpha = \{x \in Y : d(x,I) \geq \frac{\alpha}{2}\}$.*

Proof. We proceed by contradiction and assume that $\partial E \cap Y_\alpha = \emptyset$. This implies that $Y_\alpha \subset E_{int}$ or $Y_\alpha \subset \mathbb{R}^n \backslash E$. Assume that $Y_\alpha \subset E_{int}$. We define $\tilde{E} = E \cup Y$. From Lemma 5.2 it follows that $\tilde{E}$ has strictly less area than E, which is a contradiction. If we assume now that $Y_\alpha \subset \mathbb{R}^n \backslash E$, then we can define $\tilde{E} = E \setminus Y$. Again, Lemma 5.2 implies that the set $\tilde{E}$ has strictly less area than E, which is a contradiction. We conclude that $\partial E \cap \partial Y_\alpha \neq \emptyset$. □

LEMMA 5.4. *We let E denote the infimal minimizer corresponding to the class A_{S_1,S_2}, and we let $x \in \partial E \cap O$. We assume that $x \in Y$, where $Y = [0,1]^n + k$ for*

some $k \in \mathbb{Z}^n$. We assume also that Y is far away from the parallel plane restrictions Π_1 and Π_2. Then there exists a cube $\mathcal{C}_x$ of edge length 2 and a universal constant $\beta > 0$, such that $x \in \mathcal{C}_x$ and $\mathcal{C}_x$ contains at least $\beta > 0$ amount of area, where β is a universal constant.

Proof. From Lemma 5.3 there exists $y \in \partial E \cap Y$ such that $d(y, I) \geq \frac{\alpha}{2}$, where I is the exclusion contained in Y. If we make a dyadic decomposition of Y, we get 2^n cubes of side $\frac{1}{2}$ contained in Y. The point y must be contained in one of these dyadic cubes, say $\tilde{Y}$. Both Y and $\tilde{Y}$ have a common vertex, say v. We denote $\mathcal{C}_x$ as the cube of edge length 2 with its center in v. We note that $B(y, \frac{\alpha}{4})$ satisfies the hypothesis of Lemma 5.3, and thus we obtain the existence of the required constant β (in fact, $\beta = C(\frac{\alpha}{4})^n$). This completes the proof of the lemma. □

We shall use Vitali's covering lemma (see [25, Chapter 1]).

LEMMA 5.5. *Let $\mathcal{F}$ be any collection of nondegenerate closed cubes in $\mathbb{R}^n$ with edges parallel to the coordinate axis and satisfying*

$$\sup\{\text{diagonal} \quad \mathcal{C} : \mathcal{C} \in \mathcal{F}\} < \infty.$$

Then there exists a countable family $\mathcal{G}$ of disjoints cubes in $\mathcal{F}$ such that

$$\bigcup_{\mathcal{C} \in \mathcal{F}} \mathcal{C} \subset \bigcup_{\mathcal{C} \in \mathcal{G}} \hat{\mathcal{C}},$$

where $\hat{\mathcal{C}}$ is concentric with $\mathcal{C}$, and with edge length five times the edge length of $\mathcal{C}$.

Proof. The proof is the same as with balls, using the fact that the cubes are oriented in the same way as the coordinate axis. □

We have the following.

Remark 5.1. If we have a cube $\mathcal{C}$ in $\mathbb{R}^n$ of edge length l, then we can have at most $3^n - 1$ cubes of edge length l that intersect $\mathcal{C}$ without intersecting among themselves in a set of positive measure.

We now prove the following.

PROPOSITION 5.1. *There exists a universal constant $\tilde{M}$ such that for all $M \geq \tilde{M}$, if E denotes the infimal minimizer corresponding to A_{S_1,S_2}, where $S_1 = \Gamma_{\omega,0}$ and $S_2 = \Gamma_{\omega,M}$, then $d(\Pi_1, \partial E) < \tilde{M}$.*

Proof. We define $\tau = 5$, and we fix λ to be a multiple of 2τ and satisfying

$$\lambda > \frac{2^{2n} n \tau^n (3^n - 1)}{\beta}. \tag{23}$$

We let $\tilde{M} = 2\lambda\sqrt{n}$, and we note that $2\lambda\sqrt{n}$ is the length of the diagonal of the cube of edge length 2λ. We fix $M \geq \tilde{M}$ and denote E as the infimal minimizer corresponding to the class A_{S_1,S_2}, where $S_1 = \Gamma_{\omega,0}$ and $S_2 = \Gamma_{\omega,M}$. Our choice of λ allows us to fit a cube $\tilde{\mathcal{C}}$ of edge length 2λ in between $\Pi_1 = \{x \in \mathbb{R}^n : x \cdot \omega = 0\}$ and $\Pi_2 = \{x \in \mathbb{R}^n : x \cdot \frac{\omega}{|\omega|} = M\}$, with $\tilde{\mathcal{C}}$ having integer vertices, and edges parallel to the coordinate axis and intersecting Π_1 in a line. We claim that

$$d(\partial E, \Pi_1) < \tilde{M}. \tag{24}$$

We let $\mathcal{C}$ be the cube of edge length λ that is concentric with the cube $\tilde{\mathcal{C}}$. One of the following must happen:

1. $\mathcal{C} \subset \mathbb{R}^n \backslash E$. In this case, Lemma 4.2(a) implies the inequality (24).

2. $\mathcal{C} \cap E \neq \emptyset$. In this case, due to Proposition 4.1, we cannot have $\mathcal{C} \subset E$. Therefore, $\mathcal{C}$ must intersect ∂E.

For each $x \in \partial E \cap O \cap \mathcal{C}$ we denote $\mathcal{C}_x$ as the cube of edge length 2 constructed in Lemma 5.4. Therefore, we have a cover $\{\cup \mathcal{C}_x\}$ for $\partial E \cap \mathcal{C} \cap O$. By Lemma 5.5 we can extract a countable disjoint family $\{\mathcal{C}_i\}$ such that

$$\bigcup \mathcal{C}_x \subset \bigcup \hat{\mathcal{C}}_i, \tag{25}$$

where $\hat{\mathcal{C}}_i$ is concentric with $\mathcal{C}_i$ and has edge length 2τ. From Lemma 5.1 we have

$$\int_{(\cup \mathcal{C}_i) \cap O} |D\varphi_E| \leq 2n(2\lambda)^{n-1}. \tag{26}$$

From (26) and Lemma 5.4 it follows that the disjoint family has a finite number of cubes, say K, given by

$$K \leq \frac{2^n n \lambda^{n-1}}{\beta}. \tag{27}$$

Since λ is a multiple of 2τ, we can divide $\mathcal{C}$ in $\frac{\lambda^n}{(2\tau)^n}$ cubes of edge length 2τ, each cube having integer vertices and edges parallel to the coordinate axis. We note that the cubes do not intersect in sets of positive measure. Let us refer to this collection of cubes as $\mathcal{B}$. By Remark 5.1, out of the collection $\mathcal{B}$, at most

$$\frac{(3^n - 1)2^n n \lambda^{n-1}}{\beta} \tag{28}$$

intersect ∂E. Due to our choice of λ, we have

$$\frac{(3^n - 1)2^n n \lambda^{n-1}}{\beta} < \frac{\lambda^n}{(2\tau)^n}.$$

This implies that there exists $\mathcal{C}' \in \mathcal{B}$ such that $\mathcal{C}' \cap \partial E = \emptyset$. Due to Proposition 4.1 we must have $\mathcal{C}' \subset \mathbb{R}^n \backslash E$, and the inequality (24) follows from Lemma 4.2(a). ∎

We have the following.

PROPOSITION 5.2. *If E denotes the infimal minimizer corresponding to A_{S_1,S_2}, where $S_1 = \Gamma_{\omega,0}$ and $S_2 = \Gamma_{\omega,2\lambda\sqrt{n}}$, then E is an unconstrained minimizer.*

Proof. From inequality (24) we have that, for all $M > \tilde{M} = 2\lambda\sqrt{n}$, E is a minimizer for the class $A_{\Gamma_{\omega,0},\Gamma_{\omega,M}}$. We fix $\gamma > 0$. We claim that E is a minimizer for the class $A_{\Gamma_{\omega,-\gamma},\Gamma_{\omega,\tilde{M}}}$. We proceed by contradiction and assume this is not true. Therefore, the infimal minimizer, say $\tilde{E}$, corresponding to the class $A_{\Gamma_{\omega,-\gamma},\Gamma_{\omega,\tilde{M}}}$ has less perimeter than E. We choose $k \in \mathbb{Z}^n$ in such a way that $\Gamma_{\omega,0} \subset T_k\tilde{E}$. We obtain a contradiction since $T_k\tilde{E}$ is contained in the class $A_{\Gamma_{\omega,0},\Gamma_{\omega,\tilde{M}+k\cdot\frac{w}{|w|}}}$ and has less perimeter than E, which is a minimizer for this class. ∎

PROPOSITION 5.3. *If E denotes the infimal minimizer corresponding to A_{S_1,S_2}, where $S_1 = \Gamma_{\omega,0}$ and $S_2 = \Gamma_{\omega,2\lambda\sqrt{n}}$, then E is a class A minimizer.*

Proof. We let L denote any set that coincides with E outside the ball B_{R-1}. We consider E as competing in a class with a period and distance between the plane restrictions large enough so that B_{R-1} is completely contained in one period. In order to do this, we choose $M > 0$ and N in (4) large enough in such a way that B_{R-1}

is contained in the period $[-M,M]\times\mathbb{T}^{n-1}$ corresponding to the class $A_{\Gamma_{\omega,-M},\Gamma_{\omega,M}}$. Using Proposition 5.2 and Remark 4.1 it follows that E is a minimizer in this new class, and therefore $\mathrm{Per}(E,B_R\cap O)\le \mathrm{Per}(L,B_R\cap O)$. Since R is arbitrary, the proposition follows. □

This completes the proof of Theorem 5.1 for the case ω rational.

5.1. The case ω irrational. We now proceed to consider the case when the slope ω of the plane is irrational. Given $\omega\in\mathbb{R}^n\backslash\mathbb{Q}^n$, there exists a sequence $\{\omega_j\}\in\mathbb{Q}^n$ with $\omega_j\to\omega$. For each ω_j, we let $\{E_{\omega_j}\}$ denote the corresponding class A minimizers given by Theorem 5.1. From Lemma 5.1 we have

$$\mathrm{Per}(E_{\omega_j},B_R\cap O)\le CR^{n-1}.$$

Thus, $\{E_{\omega_j}\}$ has a subsequence that is convergent in $L^1(B_R\cap O)$. By applying the diagonal procedure, we obtain a subsequence of $\{E_{\omega_j}\}$ (which we will denote again as $\{E_{\omega_j}\}$) and a set E_ω such that $E_{\omega_j}\to E_\omega$ in $L^1_{loc}(\mathbb{R}^n\cap O)$. We need to check that E_ω is a class A minimizer. We let L denote any set that coincides with E_w outside the ball B_{R-1}. We define, for each j and $R\le r\le R+1$,

$$F_j^r=\begin{cases} L & \text{in} \quad B_r,\\ E_j & \text{in} \quad B_{R+1}\setminus\overline{B}_r.\end{cases}$$

Since each E_j is a class A minimizer we have

$$\begin{aligned}\int_{B_{R+1}\cap O}|D\varphi_{E_j}| &\le \int_{B_{R+1}\cap O}|D\varphi_{F_j^r}|\\ &=\int_{B_r\cap O}|D\varphi_L|+\int_{\partial B_r\cap O}|D\varphi_{F_j^r}|+\int_{(B_{R+1}\setminus\overline{B}_r)\cap O}|D\varphi_{E_j}|\\ &=\int_{B_r\cap O}|D\varphi_L|+\int_{\partial B_r\cap O}|(\varphi_L)^r_{tr}-(\varphi_{E_j})^r_{tr}|d\mathcal{H}_{n-1}\\ &\quad+\int_{(B_{R+1}\setminus\overline{B}_r)\cap O}|D\varphi_{E_j}|,\end{aligned}$$

where $(\varphi_L)^r_{tr}$ and $(\varphi_{E_j})^r_{tr}$ are the traces (see Theorem A.3) of φ_L and φ_{E_j} on ∂B_r, respectively. We recall that, for almost every $R\le r\le R+1$, the traces $(\varphi_L)^r_{tr}$ and $(\varphi_{E_j})^r_{tr}$ coincide with the corresponding characteristic functions (see [27]). Using this fact and passing the last term in the right-hand side of the previous inequality to the left we obtain, for almost every $R\le r\le R+1$,

$$\int_{B_r\cap O}|D\varphi_{E_j}|\le\int_{B_r\cap O}|D\varphi_L|+\int_{\partial B_r\cap O}|\varphi_L-\varphi_{E_j}|d\mathcal{H}_{n-1}. \tag{29}$$

We have the identity

$$\int_{(B_{R+1}\setminus\overline{B}_R)\cap O}|\varphi_{E_j}-\varphi_L|=\int_R^{R+1}\int_{\partial B_r\cap O}|\varphi_{E_j}-\varphi_L|d\mathcal{H}_{n-1}dr. \tag{30}$$

Since $E_w=L$ in $B_{R+1}\setminus\overline{B_R}$ it follows that $E_j\to L$ in $L^1((B_{R+1}\setminus\overline{B}_R)\cap O)$. This implies that (30) converges to 0 as $j\to\infty$, and therefore there exists a subsequence of $\{E_j\}$ (that we shall denote again as E_j) such that, for almost every $R\le r\le R+1$,

$$\int_{\partial B_r\cap O}|\varphi_{E_j}-\varphi_L|d\mathcal{H}_{n-1}\to 0. \tag{31}$$

From (29) and (31), it follows that, for almost every $R \le r \le R+1$,

$$\limsup_{r\to o} \int_{B_r\cap O} |D\varphi_{E_j}| \le \int_{B_r\cap O} |D\varphi_L|, \tag{32}$$

and hence

$$\begin{aligned}\int_{B_r\cap O} |D\varphi_{E_w}| &\le \liminf \int_{B_r\cap O} |D\varphi_{E_j}| \\ &\le \int_{B_r\cap O} |D\varphi_L|.\end{aligned}$$

Since $L = E_w$ in $B_{R+1} \setminus \overline{B}_R$ we conclude that

$$\int_{B_R\cap O} |D\varphi_{E_w}| \le \int_{B_R\cap O} |D\varphi_L|, \tag{33}$$

which proves that E_w is a class A minimizer. Clearly, we also have $d(E_w, \Pi_1) \le 2\lambda\sqrt{n}$. □

6. Behavior of the minimizers near the boundaries of the exclusions. It is an easy exercise to check that, for $n = 2$, minimizers must enter the exclusions *orthogonally*. In higher dimensions, the analogous result can be deduced (once we have the regularity of the minimizers up to the boundary of the exclusions) by studying the first variation of the area. An analysis of the Euler–Lagrange equation is done in [31], and we explain in this section how to use the results in [31] to obtain the fact that the minimizers must enter the exclusions *orthogonally*. For a proof of this orthogonality property, using techniques of geometric measure theory, we refer the reader to [29]. When the exclusions have C^2 boundary, the regularity of the minimizers near the boundaries of the exclusions is proven in [29].

In order to show how the orthogonality result follows from the work in [31], we must first recall that the minimal surface problem can also be studied by considering the surfaces as graphs of functions (nonparametric approach; cf. [27]). We can think of the nonparametric minimal surface problem as the problem of minimizing the energy among a class of functions with fixed boundary data and where the density at each point is one. In [31], the nonparametric minimal surface problem involving two different media is considered (the density at each point is given by a positive, piecewise smooth function), and the Euler–Lagrange equation is derived from the variational form. The solution has a jump across the interface that separates the two media, and a jump condition is derived that generalizes Snell's law to higher dimensions.

For the case $n = 2$, following [31] we consider a two-dimensional domain $D = [a, b] \times [c, d]$, and we seek a function $u(x, y)$ which minimizes the functional

$$\begin{aligned} E(u) &= \int_D c(x, y, u(x, y))\sqrt{1 + |Du(x, y)|^2}\,dx\,dy, \\ u(x, y)|_{\partial D} &= u_0(x, y), \end{aligned} \tag{34}$$

where $u_0(x, y)$ is a given boundary condition and $c(x, y, z)$ is a positive piecewise smooth function which has a finite jump across a surface $S = \{(x, y, z) : g(x, y, z) = 0\}$. We assume that the graph of the minimizer of (34) intersects the surface S at a curve Γ. We denote γ as the projection of Γ on the (x, y)-plane. The curve γ divides the set D in two regions, D_1 and D_2. It is proven in [31] that if the surface S can

be expressed locally as the graph of the function $z = \phi(x, y)$, then the jump of the derivatives of $u(x, y)$ across the surface S must satisfy the following generalized Snell's law in three dimensions (which can be extended to higher dimensions):

$$c^- \mathbf{n_1} \cdot \mathbf{m}|_\Gamma = c^+ \mathbf{n_2} \cdot \mathbf{m}|_\Gamma, \tag{35}$$

where c^- and c^+ are the weights of the two different media, $\mathbf{n_i} = \frac{(-u_x, -u_y, 1)}{\sqrt{1+u_x^2+u_y^2}}$ are the normal directions of the surface $u(x, y)$ in D_i, $i = 1, 2$, and $\mathbf{m} = \frac{(-\phi_x, -\phi_y, 1)}{\sqrt{1+\phi_x^2+\phi_y^2}}$ is the unit normal direction of S.

We note that if we consider the case $c^- = \epsilon$, $c^+ = 1$ and then compute the limit in (35) as $\epsilon \to 0$ we obtain

$$\mathbf{n_2} \cdot \mathbf{m}|_\Gamma = 0, \tag{36}$$

which implies that $\mathbf{n_2}$ and $\mathbf{m}$ are orthogonal vectors. We conclude from this that the minimizer E meets (on its regular points) the boundary of the exclusions orthogonally.

7. Connection with homogenization of Hamilton–Jacobi equations. In this section we explore some connections with the theory of homogenization of Hamilton–Jacobi equations. We first recall some of the main issues concerning the homogenization of Hamilton–Jacobi equations, and then we present the connection with the degenerate metric considered earlier.

We consider, for each $0 < \epsilon \leq 1$, the viscosity solution u^ϵ of the following problem:

$$H\left(Du^\epsilon, \frac{x}{\epsilon}\right) = 0 \text{ in } \mathbb{R}^n, \tag{37}$$

where $H : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a periodic function in the second variable. Under a suitable hypothesis (see, for instance, [22]) we can homogenize (37); i.e., the sequence of viscosity solutions $\{u^\epsilon\}$ converges as $\epsilon \to 0$ to the viscosity solution u of the averaged problem

$$\overline{H}(Du) = 0 \text{ in } \mathbb{R}^n,$$

where $\overline{H} : \mathbb{R}^n \to \mathbb{R}$ is defined as follows: for each $p \in \mathbb{R}^n$, $\overline{H}(p)$ is the unique number for which the PDE

$$\begin{aligned} H(p + D_y v, y) &= \overline{H}(p) \text{ in } \mathbb{R}^n, \\ v &\text{ is } [0, 1]^n\text{-periodic} \end{aligned} \tag{38}$$

has a viscosity solution.

As explained in the introduction, the function $\overline{H}$ is called the effective Hamiltonian, and an interesting endeavor is to study the structure of $\overline{H}$ in order to find explicit formulas for it. This is still largely an open problem, and [22, 19, 20, 24, 17, 16] contain results in this direction. The goal of this section is to provide a particular example of (37) for which we can explicitly compute the limiting function u.

We recall our earlier consideration of $\mathbb{R}^n$ as the lattice of cubes $[0, 1]^n + \mathbb{Z}^n$ where each cube of side 1 has an internal exclusion. The exclusions satisfy properties 1, 2, and 3 stated in the introduction. In this section we work with surfaces of codimension $(n - 1)$, i.e., curves, instead of surfaces of codimension 1.

We fix $x \in \mathbb{R}^n$, and for each $0 < \epsilon \leq 1$ we consider the sequence of lattices $\epsilon([0, 1]^n + \mathbb{Z}^n)$. We let $\mathcal{J}$ denote the set of all curves joining the origin with x. We

use the degenerate metric introduced in this paper to measure the length of each curve $l \in \mathcal{J}$. The length of l at the scale ϵ, that is, when we consider l as residing in the configuration $\epsilon([0,1]^n + \mathbb{Z}^n)$, is obtained by neglecting the portions inside the exclusions. This length depends on ϵ since the configuration of the lattice changes as we let $\epsilon \to 0$. We let l_ϵ denote the curve of minimal length and denote this optimal length by $d_\epsilon(0,x)$. We shall refer to the number $d_\epsilon(0,x)$ as the *smallest distance between* 0 *and* x *at the scale* ϵ.

We define, for each $0 < \epsilon \leq 1$, the sequence of functions

$$u^\epsilon(x) = d_\epsilon(0,x), \quad x \in \mathbb{R}^n. \tag{39}$$

We have the following.

THEOREM 7.1. *If* $\mathcal{I}$ *denotes the union of all exclusions and* $O = \mathbb{R}^n \backslash \mathcal{I}$, *then*

$$\begin{cases} |Du_\epsilon| = 1 & \text{in} \quad \epsilon O, \\ u_\epsilon & \text{is constant on each connected component of} \quad \epsilon\mathcal{I}. \end{cases} \tag{40}$$

Proof. Without loss of generality we can assume $\epsilon = 1$. We define

$$v(x) = d_1(x,0), \quad x \in \mathbb{R}^n. \tag{41}$$

$v(x)$ is the smallest distance from x to the origin, and since we compute the length of a path $l \in \mathcal{J}$ by neglecting the portions inside the exclusions, we have that v is constant on each exclusion, which is connected. We prove now that v solves the eikonal equation $|Dv| = 1$ in the viscosity sense outside the exclusions. We prove first that v is a viscosity subsolution of $|Du| = 1$. If φ is a C^1 function such that $v - \varphi$ has a local maximum at the point $x_0 \in O$, we need to prove that $|D\varphi(x_0)| \leq 1$. Since $v - \varphi$ has a local maximum at x_0 it follows that $v(x) - v(x_0) \leq \varphi(x) - \varphi(x_0)$ for all x in a neighborhood of x_0. Therefore, for all z satisfying $|z| = 1$ and for all h small enough, we have

$$\begin{aligned} v(x_0 + hz) - v(x_0) \leq \varphi(x_0 + hz) - \varphi(x_0) &= \int_0^h \frac{d}{ds}\varphi(x_0 + sz) \\ &= \int_0^h D\varphi(x_0 + sz) \cdot z ds \leq \int_0^h D\varphi(x_0) \cdot z ds + Ch^2. \end{aligned}$$

If we define $z_0 = -\frac{D\varphi(x_0)}{|D\varphi(x_0)|}$, then

$$v(x_0 + hz_0) - v(x_0) \leq -\int_0^h |D\varphi(x_0)| ds + Ch^2 = -h|D\varphi(x_0)| + Ch^2. \tag{42}$$

We now use the fact that v is a Lipschitz function, and from (42) we obtain

$$h|D\varphi(x_0)| \leq v(x_0) - v(x_0 + hz_0) + Ch^2 \leq |hz_0| + Ch^2,$$

and hence

$$|D\varphi(x_0)| \leq 1 + Ch.$$

By letting $h \to 0$, we conclude that $|D\varphi(x_0)| \leq 1$. We now prove that v is a supersolution. If φ is a C^1 function such that $v - \varphi$ has a local minimum at the point $x_0 \in O$,

we need to prove that $|D\varphi(x_0)| \geq 1$. Since $v-\varphi$ has a local minimum at x_0, it follows that $v(x)-v(x_0) \geq \varphi(x)-\varphi(x_0)$ for all x in a neighborhood of x_0. Therefore, if h is small enough, we have

$$v(x_0+hz)-v(x_0) \geq \varphi(x_0+hz)-\varphi(x_0) = \int_0^h \frac{d}{ds}\varphi(x_0+sz) = \int_0^h D\varphi(x_0+sz)\cdot z ds$$

$$\geq \int_0^h D\varphi(x_0)\cdot z ds - Ch^2 \geq -h|D\varphi(x_0)| - Ch^2 \tag{43}$$

for all $|z|=1$. We fix h small enough. We note that $v(x_0)=\inf_{|z|=1}\{h+v(x_0+hz)\}$, and hence there exists a point z_0 such that $v(x_0+hz_0)+h \leq v(x_0)+h^2$. From (43) we obtain $h|D\varphi(x_0)| \geq v(x_0)-v(x_0+hz_0)-Ch^2 \geq h-h^2-Ch^2$, and hence $|D\varphi(x_0)| \geq 1-h-Ch$. Letting $h \to 0$ we obtain $|D\varphi(x_0)| \geq 1$. □

From standard theory of viscosity solutions we have that $\{u_\epsilon\}$ contains a subsequence that converges uniformly to a function u_0. Constructing the PDE that u_0 solves (i.e., the homogenization of (40)) is difficult. We present in this section some partial results toward this homogenization.

We proceed to compute u_0 *for the particular case* $n=2$ and we assume, in addition to properties 1, 2, and 3 given in the introduction, that the exclusions are *balls of radius* ρ. Given two fixed points P and Q in the plane, we let $l_\epsilon^\rho(P,Q)$ denote the optimal path joining P and Q at the scale ϵ. We denote $d_\epsilon^\rho(P,Q)$ as the length of $l_\epsilon^\rho(P,Q)$. The behavior of $d_\epsilon^\rho(P,Q)$ depends on the value of ρ, where $0<\rho\leq\frac{1}{2}$ (we note that the radius of the exclusions at the scale ϵ is $\epsilon\rho$). We have, for any $0<\epsilon\leq 1$,

$$0 \leq d_\epsilon^\rho(P,Q) \leq |P-Q|_2. \tag{44}$$

Thus, fixing ρ and letting $\epsilon \to 0$ it follows that $\{d_\epsilon^\rho(P,Q)\}$ contains a subsequence that converges to a number, say $d_0^\rho(P,Q)$.

If we assume that X and Y are centers of exclusions at the scale ϵ, we can replace $l_\epsilon^\rho(X,Y)$ inside the exclusions with lines so that this optimal path is composed of a sequence of segments. We can classify (after a suitable translation and/or rotation) these segments in the following four categories:

1. a segment joining the points $(0,0)$ and $(\frac{i}{\epsilon},\frac{j}{\epsilon})$, where $i,j\in\mathbb{Z}^+$ are relatively prime and $j<i$;
2. the segment joining the points $(0,0)$ and $(\frac{1}{\epsilon},0)$;
3. the segment joining the points $(0,0)$ and $(0,\frac{1}{\epsilon})$;
4. the segment joining $(0,0)$ and $(\frac{1}{\epsilon},\frac{1}{\epsilon})$.

We identify a segment of type 1 with the pair $[i,j]$, a segment of types 2 or 3 with $[1,0]$, and a segment of type 4 with $[1,1]$. Therefore, any optimal path joining two points that are centers of exclusions is composed of a sequence of segments belonging to the set

$$\mathcal{P} = \{[i,j] : i,j\in\mathbb{Z}^+, i,j \quad \text{are relatively prime, } j<i\} \cup \{[1,1]\} \cup \{[1,0]\}.$$

We prove in the next theorem that if ρ is large enough, then the optimal path joining two centers of exclusions is composed only of segments of the type $[1,0]$.

THEOREM 7.2. *If* $\rho > \frac{2-\sqrt{2}}{2}$ *then, for any* $0<\epsilon\leq 1$, *the optimal path connecting two points that are centers of exclusions is composed only of segments of the type* $[1,0]$. *Moreover, if* P *and* Q *are any two points, we have that*

$$\lim_{\epsilon\to 0} d_\epsilon^\rho(P,Q) = (1-2\rho)|P-Q|_1.$$

Proof. We fix $\epsilon > 0$, and thus in this proof we work in the domain $\epsilon([0,1]^n + \mathbb{Z}^n)$. We denote $X = (x_1, x_2)$ and $Y = (y_1, y_2)$ as two points that are centers of exclusions. We can assume, without loss of generality, that $y_1 \geq x_1$ and $y_2 \geq x_2$. We proceed by contradiction and assume that $l_\epsilon^\rho(X,Y)$ has a segment of the type 1 or 4. Therefore, the path $l_\epsilon^\rho(X,Y)$ contains a segment joining the points $\epsilon(i+\frac{1}{2}, j+\frac{1}{2})$ and $\epsilon(i+\frac{1}{2}+m, j+\frac{1}{2}+n)$, where $n \leq m$, $n \geq 2$, and m and n are prime relative to each other. Since $l_\epsilon^\rho(X,Y)$ is the optimal path we have that

$$\begin{aligned} m(\epsilon - 2\epsilon\rho) + n(\epsilon - 2\epsilon\rho) &\geq \sqrt{\epsilon^2 m^2 + \epsilon^2 n^2} - 2\epsilon\rho, \\ m(1-2\rho) + n(1-2\rho) &\geq \sqrt{m^2+n^2} - 2\rho, \\ m + n - \sqrt{m^2+n^2} &\geq 2(m+n-1)\rho, \\ \Rightarrow \rho &\leq \frac{m+n-\sqrt{m^2+n^2}}{2(m+n-1)}. \end{aligned}$$

We claim that $\frac{m+n-\sqrt{m^2+n^2}}{2(m+n-1)} \leq \frac{2-\sqrt{2}}{2}$. To prove this, we consider the function $f(x) = \frac{x+n-\sqrt{x^2+n^2}}{2(x+n-1)}$ and its derivative $f'(x) = \frac{1}{2}\frac{n^2-\sqrt{x^2+n^2}+x-xn}{\sqrt{n^2+x^2}(x+n+1)^2}$. We note that $f'(x) \leq 0$ if $x \geq 0$. This implies that f is decreasing, and thus $f(m) \leq f(n)$. By a simple substitution it follows that $f(n) = \frac{2n-\sqrt{2}n}{2(2n-1)} = \frac{2-\sqrt{2}}{2}(\frac{n}{2n-1}) \leq \frac{2-\sqrt{2}}{2}$ (since $\frac{n}{2n-1} \leq 1$). Hence, $\rho \leq f(m) \leq \frac{2-\sqrt{2}}{2}$, which contradicts the fact that $\rho > \frac{2-\sqrt{2}}{2}$. This proves the first part of the theorem. Because of the above result, we can explicitly compute $d_\epsilon^\rho(X,Y)$:

$$\begin{aligned} d_\epsilon^\rho(X,Y) &= \frac{y_2 - x_2}{\epsilon}(\epsilon - 2\epsilon\rho) + \frac{y_1 - x_1}{\epsilon}(\epsilon - 2\epsilon\rho) \\ &= (1-2\rho)[(y_2 - x_2) + (y_1 - x_1)] \\ &= (1-2\rho)|Y - X|_1. \end{aligned} \tag{45}$$

We now denote P and Q as any two points in the plane. If P' and Q' are the closest centers of exclusions to P and Q, respectively, we have

$$d_\epsilon^\rho(P',Q') - \sqrt{2}\epsilon \leq d_\epsilon^\rho(P,Q) \leq d_\epsilon^\rho(P',Q') + \sqrt{2}\epsilon.$$

From (45) we have

$$(1-2\rho)|P'-Q'|_1 - \sqrt{2}\epsilon \leq d_\epsilon^\rho(P,Q) \leq (1-2\rho)|P'-Q'|_1 + \sqrt{2}\epsilon.$$

Using the triangle inequality, again

$$\begin{aligned} (1-2\rho)(|P-Q|_1 - 2\sqrt{2}\epsilon) - \sqrt{2}\epsilon \leq d_\epsilon^\rho(P,Q) \\ \leq (1-2\rho)(|P-Q|_1 + 2\sqrt{2}\epsilon) + \sqrt{2}\epsilon. \end{aligned}$$

Letting $\epsilon \to 0$ yields

$$\begin{aligned} 1 - 2\rho \leq \frac{d_0^\rho(P,Q)}{|P-Q|_1} &\leq 1 - 2\rho \\ \Rightarrow \\ d_0^\rho(P,Q) &= (1-2\rho)|P-Q|_1. \end{aligned}$$ □

We now wish to study the behavior of the optimal path as $\rho \to 0$. As ρ decreases, new paths (new segments of the collection $\mathcal{P}$) become available. For each

segment $[i,j]$ there exists a critical radius $\rho_{[i,j]}$, which is the largest radius for which $d_1^{\rho_{[i,j]}}((0,0),(i,j)) = \sqrt{i^2+j^2}$.

Since $\mathcal{P}$ is countable, we can enumerate the sequence $\{\rho_{[i,j]}\}$ in such a way that the coordinate i is always increasing. We have the following lemma.

LEMMA 7.1. $\lim_{i\to\infty} \rho_{[i,j]} = 0$.

Proof. We recall that $[i,j]$ represents the segment joining $(0,0)$ with (i,j). We denote by $P = (p_1,p_2)$ the closest point (other than the extremes) with integer coordinates to the segment, and we denote this distance as d. The point P satisfies the equation $|\frac{i}{j} - \frac{p_1}{p_2}| = \frac{1}{jp_2}$, that is, $|ip_2 - jp_1| = 1 = A(p)$, where $A(p)$ is the area of the parallelogram spanned by (i,j) and (p_1,p_2). Hence we have that $\frac{1}{2} = d\frac{\sqrt{i^2+j^2}}{2}$, which implies that $d = \frac{1}{\sqrt{i^2+j^2}}$. We define $l = \sqrt{i^2+j^2}$, $l_1 = \sqrt{p_1^2+p_2^2}$, and $l_2 = \sqrt{(i-p_1)^2+(j-p_2)^2}$. Solving the equation $l - 2\rho = l_1 + l_2 - 4\rho$ for ρ, we obtain the critical radius for which the segment joining $(0,0)$ with (i,j) is a better path than the one joining the points $(0,0),(p_1,p_2)$, and (i,j). We have that $\rho = \frac{l_1+l_2-l}{2}$. Since $l_1 + l_2 \le 2d + l$ it follows that $\rho \le \frac{2d+l-l}{2} = d = \frac{1}{\sqrt{i^2+j^2}}$. Since $\rho_{[i,j]} \le \rho$ the lemma holds. ∎

An easy computation gives us $\rho_{[1,1]} = \frac{2-\sqrt{2}}{2}$, $\rho_{[2,1]} = \frac{1+\sqrt{2}-\sqrt{5}}{2}$, and $\rho_{[3,1]} = \frac{1+\sqrt{5}-\sqrt{10}}{2}$. We have the following theorem.

THEOREM 7.3. *Let P, Q be any two points in the plane. Then we have the following:*

(a) *If $\frac{1+\sqrt{2}-\sqrt{5}}{2} < \rho \le \frac{2-\sqrt{2}}{2}$, we have*

$$\lim_{\epsilon\to 0} d_\epsilon^\rho(P,Q) = |P-Q|_{1,1},$$

where $|\cdot|_{1,1} : \mathbb{R}^2 \to \mathbb{R}^+$ is given by

$|(x,y)|_{1,1} = (\sqrt{2}-1)|x| + (1-2\rho)|y|$ *if* $|x| \le |y|$ *and*

$|(x,y)|_{1,1} = (\sqrt{2}-1)|y| + (1-2\rho)|x|$ *if* $|y| \le |x|$.

(b) *If $\frac{1+\sqrt{5}-\sqrt{10}}{2} < \rho \le \frac{1+\sqrt{2}-\sqrt{5}}{2}$, we have*

$$\lim_{\epsilon\to 0} d_\epsilon^\rho(P,Q) = |P-Q|_{2,1},$$

where $|\cdot|_{2,1} : \mathbb{R}^2 \to \mathbb{R}^+$ is given by

$|(x,y)|_{2,1} = (1-2\rho)|x| + (\sqrt{5}-2+2\rho)|y|$ *if* $|y| \le \frac{|x|}{2}$,

$|(x,y)|_{2,1} = (2\sqrt{2}-\sqrt{5}-2\rho)|y| + (\sqrt{5}-\sqrt{2})|x|$ *if* $\frac{|x|}{2} < |y| \le |x|$,

$|(x,y)|_{2,1} = (1-2\rho)|y| + (\sqrt{5}-2-2\rho)|x|$ *if* $|y| \ge 2|x|$, *and*

$|(x,y)|_{2,1} = (2\sqrt{2}-\sqrt{5}+2\rho)|x| + (\sqrt{5}-\sqrt{2})|y|$ *if* $|x| \le |y| < 2|x|$.

Moreover, $|\cdot|_{1,1}$ and $|\cdot|_{2,1}$ define norms in $\mathbb{R}^2$.

Proof. We denote $X = (x_1,x_2)$ and $Y = (y_1,y_2)$ as centers of exclusions (at the scale ϵ). We can assume that $x_1 \le y_1$ and $x_2 \le y_2$. In order to prove (a), we consider first the case when $y_2 - x_2 \le y_1 - x_1$. Solving the equation $\sqrt{5} - 2\rho = \sqrt{2} + 1 - 4\rho$, we obtain $\rho = \frac{\sqrt{2}+1-\sqrt{5}}{2}$, the critical radius for which the next segment $[2,1]$ becomes available. Therefore, if ρ belongs to the interval given in (a), the only paths available are $[1,0]$ and $[1,1]$. Thus, the optimal path $l_\epsilon^\rho(X,Y)$ has as many segments $[1,1]$ as possible, since for this interval $[1,1]$ is better than two segments of type $[1,0]$. Hence,

we can compute $d_\epsilon^\rho(X,Y)$ explicitly, and we obtain

$$\begin{aligned} d_\epsilon^\rho(X,Y) &= \frac{y_2-x_2}{\epsilon}(\sqrt{2}\epsilon - 2\epsilon\rho) + \frac{(y_1-x_1)-(y_2-x_2)}{\epsilon}(\epsilon - 2\epsilon\rho) \\ &= (\sqrt{2}-1)(y_2-x_2) + (1-2\rho)(y_1-x_1) \\ &= |Y-X|_{1,1}. \end{aligned}$$

The case $y_2 - x_2 \geq y_1 - x_1$ is computed in the same way, except that we interchange the roles of the coordinates. To prove (a) we can proceed now in exactly the same way (provided that $|\cdot|_{1,1}$ is a norm) as in Theorem 7.2. We need to check that $|\cdot|_{1,1}$ defines a norm in $\mathbb{R}^2$. We need only to show that the triangle inequality holds, and there are several cases to verify.

We let $(x,y),(w,z)$ be any two points in the plane, and we consider the case $|y| \leq |x|$, $|w| \leq |z|$ and $|x+w| \leq |y+z|$. We need to prove that $(\sqrt{2}-1)|x+w|+(1-2\rho)|y+z| \leq (\sqrt{2}-1)(|y|+|w|)+(1-2\rho)(|x|+|z|)$; that is, $(\sqrt{2}-1)(|x+w|-|y|-|w|)+(1-2\rho)(|y+z|-|x|-|z|) \leq 0$. Using the triangle inequality for real numbers we can see that the last inequality is true since $|x|-|y| \geq 0$ and $1-2\rho > \sqrt{2}-1$ for ρ in the interval given in (a).

Considering now the case $|y| \leq |x|$, $|w| \geq |z|$, and $|x+w| \leq |y+z|$, we need to prove that $(\sqrt{2}-1)|x+w|+(1-2\rho)|y+z| \leq (\sqrt{2}-1)(|y|+|z|)+(1-2\rho)(|x|+|w|)$; that is, $(\sqrt{2}-1)(|x+w|-|y|-|z|)+(1-2\rho)(|y+z|-|x|-|w|) \leq 0$. Using the triangle inequality for real numbers we can see that the last inequality is true.

Proceeding to the case $|y| \leq |x|$, $|w| \leq |z|$, and $|x+w| \geq |y+z|$, we need to prove that $(\sqrt{2}-1)|y+z|+(1-2\rho)|x+w| \leq (\sqrt{2}-1)(|y|+|w|)+(1-2\rho)(|x|+|z|)$; that is, $(\sqrt{2}-1)(|y+z|-|y|-|w|)+(1-2\rho)(|x+w|-|x|-|z|) \leq 0$. Using the triangle inequality for real numbers we can see that the last inequality is true since $|z|-|w| \geq 0$ and $1-2\rho > \sqrt{2}-1$ for ρ in the interval given in (a).

Finally, we check that $|y| \leq |x|$, $|w| \geq |z|$, and $|x+w| \geq |y+z|$. We need to prove that $(\sqrt{2}-1)|y+z|+(1-2\rho)|x+w| \leq (\sqrt{2}-1)(|y|+|z|)+(1-2\rho)(|x|+|w|)$; that is, $(\sqrt{2}-1)(|y+z|-|y|-|z|)+(1-2\rho)(|x+w|-|x|-|w|) \leq 0$, which is true due to the triangle inequality. There are four more cases corresponding to $|y| \geq |x|$, but they are proven in the same way. The unit ball for this norm is a polygon with eight edges as shown in Figure 2.

To prove (b) we note that by solving the equation $\sqrt{10}-2\rho = \sqrt{5}+1-4\rho$ we obtain $\rho = \frac{\sqrt{5}+1-\sqrt{10}}{2}$, the critical radius for which the next segment $[3,1]$ becomes available. If $p > 0$ and $q \geq 0$ are two integers satisfying $-p+2q \leq 0$, then, for ρ in the interval given in (b), the best path joining $(0,0)$ with (p,q) consists only of segments of the type $[2,1]$ and $[1,0]$. Furthermore, this path takes as many $[2,1]$ segments as possible and then completes the trajectory with segments $[1,0]$. If $-p+2q > 0$ and $q < p$, the best path consists only of segments of the type $[2,1]$ and $[1,1]$, and this path takes as many $[2,1]$ segments as possible and then completes the trajectory with segments $[1,1]$. Thus, we can compute $d_\epsilon^\rho(X,Y)$ exactly as before and proceed as in Theorem 7.2. The unit ball for $|\cdot|_{2,1}$ is a polygon with 16 edges as shown in Figure 2. □

The following theorem gives an asymptotic behavior of d_0^ρ.

THEOREM 7.4. *Let P, Q be any two points in the plane. Then*

$$\lim_{\rho \to 0} d_0^\rho(P,Q) = |P-Q|_2.$$

Proof. We denote X and Y as two points that are centers of exclusions (at the scale ϵ). The optimal path $l_\epsilon^\rho(X,Y)$ intersects a finite numbers of balls, say N. We

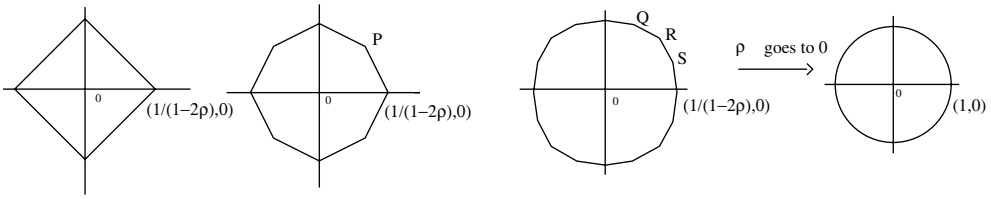

FIG. 2. *Unit balls for limiting norms.* $P = R = (\frac{1}{\sqrt{2}-2\rho}, \frac{1}{\sqrt{2}-2\rho})$, $Q = (\frac{1}{\sqrt{5}-2\rho}, \frac{2}{\sqrt{5}-2\rho})$, $S = (\frac{2}{\sqrt{5}-2\rho}, \frac{1}{\sqrt{5}-2\rho})$.

define

$$\tilde{d}^{\rho}_{\epsilon}(X,Y) = d^{\rho}_{\epsilon}(X,Y) + (N-1)(2\epsilon\rho). \tag{46}$$

Since the distance between two centers of exclusions is at least ϵ it follows that

$$\begin{aligned} N &\leq \frac{\tilde{d}^{\rho}_{\epsilon}(X,Y)}{\epsilon} + 1 \\ \Rightarrow N - 1 &\leq \frac{\tilde{d}^{\rho}_{\epsilon}(X,Y)}{\epsilon}. \end{aligned} \tag{47}$$

Hence, using (46) and (47) we obtain

$$\begin{aligned} d^{\rho}_{\epsilon}(X,Y) &\geq \tilde{d}^{\rho}_{\epsilon}(X,Y) - \frac{\tilde{d}^{\rho}_{\epsilon}(X,Y)}{\epsilon}(2\epsilon\rho) \\ &= \tilde{d}^{\rho}_{\epsilon}(X,Y)(1-2\rho) \\ &\geq (1-2\rho)|X-Y|_2. \end{aligned} \tag{48}$$

We denote P' and Q' as the closest points to P and Q, respectively (at the scale ϵ), in such a way that both P' and Q' are centers of exclusions. We have

$$d^{\rho}_{\epsilon}(P',Q') - \sqrt{2}\epsilon \leq d^{\rho}_{\epsilon}(P,Q) \leq |P-Q|_2.$$

Using (48), we obtain

$$(1-2\rho)|P'-Q'|_2 - \sqrt{2}\epsilon \leq d^{\rho}_{\epsilon}(P,Q) \leq |P-Q|_2.$$

Using the triangle inequality we have

$$(1-2\rho)(|P-Q|_2 - \sqrt{2}\epsilon) - \sqrt{2}\epsilon \leq d^{\rho}_{\epsilon}(P,Q) \leq |P-Q|_2.$$

Letting $\epsilon \to 0$ we have

$$1 - 2\rho \leq \frac{d^{\rho}_{0}(P,Q)}{|P-Q|_2} \leq 1.$$

This implies

$$\lim_{\rho \to 0} d^{\rho}_{0}(P,Q) = |P-Q|_2. \qquad \square$$

Figure 2 shows the unit balls of norms d^{ρ}_{0} for the cases $\frac{2-\sqrt{2}}{2} < \rho < 0.5$, $\frac{1+\sqrt{2}-\sqrt{5}}{2} < \rho \leq \frac{2-\sqrt{2}}{2}$, and $\frac{1+\sqrt{5}-\sqrt{10}}{2} < \rho \leq \frac{1+\sqrt{2}-\sqrt{5}}{2}$. Our results suggest that

as ρ gets smaller the behavior of the unit ball changes, though it is always polygonal with more and mores edges until it becomes a circle in the limit. That is, *as* $\rho \to 0$, *the sequence of norms converges to the Euclidean norm.*

Remark 7.1. The norms d_0^ρ can be thought of as an example of the so-called stable norms (see, for instance, [33, 28, 11, 6, 10, 5, 4] and the references therein). However, in this paper we are interested in looking at these norms in the context of Hamilton–Jacobi equations in order to provide an explicit example of homogenization of Hamilton–Jacobi equations. As mentioned earlier, finding explicit formulas for the effective Hamiltonian $\overline{H}$ is essentially still an open problem.

Remark 7.2. Theorems 7.2 and 7.3 provide, for $n = 2$, an explicit formula for u_0, which is the uniform limit of the solutions of (40). The homogenization of (40) is difficult to achieve. The construction of the corresponding effective Hamiltonian $\overline{H}$ does not follow from [30] due to the behavior of the functions u^ϵ on the boundaries of the exclusions.

Appendix A.

We refer to the standard references [27, 25, 2] for the details related to the theory of sets of finite perimeter.

DEFINITION A.1. *Throughout this paper, we denote* $B(x,r)$ *as the open ball centered at* x *and radius* r *(we shall also use the notation* B_r *when* $x = 0$*). We denote* $\mathcal{H}_k$ *as the* k*-dimensional Hausdorff measure in* $\mathbb{R}^n$*, and* $\mathcal{L}^n$ *denotes the Lebesgue measure in* $\mathbb{R}^n$*. We recall that* $\mathcal{H}_n = \mathcal{L}^n$*. At times we shall denote* $|E|$ *as the* $\mathcal{L}^n$*-Lebesgue measure of* E*.*

DEFINITION A.2 (see [27, p. 4]). *We let* $\Omega \subset \mathbb{R}^n$ *denote an open set. If* $f \in L^1(\Omega)$*, we define*

$$\int_\Omega |Df| = \sup\left\{\int_\Omega f \mathrm{div} g : g \in C_0^1(\Omega;\mathbb{R}^n), \quad |g(x)| \le 1, \quad \text{for} \quad x \in \Omega\right\}.$$

DEFINITION A.3. *A function* $f \in L^1(\Omega)$ *is said to have bounded variation in* Ω *if* $\int_\Omega |Df| < \infty$*. We define* $BV(\Omega)$ *as the space of all functions in* $L^1(\Omega)$ *with bounded variation. With the norm* $|f|_{BV} = |f|_{L^1(\Omega)} + \int_\Omega |Df|$*,* $BV(\Omega)$ *is a Banach space. If* $f \in BV(\Omega)$*, then* Df*, the gradient of* f *in the sense of distributions, is a vector valued Radon measure in* Ω *with total variation* $|Df|$*. Thus we may extend the definition of* $\int_A |Df|$ *to include cases where* $A \subset \Omega$ *is not necessarily open.*

DEFINITION A.4. *If* E *denotes a Borel set, we define the perimeter of* E *in* Ω *as*

$$\mathrm{Per}(E,\Omega) = \int_\Omega |D\varphi_E|,$$

where φ_E *is the characteristic function of the set* E*. If* $\mathrm{Per}(E,\Omega) < \infty$ *for every bounded open set* Ω*, then* E *is called a set of locally finite perimeter in* $\mathbb{R}^n$*. For simplicity, we will denote a set of locally finite perimeter in* $\mathbb{R}^n$ *simply as a* set of finite perimeter*. Also, at times we shall denote* $\mathrm{Per}(E,\mathbb{R}^n)$ *simply as* $\mathrm{Per}(E)$*.*

DEFINITION A.5 (see [27, p. 43]). *Let* E *be a set of finite perimeter. We call the* reduced boundary *of* E*, denoted as* $\partial^* E$*, the set of all points* $x \in \mathrm{supp}|D\varphi_E|$ *such that*

- $\int_{B(x,r)} |D\varphi_E| > 0$ *for all* $r > 0$*;*
- *the limit* $\nu(x) = \lim_{r\to 0} \dfrac{\int_{B(x,r)} D\varphi_E}{\int_{B(x,r)} |D\varphi_E|}$ *exists;*
- $|\nu(x)| = 1$.

DEFINITION A.6. *For every $\gamma \in [0,1]$ and every $\mathcal{L}^n$-measurable set $E \subset \mathbb{R}^n$, we define*

$$E_\gamma = \left\{ x \in \mathbb{R}^n : \lim_{r \to 0} \frac{|B(x,r) \cap E|}{|B(x,r)|} = \gamma \right\}, \tag{49}$$

the set of all points with density γ. If E is a set of finite perimeter, then (cf. [2]*) the limit in* (49) *exists for $\mathcal{H}_{n-1}$-almost every x. The sets E_1 and E_0 are the measure theoretic interior and exterior of E, respectively.*

DEFINITION A.7. *We say that the set of finite perimeter E has* least area *in the open set Ω if*

$$\int_\Omega |D\varphi_E| = \inf \left\{ \int_\Omega |D\varphi_F| : F \text{ is a set of finite perimeter, } \operatorname{support}(\varphi_F - \varphi_E) \subset \Omega \right\}.$$

DEFINITION A.8. *If E is a set of finite perimeter, we denote ∂E as the topological boundary of E. We note that $E_{int} \subset E_1$ and $E_{out} \subset E_0$, where E_{int} denotes the topological interior of the set E, and $E_{out} = (\mathbb{R}^n \setminus E)_{int}$. We define*

$$\partial^s E = \mathbb{R}^n \setminus (E_0 \cup E_1).$$

The set $\partial^s E$ is called the essential boundary *of E. We have*

$$\partial^* E \subset E_{\frac{1}{2}} \subset \partial^s E$$

and

$$\mathcal{H}_{n-1}(\partial^s E \setminus \partial^* E) = 0.$$

We have that

$$|D\varphi_E| = \mathcal{H}_{n-1}|_{\partial^* E}. \tag{50}$$

Remark A.1. When considering functions in BV we are really considering equivalence classes of functions, and changing a function on a set of measure zero gives the same function. The same is true for sets of finite perimeter, and, therefore, since we are concerned only with equivalence classes of sets, *we assume throughout this paper that a set of finite perimter E is the representative given by Theorem* A.1. With this convention, there is no ambiguity when speaking of the topological boundary of a set of finite perimeter.

Remark A.2. Standard interior regularity theory [12, 13, 26, 27, 18] implies that, if $n \leq 7$ and E is a set of finite perimeter that has least area in the open set Ω, then $\partial E \cap \Omega$ is a smooth surface. If $n > 7$, $\partial E \cap \Omega$ can have singularities, but they have zero $\mathcal{H}_k$-measure for any $k > n - 8$. At times we will use the word "surface" to denote the boundary of a set of finite perimeter, although this boundary could have singularities.

PROPOSITION A.1 (see [27, p. 7]). *If $\{f_j\}$ denote a sequence of functions in $BV(\Omega)$ that converge in $L^1_{loc}(\Omega)$ to a function f, then the following semicontinuity property holds:*

$$\int_\Omega |Df| \leq \liminf_{j \to \infty} \int_\Omega |Df_j|.$$

THEOREM A.1 (see [27, p. 42]). *If E is a Borel set, then there exists a Borel set $\tilde{E}$ equivalent to E (that is, differs only by a set of $\mathcal{L}^n$-measure zero) and such that*

$$0 < |\tilde{E} \cap B(x,r)| < \omega_n r^n$$

for all $x \in \partial\tilde{E}$ and all $r > 0$, where ω_n is the measure of the unit ball in $\mathbb{R}^n$.

THEOREM A.2 (see [27, p. 17]). *If Ω is a bounded open set in $\mathbb{R}^n$ with Lipschitz continuous boundary, then sets of functions uniformly bounded in a BV norm are relatively compact in $L^1(\Omega)$.*

Since we are regarding $BV(\Omega)$ as a subset of $L^1(\Omega)$, it makes no sense to talk about the value of a BV function on sets of measure zero. However, it is important to be able to talk about the value of a BV function on the boundary of a set even though such a boundary may have measure zero; that is, we need a notion of *trace of a BV function on the boundary of the set*. The following theorem provides such a trace, which depends on the value of the function on the surroundings of the set.

THEOREM A.3 (see [27, p. 37]). *If Ω is a bounded open set with Lipschitz continuous boundary $\partial\Omega$ and $f \in BV(\Omega)$, then there exists a function $f_{tr} \in L^1(\partial\Omega)$ such that, for $\mathcal{H}_{n-1}$-almost all $x \in \partial\Omega$,*

$$\lim_{r\to 0} \int_{B(x,r)\cap\Omega} |f(z) - f_{tr}(x)|dz = 0,$$

and f_{tr} is called the trace function.

THEOREM A.4 (see [27, p. 172]). *We let A and B denote two sets of finite perimeter. If Ω is any open set, then*

$$\mathrm{Per}(A \cap B, \Omega) + \mathrm{Per}(A \cup B, \Omega) \le \mathrm{Per}(A, \Omega) + \mathrm{Per}(B, \Omega).$$

Proof. We let f, g be two smooth functions with $0 \le f \le 1$, $0 \le g \le 1$. We define $\Psi = f + g - fg$ and $\Phi = fg$. We note that

$$\int_\Omega |D\Psi| \le \int_\Omega (1-f)|Dg| + \int_\Omega (1-g)|Df|,$$
$$\int_\Omega |D\Phi| \le \int_\Omega f|Dg| + \int_\Omega g|Df|.$$

This implies

$$\int_\Omega |D\Phi| + \int_\Omega |D\Psi| \le \int_\Omega |Df| + \int_\Omega |Dg|. \tag{51}$$

We can find [27] sequences of smooth functions f_j and g_j such that $f_j \to \varphi_A$, $g_j \to \varphi_B$ in $L^1(\Omega)$ and $\int_\Omega |Df_j| \to \int_\Omega |D\varphi_A|$, $\int_\Omega |Dg_j| \to \int_\Omega |D\varphi_B|$. Since $\Psi_j = f_j + g_j - f_j g_j \to \varphi_{A\cup B}$, $\Phi_j = f_j g_j \to \varphi_{A\cap B}$, the theorem follows from (51) and Proposition A.1. ∎

THEOREM A.5 (see [27, p. 173]). *Let $E = E_1 \cup E_2$, and let $\mathcal{H}_{n-1}(\overline{E}_1 \cap \overline{E}_2) = 0$. Then for any open set A we have*

$$\int_A |D\varphi_E| = \int_A |D\varphi_{E_1}| + \int_A |D\varphi_{E_2}|. \tag{52}$$

Moreover, if E has least area in A, the same is true for E_1 and E_2.

LEMMA A.1 (see [27, p. 28]). *Let $f \in BV(\Omega)$. If $A \subset\subset \Omega$ is an open set with Lipschitz continuous boundary ∂A, then $f|_A$ and $f|_{\Omega\setminus\overline{A}}$ belong to $BV(A)$ and $BV(\Omega\setminus\overline{A})$, respectively, and*

$$\int_{\partial A} |Df| = \int_{\partial A} |f_A^- - f_A^+| d\mathcal{H}_{n-1},$$

where $f_A^- = (f_A)_{tr}$ and $f_A^+ = (f|_{\Omega\setminus\overline{A}})_{tr}$, the traces on ∂A of $f|_A$ and $f|_{\Omega\setminus\overline{A}}$, respectively.

LEMMA A.2. *If E is a set of finite perimeter and $x \in \mathbb{R}^n$, then, for almost every r,*

$$\mathrm{Per}(E \cap B(x,r), \mathbb{R}^n) = \mathrm{Per}(E, B(x,r)) + \mathcal{H}_{n-1}(E \cap \partial B(x,r)).$$

Proof. We denote

$$F(x) = \begin{cases} \varphi_E(x), & x \in B(x,r), \\ 0, & x \in \mathbb{R}^n \setminus B(x,r). \end{cases} \tag{53}$$

From Lemma A.1 and using (53) we have

$$\int_{\mathbb{R}^n} |DF| = \int_{B(x,r)} |D\varphi_E| + \int_{\partial B(x,r)} |(\varphi_E)_{tr}| d\mathcal{H}_{n-1}. \tag{54}$$

The lemma follows from (54) since $\int_{\mathbb{R}^n} |DF| = \mathrm{Per}(E \cap B(x,r), \mathbb{R}^n)$ and $\varphi_E = (\varphi_E)_{tr}$ for almost every r. ∎

LEMMA A.3 (see [27, p. 25]). *If E is a set of finite perimeter and $x \in \mathbb{R}^n$, then, for every r,*

$$|E|^{\frac{n-1}{n}} \le C(n)\mathrm{Per}(E, \mathbb{R}^n).$$

LEMMA A.4 (see [27, p. 25]). *If E is a set of finite perimeter and $x \in \mathbb{R}^n$, then, for every r,*

$$\min\{|E \cap B(x,r)|, |(\mathbb{R}^n \setminus E) \cap B(x,r)|\}^{\frac{n-1}{n}} \le C(n) \int_{B(x,r)} |D\varphi_E|.$$

LEMMA A.5. *Let E be a set of finite perimeter that minimizes area in the open set Ω. If $x \in \partial E \cap \Omega$ has density γ_x (see Definition A.6), then $0 < \gamma_x < 1$.*

Proof. We take $x \in \partial E \cap \Omega$. Let $r_0 > 0$ such that $B(x, r_0) \subset \Omega$. We now prove that there exist universal constants C_1, C_2 such that, for all $r \le r_0$,

$$|B(x,r) \cap E| \ge C_1 r^n, \quad |B(x,r) \cap (\mathbb{R}^n \setminus E)| \ge C_2 r^n. \tag{55}$$

The computation that gives the first part of (55) is contained in the proof of Lemma 5.2. The second part (i.e., for the complement of E) is proven in the same way, and we present here again the argument since (55) is a fundamental property of minimal surfaces. We let $F = \mathbb{R}^n \setminus E$. For all $r \le r_0$ we have

$$\int_{B(x,r)} |D\varphi_F| \le \mathcal{H}_{n-1}(F \cap \partial B(x,r)). \tag{56}$$

We define $V(r) = |F \cap B(x,r)|$, $r \le r_0$. Using the isoperimetric inequality given in Lemma A.3 we have that

$$|F \cap B(x,r)| \le C[\mathrm{Per}(F \cap B(x,r), \mathbb{R}^n)]^{\frac{n}{n-1}}.$$

Proceeding as in Lemma A.2 we can prove that $\mathrm{Per}(F \cap B(x,r), \mathbb{R}^n) = \mathrm{Per}(F, B(x,r)) + \mathcal{H}_{n-1}(F \cap \partial B(x,r))$ for almost every $r \le r_0$, and hence

$$\begin{aligned} |F \cap B(x,r)| &\le C[\mathrm{Per}(F \cap B(x,r), \mathbb{R}^n)]^{\frac{n}{n-1}} \\ &= C[\mathrm{Per}(F, B(x,r)) + \mathcal{H}_{n-1}(F \cap \partial B(x,r))]^{\frac{n}{n-1}} \\ &\le C[\mathcal{H}_{n-1}(F \cap \partial B(x,r))]^{\frac{n}{n-1}}. \end{aligned}$$

Due to Remark A.1 it follows that $V(r) > 0$ for all $r \le r_0$. Since $V'(r) = \mathcal{H}_{n-1}(F \cap \partial B(x,r))$ we have , for almost every $r \le r_0$,

$$V(r) \le CV'(r)^{\frac{n}{n-1}}. \tag{57}$$

If we divide (57) by $V(r)$ and integrate we obtain $V(r)^{\frac{1}{n}} \ge Cr$; i.e.,$V(r) \ge Cr^n$. Now, from (55) we have

$$C_1 r^n \le |B(x,r) \cap E| = |B(x,r)| - |B(x,r) \cap (\mathbb{R}^n \setminus E)| \le |B(x,r)| - C_2 r^n.$$

Therefore

$$0 < \tilde{C}_1 \le \frac{|B(x,r) \cap E|}{|B(x,r)|} \le \tilde{C}_2 < 1,$$

where $\tilde{C}_1$ and $\tilde{C}_2$ are two universal constants. Taking limit as $r \to 0$ and from Definition A.6 we obtain that $0 < \gamma_x < 1$. ∎

LEMMA A.6. *If E is a minimizer corresponding to the class A_{S_1,S_2}, and if the exclusions have at least C^1 boundaries, then there exists a universal constant C such that the set $F \equiv \mathbb{R}^n \setminus (E \cap O)$ satisfies*

$$|F \cap B(x,r)| \ge Cr^n \tag{58}$$

for all $x \in \partial F$, $r \le r_0$, where r_0 is a universal constant.

Proof. We take $x \in \partial F$ and $r < \frac{\alpha}{2}$. We have different situations according to the location of $B(x,r)$. In each case, however, the density estimate (58) can be obtained as in Lemma A.5 from the isoperimetric inequality given in Lemma A.3. In fact, if $B(x,r)$ does not intersect any exclusion or the parallel plane restrictions, then we proceed exactly as in Lemma A.5. We consider now the cases

1. $B(x,r)$ intersects Π_1 (the lower parallel plane restriction) and/or an exclusion.
2. $B(x,r)$ intersects Π_2 (the upper parallel plane restriction) and/or an exclusion.

In case 1, we proceed as in Lemma A.5 with $V(r) = |F \cap B(x,r) \cap O|$, applying the isoperimetric inequality given in Lemma A.3 to the domain $|F \cap B(x,r) \cap O|$. In order to estimate $\mathrm{Per}(F \cap B(x,r) \cap O)$ we use the fact that ∂E is a free boundary (in the sense that we do not impose any restriction as to how the minimizer E meets the exclusions), and hence $\mathrm{Per}(F, B(x,r) \cap O) \le \mathcal{H}_{n-1}(\partial B(x,r) \cap F \cap O)$. If $B(x,r)$ intersects the exclusion I, then (while computing $\mathrm{Per}(F \cap B(x,r) \cap O)$) we can estimate $H_{n-1}(\partial I \cap B(x,r) \cap F)$ by performing a change of variables to flatten the boundary of the exclusion. In case 2, if $B(x,r)$ intersects Π_2 and more than half the ball $B(x,r)$ is

outside the restrictions, then (58) is clear, but, if not, then we consider $B(x, \frac{r}{2})$, and we proceed as in case 1. □

LEMMA A.7. *Let E be a set of finite perimeter in $\mathbb{R}^n$, and let $F = \mathbb{R}^n \setminus E$. If there exists a universal constant C such that*

$$|F \cap B(x,r)| \geq Cr^n \tag{59}$$

for all $x \in \partial F$ and all $r \leq \tilde{r}$, then there exists a sequence of C^∞ sets $E_{\epsilon_k} \subset\subset E$ converging in measure to E and such that

$$\lim_{\epsilon_k \to 0} \mathrm{Per}(E_{\epsilon_k}, \mathbb{R}^n) = \mathrm{Per}(E, \mathbb{R}^n).$$

Proof. The proof of this lemma is an improvement of Theorem 3.42 in [2] under the extra condition (59). In fact, we consider the standard mollified functions $u_\epsilon = \varphi_E * \rho_\epsilon$ and $v_\epsilon = \varphi_F * \rho_\epsilon$, where spt $\rho \subset B_1$, $\rho \equiv 1$ on $B(x, \frac{1}{2})$, and $\rho_\epsilon = \frac{1}{\epsilon^n}\rho(\frac{\cdot}{\epsilon})$. We note that $u_\epsilon + v_\epsilon = 1$. If $x \in \partial F$ and $\epsilon < \tilde{r}$, we obtain from (59)

$$\begin{aligned} v_\epsilon(x) &= \frac{1}{\epsilon^n} \int_{B(x,\epsilon)\cap F} \rho\left(\frac{x-y}{\epsilon}\right) dy \\ &\geq \frac{1}{\epsilon^n} \int_{B(x,\frac{\epsilon}{2})\cap F} \rho\left(\frac{x-y}{\epsilon}\right) dy \\ &= \frac{1}{\epsilon^n} \left| B\left(x, \frac{\epsilon}{2}\right) \cap F \right| \\ &\geq C\epsilon^{-n}\epsilon^n = C. \end{aligned}$$

Therefore, we can choose t close enough to 1 so that

$$\{v_\epsilon < 1 - t\} = \{u_\epsilon > t\} \subset\subset E. \tag{60}$$

Using an exercise problem in [2, p. 39] we have that, for almost every $t \in (0, 1)$,

$$\lim_{\epsilon \to 0} \mathrm{Per}(\{u_\epsilon > t\}, \mathbb{R}^n) = \mathrm{Per}(E, \mathbb{R}^n). \tag{61}$$

Hence, we choose t such that (60) and (61) holds, and we define $E_{\epsilon_k} \equiv \{u_{\epsilon_k} > t\}$. We can now conclude as in [2]. □

Acknowledgments. I am greatly indebted to Luis A. Caffarelli for introducing me to this subject. I am also very thankful to Lawrence C. Evans, Rafael de la Llave, Ovidiu Savin, and the referees for many useful comments. The proof of Lemma A.7 was communicated by Luigi Ambrosio.

REFERENCES

[1] O. ALVAREZ, *Homogenization of Hamilton-Jacobi equations in perforated sets*, J. Differential Equations, 159 (1999), pp. 543–577.

[2] L. AMBROSIO, N. FUSCO, AND D. PALLARA, *Functions of bounded variation and free discontinuity problems*, Oxford Mathematical Monographs, The Clarendon Press, Oxford University Press, New York, 2000.

[3] F. AUER, *Minimale hyperflächen in riemannsche n-torus*, Albert-Ludwigs Univ. thesis, Freiburg, Germany, 1997.

[4] F. AUER AND V. BANGERT, *Minimising currents and the stable norm in codimension one*, C. R. Acad. Sci. Paris Sér. I Math., 333 (2001), pp. 1095–1100.

[5] V. Bangert, *Minimal geodesics*, Ergodic Theory Dynam. Systems, 10 (1990), pp. 263–286.
[6] V. Bangert, *Minimal foliations and laminations*, in Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Zürich, 1994), Birkhäuser, Basel, 1995, pp. 453–464.
[7] M. Bardi, M. G. Crandall, L. C. Evans, H. M. Soner, and P. E. Souganidis, *Viscosity Solutions and Applications*, Lecture Notes in Math. 1660, I. Capuzzo Dolcetta and P. L. Lions, eds., Springer-Verlag, Berlin, 1997.
[8] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser Boston, Boston, MA, 1997.
[9] G. Barles and P. E. Souganidis, *On the large time behavior of solutions of Hamilton–Jacobi equations*, SIAM J. Math. Anal., 31 (2000), pp. 925–939.
[10] D. Burago, *Periodic metrics*, in Seminar on Dynamical Systems (St. Petersburg, 1991), Progr. Nonlinear Differential Equations Appl. 12, Birkhäuser, Basel, 1994, pp. 90–95.
[11] D. Burago, S. Ivanov, and B. Kleiner, *On the structure of the stable norm of periodic metrics*, Math. Res. Lett., 4 (1997), pp. 791–808.
[12] L. A. Caffarelli and A. Córdoba, *An elementary regularity theory of minimal surfaces*, Differential Integral Equations, 6 (1993), pp. 1–13.
[13] L. A. Caffarelli and A. Córdoba, *Correction: "An elementary regularity theory of minimal surfaces"* [Differential Integral Equations, 6 (1993), pp. 1–13], Differential Integral Equations, 8 (1995), p. 223.
[14] L. A. Caffarelli and R. de la Llave, *Planelike minimizers in periodic media*, Comm. Pure Appl. Math., 54 (2001), pp. 1403–1441.
[15] I. Capuzzo-Dolcetta and H. Ishii, *On the rate of convergence in homogenization of Hamilton-Jacobi equations*, Indiana Univ. Math. J., 50 (2001), pp. 1113–1129.
[16] M. C. Concordel, *Periodic homogenization of Hamilton-Jacobi equations: Additive eigenvalues and variational formula*, Indiana Univ. Math. J., 45 (1996), pp. 1095–1117.
[17] M. C. Concordel, *Periodic homogenisation of Hamilton-Jacobi equations.* II. *Eikonal equations*, Proc. Roy. Soc. Edinburgh Sect. A, 127 (1997), pp. 665–689.
[18] E. De Giorgi, *Frontiere orientate di misura minima*, Seminario di Matematica della Scuola Normale Superiore di Pisa, 1960-61, Editrice Tecnico Scientifica, Pisa, 1961.
[19] L. C. Evans and D. Gomes, *Effective Hamiltonians and averaging for Hamiltonian dynamics.* I, Arch. Ration. Mech. Anal., 157 (2001), pp. 1–33.
[20] L. C. Evans and D. Gomes, *Effective Hamiltonians and averaging for Hamiltonian dynamics.* II, Arch. Ration. Mech. Anal., 161 (2002), pp. 271–305.
[21] L. C. Evans, *The perturbed test function method for viscosity solutions of nonlinear PDE*, Proc. Roy. Soc. Edinburgh Sect. A, 111 (1989), pp. 359–375.
[22] L. C. Evans, *Periodic homogenisation of certain fully nonlinear partial differential equations*, Proc. Roy. Soc. Edinburgh Sect. A, 120 (1992), pp. 245–265.
[23] L. C. Evans, *Partial differential equations*, Graduate Stud. Math. 19, AMS, Providence, RI, 1998.
[24] L. C. Evans, *Effective Hamiltonians and quantum states*, in Séminaire: Équations aux Dérivées Partielles, 2000–2001, Sémin. Équ. Dériv. Partielles, Exp. No. XXII, École Polytech., Palaiseau, 2001.
[25] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.
[26] H. Federer, *Geometric Measure Theory*, Die Grundlehren der mathematischen Wissenschaften, Band 153, Springer-Verlag, New York, 1969.
[27] E. Giusti, *Minimal Surfaces and Functions of Bounded Variation*, Birkhäuser Verlag, Basel, 1984.
[28] M. Goldberg, *Stable norms—examples and remarks*, in General inequalities, 7 (Oberwolfach, 1995), Internat. Ser. Numer. Math. 123, Birkhäuser, Basel, 1997, pp. 61–64.
[29] M. Grüter, *Optimal regularity for codimension one minimal surfaces with a free boundary*, Manuscripta Math., 58 (1987), pp. 295–343.
[30] K. Horie and H. Ishii, *Homogenization of Hamilton-Jacobi equations on domains with small scale periodic structure*, Indiana Univ. Math. J., 47 (1998), pp. 1011–1058.
[31] Z. Lin, L. Xiao-Biao, M. Torres, and Z. Hongkai, *Theoretical and numerical analysis of the weighted minimal surface problem*, Methods Appl. Anal., 10 (2003), pp. 199–214.
[32] P. L. Lions, G. Papanicolaou, and S. R. S. Varadhan, *Homogenization of Hamilton-Jacobi Equations*, manuscript.
[33] D. Massart, *Stable norms of surfaces: Local structure of the unit ball of rational directions*, Geom. Funct. Anal., 7 (1997), pp. 996–1010.
[34] L. Simon, *Lectures on Geometric Measure Theory*, Australian National University, Centre for Mathematical Analysis, Canberra, 1983.

SIAM J. MATH. ANAL.
Vol. 36, No. 2, pp. 552–579

CONVEXIFIED GAUSS CURVATURE FLOW OF SETS: A STOCHASTIC APPROXIMATION*

HITOSHI ISHII† AND TOSHIO MIKAMI‡

Abstract. We construct a discrete stochastic approximation of a convexified Gauss curvature flow of boundaries of bounded open sets in an anisotropic external field. We also show that a weak solution to the PDE which describes the motion of a bounded open set is unique and is a viscosity solution of it.

Key words. convexified Gauss curvature flow, stochastic approximation, weak solution, viscosity solution

AMS subject classifications. 53C44, 35K65, 60D05

DOI. 10.1137/S0036141002420509

1. Introduction.

Gauss curvature flow is known as a mathematical model of the wearing process of a convex stone rolling on a beach and has been studied by many authors (see, e.g., [2, 3, 6, 7, 11, 14, 16, 23]).

In the last few years we have been generalizing the theory of Gauss curvature flow to a class of nonconvex sets.

In [17] we studied the existence and the uniqueness of a viscosity solution to the PDE that describes the time evolution of a nonconvex graph by a convexified Gauss curvature (see (1.10) for the PDE).

In [20] we proposed and studied the discrete stochastic approximations of evolving functions which are generalizations of those considered in [17], proved the existence and the uniqueness of a weak solution to the PDE which appears as the continuum limit of discrete stochastic processes, and discussed under what conditions a weak solution to the PDE is a viscosity solution of it.

In [19] we studied the existence and the uniqueness of the motion (or time evolution) of a nonconvex compact set which evolves by a convexified Gauss curvature in $\mathbf{R}^N$ ($N \geq 2$), by the level set approach in the theory of viscosity solutions (see, e.g., [5, 10, 23] for the level set approach; also see [1]).

We introduce the notion of the motion of a smooth oriented closed hypersurface by a convexified Gauss curvature.

Let M be a smooth oriented closed hypersurface in $\mathbf{R}^N$ and ν be a smooth vector field over M of unit normal vectors. For $x \in M$, let T_xM denote the tangent space of M at x, and let $A_x : T_xM \mapsto T_xM$ denote the Weingarten map at x defined by the following:

$$A_x(e) = -D_e\nu \quad \text{for } e \in T_xM, \tag{1.1}$$

where $D_e\nu$ denotes the derivative of ν with respect to e. Recall that the principal

*Received by the editors December 27, 2002; accepted for publication (in revised form) November 14, 2003; published electronically August 6, 2004.
http://www.siam.org/journals/sima/36-2/42050.html

†Department of Mathematics, School of Education, Waseda University, 1-6-1 Nishiwaseda, Shinjuku-ku, Tokyo 169-8050, Japan (ishii@edu.waseda.ac.jp). This author was supported in part by Grants-in-Aid for Scientific Research 14654032 and 15340051, JSPS.

‡Department of Mathematics, Hokkaido University, Sapporo 060-0810, Japan (mikami@math.sci.hokudai.ac.jp). This author was supported in part by Grants-in-Aid for Scientific Research 13640096 and 14654032, JSPS.

curvatures $\kappa_1, \dots, \kappa_{N-1}$ of M at x are the eigenvalues of the symmetric map A_x and the Gauss curvature $K(x)$ of M at x is given by $\det A_x$.

Let C be the convex hull $\operatorname{co} M$ of M. We define $\sigma : M \mapsto \{0, 1\}$ by

$$\sigma(x) = \begin{cases} 1 & \text{if } x \in M \cap \partial C, \\ 0 & \text{otherwise} \end{cases} \tag{1.2}$$

and call $\sigma(x)K(x)$ the *convexified* Gauss curvature of M at x.

The motion of a smooth oriented closed hypersurface by a convexified Gauss curvature is the curvature flow:

$$v = -\sigma K \nu, \tag{1.3}$$

where ν denotes the unit outward normal vector field on the hypersurface and v denotes the velocity of the hypersurface.

Let $(A_x)_+$ denote the positive part of the symmetric map A_x. $K_+(x) := \det\{(A_x)_+\}$ is called the *positive part* of the Gauss curvature of M at x, and the following holds:

$$\sigma(x)K(x) = \sigma(x)K_+(x). \tag{1.4}$$

Remark 1.1. For $x \in M$,

$$\det\{(A_x)_+\} = \operatorname{Det}_+(A_x) := \begin{cases} \det A_x & \text{if } A_x \text{ is nonnegative definite,} \\ 0 & \text{otherwise.} \end{cases}$$

The discrete approximation of a smooth simple closed convex curve which evolves as the curvature flow was considered by Girão and is useful in the numerical analysis (see [13] and the references therein). We refer to [12] and the references therein for the recent development of this topic.

The discrete stochastic approximation of the curvature flow of smooth simple closed convex curves is given in [18], where the model and the approach are completely different from those in this paper.

In this paper we propose and study the discrete stochastic approximation of a convexified Gauss curvature flow of boundaries of bounded open sets in an anisotropic external field:

$$v = -R(\nu)\sigma K \nu, \tag{1.5}$$

where $R : \mathbf{S}^{N-1} \mapsto [0, \infty)$ controls the anisotropy of an external field (see (1.3) for notation).

This is important since a stone on a beach is not always convex. Equation (1.5) also gives a mathematical model of the wearing process of a stone which is hit by several kinds of matter from different directions or a stone which has a fine microstructure consisting of anisotropic materials.

Our result in this paper is the first one in the case $N \geq 3$, among random and nonrandom results, which gives a discrete approximation of the motion of a bounded open set, in $\mathbf{R}^N$, by Gauss curvature.

Therefore the construction of a nonrandom version of such an approximation as above in the case $N \geq 3$ is an open problem, although the stochastic approximation is more realistic than a nonrandom one in that a stone rolling on a beach moves randomly.

We briefly describe what we proved in [20] and then discuss the results in this paper more precisely to compare a convexified Gauss curvature flow of graphs with that of closed hypersurfaces. Put $n := N - 1$.

For $x \in \mathbf{R}^n$ and $u : \mathbf{R}^n \mapsto \mathbf{R}$, the following set is called the subdifferential of u at x:

$$\partial u(x) := \{p \in \mathbf{R}^n : u(y) - u(x) \geq p \cdot (y - x) \text{ for all } y \in \mathbf{R}^n\}, \tag{1.6}$$

where $\cdot$ denotes the inner product in $\mathbf{R}^n$. Let $L^1(\mathbf{R}^n : [0, \infty), dx)$ denote the set of measurable functions $\varphi : \mathbf{R}^n \mapsto [0, \infty)$ for which $\int_{\mathbf{R}^n} |\varphi(x)| dx < \infty$.

Alexandrov–Bakelman's generalized curvature introduced in the following played a crucial role in [20].

DEFINITION 1.1 (see, e.g., [4, section 9.6]). *Let $\overline{R} \in L^1(\mathbf{R}^n : [0, \infty), dx)$ and $u \in C(\mathbf{R}^n)$. For $A \in B(\mathbf{R}^n)$(:=Borel σ-field of $\mathbf{R}^n$), put*

$$\overline{w}(\overline{R}, u, A) := \int_{\cup_{x \in A} \partial u(x)} \overline{R}(y) dy. \tag{1.7}$$

(It is known that $\overline{w}(\overline{R}, u, \cdot) : B(\mathbf{R}^n) \mapsto [0, \infty)$ is completely additive.)

For $\overline{R} \in L^1(\mathbf{R}^n : [0, \infty), dx)$, we showed the existence and the uniqueness of a solution $u \in C([0, \infty) \times \mathbf{R}^n)$ to the following equation (see [20, Theorem 1]): for any $\varphi \in C_o(\mathbf{R}^n)$ and any $t \geq 0$,

$$\int_{\mathbf{R}^n} \varphi(x)(u(t, x) - u(0, x)) dx = \int_0^t ds \int_{\mathbf{R}^n} \varphi(x) \overline{w}(\overline{R}, u(s, \cdot), dx). \tag{1.8}$$

The existence of a continuous solution to (1.8) was given by the continuum limit of the infinite particle systems $\{(Z_m(t, z))_{z \in \mathbf{Z}^n/m}\}_{t \geq 0}$ that satisfies the following: for any $t \geq 0$ and any $z \in \mathbf{Z}^n/m$,

$$P(Z_m(t + \Delta t, z) - Z_m(t, z) > 0) = m^n E[\overline{w}(\overline{R}, \hat{Z}_m(t, \cdot), \{z\})] \Delta t + o(\Delta t) \tag{1.9}$$

as $\Delta t \to 0$ ($m \geq 1$), where $\mathbf{Z}^n/m := \{z/m | z \in \mathbf{Z}^n\}$ and $\hat{Z}_m(t, \cdot)$ denotes a convex envelope of the function $z \mapsto Z_m(t, z)$, i.e., the graph of the boundary of the convex hull, in $\mathbf{R}^N$, of the set $\{(z, y) | z \in \mathbf{Z}^n/m, y \geq Z_m(t, z)\}$.

In [20, Theorem 2], we proved that a continuous solution u to (1.8) sweeps in time $t > 0$ a region with volume given by $t \cdot \overline{w}(\overline{R}, u(0, \cdot), \mathbf{R}^n)$ and that, for continuous solutions u and v to (1.8) with $v(0, \cdot) = \hat{u}(0, \cdot)$, $\hat{u}(t, \cdot)$ is different from $v(t, \cdot)$ at time $t > 0$ in general if $u(0, \cdot) \neq \hat{u}(0, \cdot)$.

We also showed that a continuous solution to (1.8) is a viscosity solution of the following PDE (see [20, Theorem 3]):

$$\partial_t u(t, x) = \chi(u, Du(t, x), t, x) \overline{R}(Du(t, x)) \mathrm{Det}_+(D^2 u(t, x)) \text{ on } (0, \infty) \times \mathbf{R}^n, \tag{1.10}$$

where $Du(t, x) := (\partial u(t, x)/\partial x_i)_{i=1}^n$, $D^2 u(t, x) := (\partial^2 u(t, x)/\partial x_i \partial x_j)_{i,j=1}^n$, and

$$\chi(u, p, t, x) := \begin{cases} 1 & \text{if } p \in \partial u(t, x), \\ 0 & \text{otherwise} \end{cases}$$

(see Remark 1.1 for notation). Here $\partial u(t, x)$ denotes the subdifferential of the function $x \mapsto u(t, x)$. Conversely, we discussed under what conditions a viscosity solution to (1.10) is a solution to (1.8).

Remark 1.2. Suppose that u in (1.10) is twice differentiable in x. Then putting

$$\nu = \left(\frac{(Du(t,x), -1)}{(|Du(t,x)|^2 + 1)^{1/2}} \right)_{x \in \mathbf{R}^n}$$

in (1.3),

$$\chi(u, Du(t,x), t, x)(1 + |Du(t,x)|^2)^{-(n+2)/2} \mathrm{Det}_+(D^2 u(t,x))$$

is the convexified Gauss curvature of $\{(y, u(t,y)) | y \in \mathbf{R}^n\}$ at x.

Next we briefly discuss what we study in this paper.

Let F be a closed convex set in $\mathbf{R}^N$. For $x \in \partial F$, put

$$N_F(x) := \{p \in \mathbf{S}^{N-1} | F \subset \{y | \langle y - x, p \rangle \leq 0\}\},$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product in $\mathbf{R}^N$.

Let $L^1(\mathbf{S}^{N-1} : [0, \infty), d\mathcal{H}^{N-1})$ denote the set of measurable functions $\phi : \mathbf{S}^{N-1} \mapsto [0, \infty)$ for which $\int_{\mathbf{S}^{N-1}} |\varphi(p)| d\mathcal{H}^{N-1}(p) < \infty$, where $d\mathcal{H}^{N-1}$ denotes the $(N-1)$-dimensional Hausdorff outer measure.

To consider a convexified Gauss curvature flow of bounded open sets by the level set approach, we introduce new types of measures.

DEFINITION 1.2. *Let u be a bounded function from a subset of $\mathbf{R}^N$ to $\mathbf{R}$, and $R \in L^1(\mathbf{S}^{N-1} : [0, \infty), d\mathcal{H}^{N-1})$.*

(i) *Let $r \in \mathbf{R}$. For $B \in B(\mathbf{R}^N)$, put*

$$\omega_r(R, u, B) := \int_{\cup_{x \in B \cap \partial A} N_{A^-}(x)} R(p) d\mathcal{H}^{N-1}(p), \tag{1.11}$$

where $A = \mathrm{co}\ u^{-1}([r, \infty))$ and A^- denotes the closure of the set A.

(ii) *For $B \in B(\mathbf{R}^N)$, put*

$$\mathbf{w}(R, u, B) := \int_{\mathbf{R}} dr \omega_r(R, u, B), \tag{1.12}$$

provided the right-hand side is well defined.

Remark 1.3. (i) If $\partial(\mathrm{co}\ u^{-1}([r, \infty)))$ is smooth at x, then $N_{(\mathrm{co}\ u^{-1}([r,\infty)))^-}(x)$ is the set of a unit outward normal vector on $\partial(\mathrm{co}\ u^{-1}([r, \infty)))$ at x. Otherwise $N_{(\mathrm{co}\ u^{-1}([r,\infty)))^-}(x)$ consists of more than one point. (ii) $\omega_r(R, u, \cdot) : B(\mathbf{R}^N) \mapsto [0, \infty)$ in (1.11) is completely additive (see [4, p. 31, Theorem 5.1]). (iii) For u and r in Definition 1.2, the $(N-1)$-dimensional Hausdorff outer measure of the following set is zero (see [4, p. 30, Lemma 5.2]):

$$\cup_{x \in \partial(\mathrm{co}\ u^{-1}([r,\infty)))} \{p \in N_{(\mathrm{co}\ u^{-1}([r,\infty)))^-}(x) : \{x\} \subsetneq \{y \in \mathbf{R}^N : \langle y - x, p \rangle = 0\} \cap (\mathrm{co}\ u^{-1}([r, \infty)))^-\}.$$

When it is not confusing, we write $\omega_r(R, u, dx) = \omega_r(u, dx)$ and $\mathbf{w}(R, u, dx) = \mathbf{w}(u, dx)$ for the sake of simplicity.

The existence and the uniqueness of a solution to the following equation are given in section 2.

DEFINITION 1.3. *Let $T \in [0, \infty]$ and $R \in L^1(\mathbf{S}^{N-1} : [0, \infty), d\mathcal{H}^{N-1})$. A family of bounded open sets $\{D(t)\}_{t \in [0,T)}$ in $\mathbf{R}^N$ is called a convexified Gauss curvature flow in an (R-)anisotropic external field on $[0, T)$ if*

$$D(t) = (\mathrm{co}\ D(t)) \cap D(0) \quad \text{for } t \in [0, T) \tag{1.13}$$

and if the following holds: for any $\varphi \in C_o(\mathbf{R}^N)$ *and any* $t \in [0,T)$,

$$\int_{\mathbf{R}^N} \varphi(x)(I_{D(0)}(x) - I_{D(t)}(x))dx = \int_0^t ds \int_{\mathbf{R}^N} \varphi(x)\omega_1(R, I_{D(s)}(\cdot), dx), \tag{1.14}$$

where $I_A(x) = 1$ *if* $x \in A$ *and* $= 0$ *if* $x \notin A$ *for the set* A.

We also show the existence and the uniqueness of a solution $u \in C_b([0,T) \times \mathbf{R}^N)$ to the following: for any $\varphi \in C_o(\mathbf{R}^N)$ and any $t \in [0,T)$,

$$\int_{\mathbf{R}^N} \varphi(x)(u(0,x) - u(t,x))dx = \int_0^t ds \int_{\mathbf{R}^N} \varphi(x)\mathbf{w}(R, u(s,\cdot), dx). \tag{1.15}$$

The existence of $\{I_{D(t)}\}_{t\geq 0}$ in Definition 1.3 is shown by the continuum limit of a class of particle systems $\{(Y_m(t,z))_{z\in\mathbf{Z}^N/m}\}_{t\geq 0}$ that satisfies the following: for any $t \geq 0$ and any $z \in \mathbf{Z}^N/m$,

$$P(Y_m(t+\Delta t, z) - Y_m(t,z) < 0) = m^N E[\omega_1(Y_m(t,\cdot), \{z\})]\Delta t + o(\Delta t) \tag{1.16}$$

as $\Delta t \to 0$ $(m \geq 1)$ (see Theorem 2.1 in section 2).

The existence and the uniqueness of a solution to (1.15) are given by the continuum limit of the linear combinations of solutions to (1.14) with $D(0) = u(0,\cdot)^{-1}((r,\infty))$ for $r \in \mathbf{R}$ (see Corollary 2.3 in section 2).

We also discuss the properties of $\{D(t)\}_{t\geq 0}$ in Definition 1.3 (see Theorem 2.4 in section 2).

For $p \in \mathbf{R}^N$ and an $(N \times N)$-symmetric real matrix X, put

$$G(p,X) := \begin{cases} |p| \det\left\{\left(-(I_N - \bar{p}\otimes\bar{p})\frac{X}{|p|}(I_N - \bar{p}\otimes\bar{p}) + \bar{p}\otimes\bar{p}\right)_+\right\} & \text{if } p \neq o, \\ 0 & \text{if } p = o \end{cases} \tag{1.17}$$

(see Remark 1.1 for notation), where I_N denotes the $(N \times N)$-identity matrix and $\bar{p} := p/|p|$.

Suppose that a smooth oriented hypersurface M in $\mathbf{R}^N$ is given by $M = \{y \in \mathbf{R}^N \mid \varphi(y) = a,\ D\varphi(y) \neq o\}$ for some $\varphi \in C^2(\mathbf{R}^N)$ and $a \in \mathbf{R}$ and that the vector field ν is given by $\nu_x = D\varphi(x)/|D\varphi(x)|$. Regard the tangent space, T_xM, as the orthogonal complement of ν_x, and let $E_x := \operatorname{span} \nu_x$ and id_{E_x} denote the identity map on E_x. Then the map

$$A_x \oplus \mathrm{id}_{E_x} : \mathbf{R}^N \equiv T_xM \oplus E_x \to T_xM \oplus E_x$$

has a matrix representation

$$-(I_N - \bar{p}\otimes\bar{p})\frac{X}{|p|}(I_N - \bar{p}\otimes\bar{p}) + \bar{p}\otimes\bar{p},$$

with $p = D\varphi(x)$ and $X = D^2\varphi(x)$. Therefore, if $D\varphi(x) \neq o$, then

$$K(x) = \det\left(-(I_N - \bar{p}\otimes\bar{p})\frac{X}{|p|}(I_N - \bar{p}\otimes\bar{p}) + \bar{p}\otimes\bar{p}\right), \tag{1.18}$$

$$K_+(x) = \frac{G(p,X)}{|p|}. \tag{1.19}$$

For $\{D(t)\}_{t\geq 0}$ in Definition 1.3, we show that $I_{D(t)}(x)$ and $I_{D(t)^-}(x)$ are, respectively, a viscosity supersolution and a viscosity subsolution of the following PDE (see Theorem 2.5 in section 2):

$$\partial_t u(t,x) + R\left(\frac{Du(t,x)}{|Du(t,x)|}\right)\sigma^-(u, Du(t,x), t, x)G(Du(t,x), D^2u(t,x)) = 0 \tag{1.20}$$

(where $(t,x) \in (0,\infty)\times \mathbf{R}^N$). Here

$$\sigma^-(u,p,t,x) := \begin{cases} 1 & \text{if } u(t,\cdot) < u(t,x) \text{ on } H(p,x) \text{ and } p \in \mathbf{R}^N \setminus \{o\}, \\ 0 & \text{otherwise,} \end{cases} \tag{1.21}$$

where

$$H(p,x) := \{y \in \mathbf{R}^N \setminus \{x\} \mid \langle y - x, p\rangle \leq 0\}. \tag{1.22}$$

Moreover, we show that a continuous solution to (1.15) is a viscosity solution of (1.20) (see Corollary 2.6 in section 2).

$G(p, -I_N) = |p|^{2-N}$ for $p \neq o$. Indeed, take $p_1, \ldots, p_{N-1} \in \mathbf{S}^{N-1}$ so that $\{p_1, \ldots, p_{N-1}, \bar{p}\}$ is an orthonormal basis of $\mathbf{R}^N$. Then $p_1, \ldots, p_{N-1}$ and $\bar{p}$ are eigenvectors of $(I_N - \bar{p}\otimes\bar{p})^2 = I_N - \bar{p}\otimes\bar{p}$ and $\bar{p}\otimes\bar{p}$, with an eigenvalue 1, respectively. Therefore $G(p,X)$ is not continuous at $p = o$. However, $G(p,X)$ should be continuous if we use the standard approach for a viscosity solution (see [8]). This is why we use the modified definition of a viscosity solution to (1.20) from [21] (see Remark 1.5).

We first introduce the set of admissible test functions. We denote by $\mathcal{F}$ the set of all functions $f \in C^2([0,\infty))$ for which $f'' > 0$ on $(0,\infty)$ and

$$\lim_{r\downarrow 0} \frac{f(r)}{r^N} = 0. \tag{1.23}$$

Let Ω be an open subset of $(0,\infty)\times\mathbf{R}^N$. A function $\varphi \in C^2(\Omega)$ is called admissible in Ω if for any $(\hat{t},\hat{x}) \in \Omega$ for which $D\varphi$ vanishes, there exists $f \in \mathcal{F}$ such that as $(t,x) \to (\hat{t},\hat{x})$,

$$|\varphi(t,x) - \varphi(\hat{t},\hat{x}) - \partial_t\varphi(\hat{t},\hat{x})(t-\hat{t})| \leq f(|x-\hat{x}|) + o(|t-\hat{t}|). \tag{1.24}$$

We denote by $\mathcal{A}(\Omega)$ the set of all admissible functions in Ω.

Remark 1.4. $f(r) = r^{N+1} \in \mathcal{F}$ and $\varphi(t,x) = f(|x-\hat{x}|) \in \mathcal{A}((0,\infty)\times\mathbf{R}^N)$ for any $\hat{x} \in \mathbf{R}^N$.

DEFINITION 1.4 (viscosity solution). *Let $0 < T \leq \infty$ and set $\Omega := (0,T)\times\mathbf{R}^N$, and put $R(o/|o|) := 0$.*

(i) *A function $u \in LSC(\Omega)$ is called a viscosity supersolution of* (1.20) *in Ω if whenever $\varphi \in \mathcal{A}(\Omega)$, $(s,y) \in \Omega$, and $u - \varphi$ attains a local minimum at (s,y); then*

$$\partial_t\varphi(s,y) + R\left(\frac{D\varphi(s,y)}{|D\varphi(s,y)|}\right)\sigma^+(u, D\varphi(s,y), s, y)G(D\varphi(s,y), D^2\varphi(s,y)) \geq 0, \tag{1.25}$$

where

$$\sigma^+(u,p,s,y) := \begin{cases} 1 & \text{if } u(s,\cdot) \leq u(s,y) \text{ on } H(p,y) \text{ and } p \in \mathbf{R}^N \setminus \{o\}, \\ 0 & \text{otherwise.} \end{cases} \tag{1.26}$$

(ii) *A function $u \in USC(\Omega)$ is called a viscosity subsolution of* (1.20) *in* Ω *if whenever* $\varphi \in \mathcal{A}(\Omega)$, $(s,y) \in \Omega$, *and* $u - \varphi$ *attains a local maximum at* (s,y), *then*

$$\partial_t \varphi(s,y) + \sigma^-(u, D\varphi(s,y), s, y) R\left(\frac{D\varphi(s,y)}{|D\varphi(s,y)|}\right) G(D\varphi(s,y), D^2\varphi(s,y)) \le 0. \tag{1.27}$$

(iii) *A function $u \in C(\Omega)$ is called a viscosity solution of* (1.20) *in* Ω *if it is a viscosity supersolution and a viscosity subsolution of* (1.20) *in* Ω.

Remark 1.5. (i) $\sigma^+(u,p,s,y) \ge \sigma^-(u,p,s,y)$ for all $u : \Omega \mapsto \mathbf{R}$ and all $(p,s,y) \in \mathbf{R}^N \times \Omega$. (ii) Let $\mathcal{A}_0(\Omega)$ denote the set of all $\phi_1(t) + \phi_2(x) \in \mathcal{A}(\Omega)$ such that $x \mapsto G(D\phi_2(x), D^2\phi_2(x))$ is continuous in Ω. Then one can replace, in Definition 1.4, $\mathcal{A}(\Omega)$ by $\mathcal{A}_0(\Omega)$ (see [19]). (iii) For any $f \in \mathcal{F}$ and $\hat{x} \in \mathbf{R}^N$, $\varphi(t,x) = f(|x - \hat{x}|) \in \mathcal{A}_0((0,\infty) \times \mathbf{R}^N)$.

In section 2 we state our main results, which will be proved in section 4. In section 3 we give technical lemmas.

2. Main result. In this section we give our main result.

We give two assumptions to state the stochastic process which approximates the solution to (1.13)–(1.14).

(A.0). D is a nonempty bounded open set in $\mathbf{R}^N$ such that Vol(∂D) :=the Lebesgue measure of the boundary ∂D of D is zero.

(A.1). $R \in L^1(\mathbf{S}^{N-1} : [0,\infty), d\mathcal{H}^{N-1})$ and $||R||_{L^1(\mathbf{S}^{N-1})} = 1$.

Take $K > 0$ so that co $D \subset [-K+1, K-1]^N$ (see (1.2) for notation). For $m \ge 1$, put

$$\mathcal{S}_m := \{I_A : [-K,K]^N \cap (\mathbf{Z}^N/m) \mapsto \{0,1\} | A \subset [-K,K]^N \cap \mathbf{Z}^N/m\} \tag{2.1}$$

(see (1.14) for notation), and

$$D_m := D \cap (\mathbf{Z}^N/m). \tag{2.2}$$

For x, $z \in \mathbf{Z}^N/m$, and $v \in \mathcal{S}_m$, put

$$v_{m,z}(x) := \begin{cases} v(x) & \text{if } x \ne z, \\ 0 & \text{if } x = z, \end{cases}$$

and for $f : \mathcal{S}_m \mapsto \mathbf{R}$, put

$$A_m f(v) := m^N \sum_{z \in [-K,K]^N \cap (\mathbf{Z}^N/m)} \omega_1(R, v, \{z\})\{f(v_{m,z}) - f(v)\}. \tag{2.3}$$

Let $m \ge 1$ and $\{y_m(k,\cdot)\}_{k \ge 0}$ be a Markov chain on $\mathcal{S}_m$ such that $y_m(0,\cdot) = I_{D_m}(\cdot)$ and such that the following holds: for any $k \ge 0$ and any $z \in [-K,K]^N \cap (\mathbf{Z}^N/m)$,

$$P(y_m(k+1,\cdot) = (y_m(k,\cdot))_{m,z} | y_m(0,\cdot), \dots, y_m(k,\cdot)) = \omega_1(R, y_m(k,\cdot), \{z\}). \tag{2.4}$$

Let $\{\Delta_k\}_{k \ge 0}$ be independent, exponentially distributed random variables with parameter one and be independent of $\{y_m(k,\cdot)\}_{k \ge 0}$. Put

$$Y_m(t,\cdot) := y_m(k,\cdot) \qquad \text{if} \quad \frac{1}{m^N} \sum_{i=-1}^{k-1} \Delta_i \le t < \frac{1}{m^N} \sum_{i=-1}^{k} \Delta_i, \tag{2.5}$$

where $\Delta_{-1} := 0$.

Then $\{Y_m(t,\cdot)\}_{t\geq 0}$ is a Markov process on $\mathcal{S}_m$ $(m \geq 1)$, with the generator A_m, such that $Y_m(0,z) = I_{D_m}(z)$ $(z \in [-K,K]^N \cap (\mathbf{Z}^N/m))$ (see [9, p. 162]).

Remark 2.1. (i) If $y_m(k,\cdot) \not\equiv 0$, then

$$\sum_{z\in[-K,K]^N\cap(\mathbf{Z}^N/m)} \omega_1(R, y_m(k,\cdot), \{z\}) = 1.$$

(ii) If $\omega_1(R, y_m(k,\cdot), \{z\}) > 0$, then $z \in \partial\{\text{co } y_m(k,\cdot)^{-1}(1)\} \cap y_m(k,\cdot)^{-1}(1)$.

(iii) $y_m(k,\cdot) \equiv 0$ if and only if $k \geq \#D_m$:=the cardinal number of the set D_m, a.s.

For $(t,x) \in [0,\infty) \times [-K,K]^N$, put also

$$X_m(t,x) := I_{(\text{co } Y_m(t,\cdot)^{-1}(1))^o \cap D}(x), \tag{2.6}$$

where A^o denotes the interior of the set $A \subset \mathbf{R}^N$.

Then $\{X_m(t,\cdot)\}_{t\geq 0}$ is a stochastic process on

$$\mathcal{S} := \{f \in L^2([-K,K]^N) : ||f||_{L^2([-K,K]^N)} \leq (2K)^N\}, \tag{2.7}$$

which is a complete separable metric space by the metric

$$d_{\mathcal{S}}(f,g) := \sum_{k=1}^{\infty} \frac{\min(|\langle f-g, e_k\rangle_{L^2([-K,K]^N)}|, 1)}{2^k}. \tag{2.8}$$

Here $\{e_k\}_{k\geq 1}$ denotes a complete orthonomal basis of $L^2([-K,K]^N)$.

The following is our main result.

THEOREM 2.1. *Suppose that* (A.0)–(A.1) *hold. Then there exists a unique solution* $\{D(t)\}_{t\geq 0}$ *to* (1.13)–(1.14) *with* $D(0) = D$ *on* $[0,\infty)$ *such that* $I_{D(\cdot)}(\cdot) \in C([0,\infty) : L^2([-K,K]^N))$ *and such that the following holds: for any* $\gamma > 0$,

$$\lim_{m\to\infty} P\left(\sup_{t\geq 0} ||X_m(t,\cdot) - I_{D(t)}(\cdot)||_{L^2([-K,K]^N)} \geq \gamma\right) = 0. \tag{2.9}$$

We recall the definition of Hausdorff metric: for compact sets A and $B \subset \mathbf{R}^N$,

$$d_H(A,B) := \max(\max_{p\in A} dist(p,B), \max_{q\in B} dist(q,A)). \tag{2.10}$$

As a corollary, we obtain the following.

COROLLARY 2.2. *Suppose that* (A.0)–(A.1) *hold and that* D *is convex. Then for a unique solution* $\{D(t)\}_{t\geq 0}$ *to* (1.13)–(1.14) *with* $D(0) = D$ *on* $[0,\infty)$, *the following holds: for any* $T \in [0, \text{Vol}(D))$ *and any* $\gamma > 0$,

$$\lim_{m\to\infty} P\left(\sup_{0\leq t\leq T} d_H(\partial(\text{co } Y_m(t,\cdot)^{-1}(1)), \partial D(t)) \geq \gamma\right) = 0. \tag{2.11}$$

We introduce the assumption on the initial function in (1.15).

(A.2). $h \in C_b(\mathbf{R}^N)$. For any $r \in \mathbf{R}$, the set $h^{-1}((r,\infty))$ is bounded or $\mathbf{R}^N$.

Then one can easily obtain the following from Theorem 2.1.

COROLLARY 2.3. *Suppose that* (A.1)–(A.2) *hold. Then there exists a unique continuous solution* $\{u(t,\cdot)\}_{t\geq 0}$ *to* (1.15) *with* $u(0,\cdot) = h(\cdot)$ *on* $[0,\infty)$. *In addition, for any* $r \in \mathbf{R}$, $\{u(t,\cdot)^{-1}((r,\infty))\}_{t\geq 0}$ *is a unique solution to* (1.13)–(1.14) *with* $D(0) =$

$h^{-1}((r,\infty))$ *on* $[0,\infty)$. *In particular,* $\{u(t,\cdot)^{-1}((r,\infty))\}_{t\ge 0}$ *depends only on the set* $u(0,\cdot)^{-1}((r,\infty))$.

The following theorem collects some elementary properties of solutions to (1.13)–(1.14).

THEOREM 2.4. *Suppose that* (A.0)–(A.1) *hold. Let* $\{D(t)\}_{t\ge 0}$ *be a unique solution to* (1.13)–(1.14) *with* $D(0)=D$ *on* $[0,\infty)$. *Then the following holds.*

(a) $t\mapsto D(t)$ *is nonincreasing on* $[0,\infty)$.

(b) *For any* $t\le T^*:=\mathrm{Vol}(D(0))$,

$$\mathrm{Vol}(D(0)\backslash D(t))=t. \tag{2.12}$$

(c) $D(t)=\emptyset$ *for* $t\ge T^*$.

(d) *Let* $\{D_1(t)\}_{t\ge 0}$ *be a solution to* (1.13)–(1.14) *on* $[0,\infty)$ *such that* $D_1(0)$ *is a bounded, convex, open set which contains* D. *Then*

$$D(t)\subset D_1(t) \quad \text{for all } t\ge 0, \tag{2.13}$$

where the equality holds if and only if $D(0)=D_1(0)$.

Under

$$R\in C(S^{N-1}:[0,\infty)), \tag{A.3}$$

we give the relation between the solution to (1.13)–(1.14) and the viscosity solution of (1.20).

THEOREM 2.5. *Suppose that* (A.0)–(A.1) *and* (A.3) *hold. Then for a unique solution* $\{D(t)\}_{t\ge 0}$ *to* (1.13)–(1.14) *with* $D(0)=D$ *on* $[0,\infty)$, $I_{D(t)}(x)$ *and* $I_{D(t)^-}(x)$ *are a viscosity supersolution and a viscosity subsolution to* (1.20) *in* $(0,\infty)\times\mathbf{R}^N$, *respectively.*

As a corollary, we obtain the following.

COROLLARY 2.6. *Suppose that* (A.1)–(A.3) *hold. Then a continuous solution* $\{u(t,\cdot)\}_{t\ge 0}$ *to* (1.15) *with* $u(0,\cdot)=h(\cdot)$ *on* $[0,\infty)$ *is a viscosity solution to* (1.20) *in* $(0,\infty)\times\mathbf{R}^N$.

3. Lemmas. In this section we give lemmas which will be used in the next section.

We extend $Y_m(t,\cdot)$ as a function on $\mathbf{R}^N$ so that

$$\overline{Y}_m(t,x)=\begin{cases} 0 & (x\in D^c\cap(\mathbf{Z}^N/m)),\\ Y_m(t,[mx]/m) & (x=(x_i)_{i=1}^N\in\mathbf{R}^N),\end{cases} \tag{3.1}$$

where $[mx]:=([mx_i])_{i=1}^N$ and $[mx_i]$ denotes an integer for which $[mx_i]\le mx_i<[mx_i]+1$.

Remark 3.1. For $z\in[-K,K]^N\cap(\mathbf{Z}^N/m)$,

$$Y_m(t,z)=m^N\int_{\{x\in\mathbf{R}^N\,|\,[mx]=mz\}}\overline{Y}_m(t,x)dx.$$

LEMMA 3.1. *Suppose that* (A.0)–(A.1) *hold. Then* $\{\overline{Y}_m(\cdot,\cdot)\}_{m\ge 1}$ *is tight in* $D([0,\infty):\mathcal{S})$, *and any weak limit point of* $\{\overline{Y}_m(\cdot,\cdot)\}_{m\ge 1}$ *belongs to the set* $C([0,\infty):\mathcal{S})$.

Proof. Since $\mathcal{S}$ is compact and since $t\mapsto\overline{Y}_m(t,x)$ is nonincreasing for any $x\in\mathbf{R}^N$, we only have to show the following (see [9, p. 129, Corollary 7.4, and p. 148, Theorem

10.2]): for any $\eta > 0$ and $T > 0$, there exists $\delta > 0$ such that for any i for which $1 \le i \le [T/\delta] + 1$,

$$\lim_{m\to\infty} P(||\overline{Y}_m(i\delta,\cdot) - \overline{Y}_m((i-1)\delta,\cdot)||_{L^1([-K,K]^N)} \ge \eta) = 0. \tag{3.2}$$

Indeed, for any s and t for which $(i-1)\delta \le s \le t \le i\delta$,

$$\overline{Y}_m(s,x) - \overline{Y}_m(t,x) = 0 \text{ or } 1$$

and

$$\begin{aligned} d_{\mathcal{S}}(\overline{Y}_m(t,\cdot),\overline{Y}_m(s,\cdot))^2 &\le ||\overline{Y}_m(t,\cdot) - \overline{Y}_m(s,\cdot)||^2_{L^2([-K,K]^N)} \\ &= ||\overline{Y}_m(i\delta,\cdot) - \overline{Y}_m((i-1)\delta,\cdot)||_{L^1([-K,K]^N)}. \end{aligned}$$

For $\delta < \eta/2$ and $m \ge 1$, by Chebyshev's inequality and Ito's formula (see [15]),

$$\begin{aligned} &P(||\overline{Y}_m(i\delta,\cdot) - \overline{Y}_m((i-1)\delta,\cdot)||_{L^1([-K,K]^N)} \ge \eta) \\ &\le \left(\frac{2}{\eta}\right)^2 E\left[\left| \sum_{z\in D_m} (Y_m(i\delta,z) - Y_m((i-1)\delta,z))\frac{1}{m^N} \right.\right. \\ &\qquad \left.\left. + \int_{(i-1)\delta}^{i\delta} \omega_1(Y_m(s,\cdot),D_m)ds \right|^2\right] \\ &= \left(\frac{2}{\eta}\right)^2 m^{-N} E\left[\int_{(i-1)\delta}^{i\delta} \omega_1(Y_m(s,\cdot),D_m)ds\right] \\ &\le \left(\frac{2}{\eta}\right)^2 m^{-N}\delta \to 0 \quad \text{as } m\to\infty \end{aligned} \tag{3.3}$$

(see (2.2) for notation). Indeed,

$$\begin{aligned} &||\overline{Y}_m(i\delta,\cdot) - \overline{Y}_m((i-1)\delta,\cdot)||_{L^1([-K,K]^N)} \\ &= -\sum_{z\in D_m} (Y_m(i\delta,z) - Y_m((i-1)\delta,z))\frac{1}{m^N} - \int_{(i-1)\delta}^{i\delta} \omega_1(Y_m(s,\cdot),D_m)ds \\ &\qquad + \int_{(i-1)\delta}^{i\delta} \omega_1(Y_m(s,\cdot),D_m)ds. \qquad \square \end{aligned}$$

LEMMA 3.2. *Suppose that* (A.0)–(A.1) *hold. Then there exist a subsequence* $\{m_k\}_{k\ge1} \subset \mathbf{N}$ *and stochastic processes* $\{\overline{Y}_{1,m_k}(\cdot,\cdot)\}_{k\ge1}$ *on a probability space* $(\Omega_1, \mathbf{B}_1, P_1)$ *such that the probability law of* $\overline{Y}_{1,m_k}(\cdot,\cdot)$ *is the same as that of* $\overline{Y}_{m_k}(\cdot,\cdot)$ *for all* $k \ge 1$, *and such that* $\{\overline{Y}_{1,m_k}(\cdot,\cdot)\}_{k\ge1}$ *is convergent in* $D([0,\infty) : \mathcal{S})$, P_1*-a.s., and such that the following holds* P_1*-a.s.: for any* $T > 0$ *and* $\varphi \in C([-K,K]^N)$,

$$\begin{aligned} \sup_{0\le t\le T} &\left| \sum_{z\in D_{m_k}} \varphi + (z)(Y_{1,m_k}(t,z) - Y_{1,m_k}(0,z))\left(\frac{1}{m_k}\right)^N \right. \\ &\left. + \int_0^t \sum_{z\in D_{m_k}} \varphi(z)\omega_1(Y_{1,m_k}(s,\cdot),\{z\})ds \right| \to 0 \quad \text{as } k\to\infty. \end{aligned} \tag{3.4}$$

Here Y_{1,m_k} is defined from $\overline{Y}_{1,m_k}$ in the same way as in Remark 3.1.

Proof. By Lemma 3.1 and Skorohod's theorem (see [9, p. 102, Theorem 1.8]), there exist a subsequence $\{m_{0,k}\}_{k\geq 1} \subset \mathbf{N}$ and stochastic processes $\{\overline{Y}_{1,m_{0,k}}(\cdot,\cdot)\}_{k\geq 1}$ on a probability space $(\Omega_1, \mathbf{B}_1, P_1)$ such that the probability law of $\overline{Y}_{1,m_{0,k}}(\cdot,\cdot)$ is the same as that of $\overline{Y}_{m_{0,k}}(\cdot,\cdot)$ for all $k \geq 1$, and such that $\{\overline{Y}_{1,m_{0,k}}(\cdot,\cdot)\}_{k\geq 1}$ is convergent in $D([0,\infty) : \mathcal{S})$, P_1-a.s.

As in (3.3), by Doob–Kolmogorov's inequality (see [15]), for any $T > 0$ and $\varphi \in C([-K,K]^N)$,

$$E_1\left[\sup_{0\leq t\leq T}\left|\sum_{z\in D_{m_{0,k}}}\varphi(z)(Y_{1,m_{0,k}}(t,z)-Y_{1,m_{0,k}}(0,z))\left(\frac{1}{m_{0,k}}\right)^N + \int_0^t \sum_{z\in D_{m_{0,k}}}\varphi(z)\omega_1(Y_{1,m_{0,k}}(s,\cdot),\{z\})ds\right|^2\right] \to 0 \quad \text{as } k\to\infty. \tag{3.5}$$

Since an L^2-convergent sequence of random variables has an a.s. convergent subsequence, and since $C([-K,K]^N)$ is separable, one can complete the proof by the diagonal method. □

When it is not confusing, we write $\overline{Y}_{1,m_k} = \overline{Y}_{m_k}$ and $Y_{1,m_k} = Y_{m_k}$ on $(\Omega_1, \mathbf{B}_1, P_1)$ for the sake of simplicity.

Take $x_0 \in D$ and $r_0 > 0$ so that $U_{4r_0}(x_0) := \{y \in \mathbf{R}^N : |x_0 - y| < 4r_0\} \subset D$, and put $U_0 := U_{2r_0}(x_0)$. Then

$$V_0 := \inf_{x\in\partial U_0} \mathrm{Vol}(U_{3r_0}(x_0) \cap H(x_0 - x, x)) > 0. \tag{3.6}$$

It is easy to see that $V_0 = \mathrm{Vol}(U_{3r_0}(x_0) \cap H(p, x_0 - 2r_0 p))$ for any $p \in \mathbf{S}^{N-1}$.

Put also, on $(\Omega_1, \mathbf{B}_1, P_1)$,

$$\tau_m := \inf\{t > 0 | Y_{1,m}(t,z) = 0 \text{ for some } z \in (\mathbf{Z}^N/m) \cap U_0\}. \tag{3.7}$$

Then for any $k \geq 1$, there exists a random point $z_{m_k} \in (\mathbf{Z}^N/m) \cap U_0$ such that

$$Y_{1,m_k}(\tau_{m_k}-, z_{m_k}) - Y_{1,m_k}(\tau_{m_k}, z_{m_k}) = 1.$$

Additionally, the following holds:

$$Y_{1,m_k}(0,z) - Y_{1,m_k}(\tau_{m_k}, z) = 1 \quad \text{for all } z \in D_{m_k} \cap H(x_0 - z_{m_k}, z_{m_k}).$$

Since $U_{3r_0}(x_0) \subset D$, $Y_{1,m_k}(0,z) - Y_{1,m_k}(\tau_{m_k}, z) = 0$ or 1, and

$$\sum_{z\in D_{m_k}} (Y_{1,m_k}(0,z) - Y_{1,m_k}(\tau_{m_k}, z))\left(\frac{1}{m_k}\right)^N \sim \tau_{m_k}$$

for sufficiently large $k \geq 1$ from (3.4), the following holds.

LEMMA 3.3. *Suppose that* (A.0)–(A.1) *hold. Then*

$$P_1\left(V_0 \leq \liminf_{k\to\infty} \tau_{m_k} \leq \limsup_{k\to\infty} \tau_{m_k} \leq \mathrm{Vol}(D)\right) = 1. \tag{3.8}$$

Proof. By (3.4), for any $t > \mathrm{Vol}(D)$,

$$\begin{aligned}(3.9)\qquad &\limsup_{k\to\infty}\{\min(\tau_{m_k}, t)\}\\ &= \limsup_{k\to\infty}\int_0^{\min(\tau_{m_k},t)} \omega_1(Y_{m_k}(s,\cdot), D_{m_k})ds\\ &\le \limsup_{k\to\infty}\sum_{z\in D_{m_k}}(Y_{m_k}(0,z) - Y_{m_k}(\min(\tau_{m_k},t),z))\left(\frac{1}{m_k}\right)^N \le \mathrm{Vol}(D)\end{aligned}$$

P_1-a.s. We also have

$$\begin{aligned}(3.10)\qquad V_0 &\le \liminf_{k\to\infty}\sum_{z\in D_{m_k}}(Y_{m_k}(0,z) - Y_{m_k}(\tau_{m_k},z))\left(\frac{1}{m_k}\right)^N\\ &\le \liminf_{k\to\infty}\int_0^{\tau_{m_k}} \omega_1(Y_{m_k}(s,\cdot), D_{m_k})ds = \liminf_{k\to\infty}\tau_{m_k} \quad P_1\text{-a.s.} \qquad \square\end{aligned}$$

The following lemma can be proved in the same way as in [4, section 5.2], and the proof is omitted.

LEMMA 3.4. *Suppose that* (A.1) *holds. Let F and $F_m (m \ge 1)$ be closed convex sets in $\mathbf{R}^N$ such that ∂F and $\partial F_m (m \ge 1)$ are closed hypersurfaces and such that $d_H(F_m, F) \to 0$ as $m \to \infty$. Then $\omega_1(I_{F_m}(\cdot), dx)$ weakly converges to $\omega_1(I_F(\cdot), dx)$ as $m \to \infty$, that is, for any $\varphi \in C_o(\mathbf{R}^N)$,*

$$(3.11)\qquad \lim_{m\to\infty}\int_{\mathbf{R}^N}\varphi(x)\omega_1(I_{F_m}(\cdot), dx) = \int_{\mathbf{R}^N}\varphi(x)\omega_1(I_F(\cdot), dx).$$

We denote by $X(\cdot,\cdot) \in C([0,\infty) : \mathcal{S})$ the P_1-a.s. limit of $\overline{Y}_{1,m_k}(\cdot,\cdot)$ as $k \to \infty$ (see Lemma 3.2 for notation). Then we have the following.

LEMMA 3.5. *Suppose that* (A.0)–(A.1) *hold. Then there exists a solution $\{D(t)\}_{t\in[0,V_0)}$ to* (1.13)–(1.14) *on $[0, V_0)$ such that the following holds P_1-a.s.:*

$$(3.12)\qquad X(t,x) = I_{D(t)}(x), \quad dx\text{-a.e.} \quad \text{for all } t \in [0, V_0)$$

(see (3.6) *for notation).*

Proof. We introduce local coordinates so that we can reduce the study of the time evolution of $\partial(\mathrm{co}\, Y_{m_k}(t,\cdot)^{-1}(1))$ on $[0, V_0)$ to that of convex functions in all local coordinates.

For (x_0, r_0) in (3.6) and any $p \in \mathbf{S}^{N-1}$, let $C(x_0, r_0; p)$ denote a semi-infinite cylinder

$$\{x_0 + rp + x : r \ge 0, |x| \le r_0, \langle x, p\rangle = 0, x \in \mathbf{R}^N\}$$

which can be obtained by moving an $(N-1)$-dimensional ball

$$\{x_0 + x : |x| \le r_0, \langle x, p\rangle = 0, x \in \mathbf{R}^N\}$$

in the positive direction of p.

Take $p_1, \ldots, p_{k_0} \in \mathbf{S}^{N-1}$ for some $k_0 \in \mathbf{N}$ so that

$$\mathrm{co}\ D \subset \cup_{i=1}^{k_0} C(x_0, r_0; p_i).$$

For $i = 1, \dots, k_0$, take $\{q_{i1}, \dots, q_{i(N-1)}\}$ so that $\{q_{i1}, \dots, q_{i(N-1)}, p_i\}$ is an orthonormal basis in $\mathbf{R}^N$, and put

$$C_{m_k}(t) := \text{co } Y_{m_k}(t, \cdot)^{-1}(1). \tag{3.13}$$

For $x = (x_j)_{j=1}^{N-1} \in \mathbf{R}^{N-1}$ for which $|x| \le 2r_0$, also put

$$\tilde{X}_{m_k,i}(t,x) := -\sup\left\{ r > 0 | x_0 + rp_i + \sum_{j=1}^{N-1} q_{ij}x_j \in C_{m_k}(t) \right\}. \tag{3.14}$$

Since the set $C_{m_k}(t)$ is convex and since

$$\inf\left\{ \left(\sum_{j=1}^{N-1} |x_j|^2 \right)^{1/2} : x_0 + rp_i + \sum_{j=1}^{N-1} q_{ij}x_j \in C_{m_k}(t) \text{ for some } r > 0 \right\} \ge \frac{7r_0}{4}$$

for $m_k \ge 8\sqrt{N}/r_0$ and $t \in [0, \tau_{m_k})$, and since $t \mapsto Y_{m_k}(t, \cdot)$ is nonincreasing, $\tilde{X}_{m_k,i}(t, \cdot)$ ($m_k \ge 8\sqrt{N}/r_0$, $t \in [0, \tau_{m_k})$) are bounded convex functions on $\{x \in \mathbf{R}^{N-1} : |x| \le 7r_0/4\}$.

It is known that a uniformly bounded set of convex functions with a common domain is compact as the set of continuous functions defined on K for every compact subset K of the interior of their domain (see [4, section 3.3]).

Therefore, by Lemma 3.3 and the diagonal method, there exists a subsequence $\{\tilde{X}_{\tilde{m}_k,i}(t, \cdot)\}_{k\ge1}$ of $\{\tilde{X}_{m_k,i}(t, \cdot)\}_{k\ge1}$ and a convex function $\tilde{X}_i(t, \cdot)$ such that for any $t \in \mathbf{Q} \cap [0, V_0)$ and $i = 1, \dots, k_0$,

$$\lim_{k\to\infty} \sup_{x\in\mathbf{R}^{N-1}, |x|\le 3r_0/2} |\tilde{X}_{\tilde{m}_k,i}(t,x) - \tilde{X}_i(t,x)| = 0. \tag{3.15}$$

(Notice that $\{\tilde{m}_k\}_{k\ge1}$ can be random.)

It is clear that there exists a nonincreasing family of compact convex sets $\{\tilde{C}(t)\}_{t\in\mathbf{Q}\cap[0,V_0)}$ such that for any $t \in \mathbf{Q} \cap [0, V_0)$,

$$\lim_{k\to\infty} d_H(C_{\tilde{m}_k}(t), \tilde{C}(t)) = 0, \tag{3.16}$$

$$\tilde{X}_i(t,x) = -\sup\left\{ r > 0 | x_0 + rp_i + \sum_{j=1}^{N-1} q_{ij}x_j \in \tilde{C}(t) \right\}$$

for all $i = 1, \dots, k_0$, and $x = (x_k)_{k=1}^{N-1} \in \mathbf{R}^{N-1}$ for which $|x| \le 3r_0/2$.

In particular,

$$D \subset \tilde{C}(0) \quad \text{(by (2.2))}, \tag{3.17}$$

$$\lim_{k\to\infty} ||X_{1,\tilde{m}_k}(t, \cdot) - I_{\tilde{C}(t)^o \cap D}(\cdot)||_{L^2([-K,K]^N)} = 0$$

for all $t \in \mathbf{Q} \cap [0, V_0)$, where $X_{1,\tilde{m}_k}$ is defined by $Y_{1,\tilde{m}_k}$ in the same way as in (2.6). When it is not confusing, we write $X_{1,\tilde{m}_k} = X_{\tilde{m}_k}$ on $(\Omega_1, \mathbf{B}_1, P_1)$ for the sake of simplicity.

The following also holds: for all $t \in [0, V_0) \cap \mathbf{Q}$,

$$\lim_{k\to\infty} ||\overline{Y}_{\tilde{m}_k}(t, \cdot) - X_{\tilde{m}_k}(t, \cdot)||_{L^2([-K,K]^N)} = 0. \tag{3.18}$$

Indeed, if $X_m(t,x) \neq \overline{Y}_m(t,x)$, then

$$dist(x, \partial(C_m(t)^o \cap D)) \leq \frac{N^{1/2}}{m},$$

and by (3.16), the volume of the $(N^{1/2}/\tilde{m}_k)$-neighborhood of the set $\partial D \cup \partial C_{\tilde{m}_k}(t)$ converges to zero as $k \to \infty$ for $t \in [0, V_0) \cap \mathbf{Q}$.

For $t \in [0, V_0)\backslash\mathbf{Q}$, put

$$\tilde{C}(t) := \cap_{s \in \mathbf{Q} \cap [0,t)} \tilde{C}(s). \tag{3.19}$$

Then, by (3.17)–(3.19), the following holds P_1-a.s.:

$$X(t,x) = I_{\tilde{C}(t)^o \cap D}(x), \quad dx\text{-a.e.}, \quad \text{for all } t \in [0, V_0), \tag{3.20}$$

since $\{\overline{Y}_{\tilde{m}_k}\}_{k \geq 1}$ is a subsequence of a convergent sequence $\{\overline{Y}_{m_k}\}_{k \geq 1}$ and since $X \in C([0,\infty) : \mathcal{S})$ is the P_1-a.s. limit, in $D([0,\infty) : \mathcal{S})$, of $\overline{Y}_{m_k}$ as $k \to \infty$, and since $\{\tilde{C}(t)\}_{t \in [0,V_0) \cap \mathbf{Q}}$ is nonincreasing in t.

Put

$$D(t) := \tilde{C}(t)^o \cap D. \tag{3.21}$$

Then (1.13) holds for all $t \in [0, V_0)$, since $D = D(0)$ by (3.17) and since

$$D(t) \supset \{\text{co } (\tilde{C}(t)^o \cap D)\} \cap D = (\text{co } D(t)) \cap D \supset D(t) \cap D = D(t).$$

On $[0, V_0)$,

$$\omega_1(I_{\tilde{C}(t)}(\cdot), dx) = \omega_1(I_{D(t)}(\cdot), dx), \quad dt\text{-a.e.}, \tag{3.22}$$

since

$$\tilde{C}(t)\backslash(\text{co } D(t))^- \subset \tilde{C}(t)\backslash D(t)^- \subset D^c$$

by (3.21), where D^c denotes a complement of D, and since

$$\int_0^{V_0} \omega_1(I_{\tilde{C}(s)}(\cdot), D^c) ds = \int_{D^c} (I_{D(0)}(x) - I_{D(V_0)}(x)) dx = 0$$

by (3.4), (3.20)–(3.21), and Lemma 3.4.

Here we used the fact that (3.16) holds except for at most countably many $t \in [0, V_0)$. Indeed, $t \mapsto C_{\tilde{m}_k}(t)$ is nonincreasing and (3.16) holds for all $t \in \mathbf{Q} \cap [0, V_0)$. Therefore, if $C_{\tilde{m}_k}(t)$ does not converge to $\tilde{C}(t)$ as $k \to \infty$, then $(\tilde{C}(t)\backslash\tilde{C}(t+))^o$ is not empty and has a positive Lebesgue measure by (3.19), where $\tilde{C}(t+) := \cup_{s>t} \tilde{C}(s)$. Besides, $(\tilde{C}(t)\backslash\tilde{C}(t+))^o$ are disjoint for different t.

Hence, (1.14) holds for all $t \in [0, V_0)$ from (3.4), Lemma 3.4, and (3.20)–(3.22). □

The following lemma, which can be proved by the translation invariance of the solution to (1.13)–(1.14), implies the uniqueness of the solution to (1.13)–(1.14).

LEMMA 3.6. *Suppose that* (A.1) *holds. For* $T > 0$, *if* $\{D_i(t)\}_{0 \leq t < T}$ $(i = 1, 2)$ *are solutions to* (1.13)–(1.14) *on* $[0, T)$ *for which* $D_1(0) \subset D_2(0)$ *and if* (A.0) *holds for* $D = D_i(0)$ $(i = 1, 2)$, *then* $D_1(t) \subset D_2(t)$ *for all* $t \in [0, T)$. *In particular, for all* $t \in [0, \min(\text{Vol}(D_1(0)), T))$,

$$dist(D_1(t), D_2(t)^c) \geq dist(D_1(0), D_2(0)^c). \tag{3.23}$$

Proof. For each $t \geq 0$, put

$$\tilde{D}(t) := D_1(t)^- \cap D_2(t)^c, \quad u_i(t,\cdot) := I_{D_i(t)}(\cdot), \quad u_i^-(t,\cdot) := I_{D_i(t)^-}(\cdot),$$

$$N_i(t) := \cup_{x \in \partial \tilde{D}(t) \cap \partial D_i(t)} \{p \in S^{N-1} | \sigma^+(u_i, -p, t, x) = 1\}$$

($i = 1, 2$). Then $N_2(t) \subset N_1(t)$.

Take a nondecreasing sequence $\{\eta_m\}_{m \geq 1}$ of nondecreasing C^1-functions such that

$$\eta_m(r) = 0 \quad \text{for all } r \leq 0, \quad \eta_m(r) = 1 \quad \text{for all } r \geq \frac{1}{m}, \tag{3.24}$$

and for $r \in \mathbf{R}$, put

$$\zeta_m(r) = \int_0^r \eta_m(s) ds. \tag{3.25}$$

Then since $t \mapsto u_i(t,x)$ and $t \mapsto u_i^-(t,x)$ are, respectively, right and left continuous for any $x \in \mathbf{R}^N$, for $t \in [0, \min(\mathrm{Vol}(D_1(0)), T))$ and $x \in \mathbf{R}^N$,

$$\begin{aligned} &\zeta_m(u_1^-(t,x) - u_2(t,x) - 1) - \zeta_m(u_1^-(0,x) - u_2(0,x)) \\ &= \int_0^t \zeta_m(u_1^-(s,x) - u_2(s,x) - s/t)(u_1^-(ds,x) - u_2(ds,x)) \\ &\qquad -\frac{1}{t}\int_0^t \eta_m(u_1^-(s,x) - u_2(s,x) - s/t) ds. \end{aligned} \tag{3.26}$$

Since $\zeta_m \geq 0$, $D_1(0) \subset D_2(0)$, and $N_2(s) \subset N_1(s)$, we have

$$\begin{aligned} 0 \leq &\int_0^t ds \int_{\mathbf{R}^N} \zeta_m(u_1^-(s,x) - u_2(s,x) - s/t) \\ &\qquad \times (\omega_1(u_2(s,\cdot), dx) - \omega_1(u_1(s,\cdot), dx)) \\ &\qquad -\frac{1}{t}\int_0^t ds \int_{\mathbf{R}^N} \eta_m(u_1^-(s,x) - u_2(s,x) - s/t) dx \\ \to &\int_0^t (1 - s/t)(\omega_1(u_2(s,\cdot), \tilde{D}(s)) - \omega_1(u_1(s,\cdot), \tilde{D}(s))) ds \\ &\qquad -\frac{1}{t}\int_0^t ds \int_{\tilde{D}(s)} dx \quad (\text{as } m \to \infty) \\ \leq &-\frac{1}{t}\int_0^t ds \int_{\tilde{D}(s)} dx, \end{aligned} \tag{3.27}$$

which implies the first assertion of this lemma.

Suppose that (3.23) does not hold. Then there exists $a \in (0, dist(D_1(0), D_2(0)^c))$ such that

$$\inf\{dist(D_1(t), D_2(t)^c) | t \in [0, \min(\mathrm{Vol}(D_1(0)), T))\} < a.$$

Take $p_a \in \mathbf{S}^{N-1}$ and $t_a \in [0, \min(\mathrm{Vol}(D_1(0)), T))$ so that

$$ap_a + D_1(t_a) \not\subset D_2(t_a).$$

Since $ap_a + D_1(0) \subset D_2(0)$ and $\{ap_a + D_1(t)\}_{0\le t<T}$ is a solution to (1.13)–(1.14) on $[0,T)$, this contradicts the first assertion of this lemma. □

Take $\varphi \in C^2(\mathbf{R}^N)$ for which $D\varphi(x_o) \neq 0$ for some $x_o \in \mathbf{R}^N$. Put

$$f_N := \frac{D\varphi(x_o)}{|D\varphi(x_o)|}, \quad (g_1 \cdots g_N) := I_N - f_N \otimes f_N. \tag{3.28}$$

Take $\{f_1, \ldots, f_{N-1}\}$ so that $\{f_1, \ldots, f_N\}$ is an orthonormal basis of $\mathbf{R}^N$. Then the following holds.

LEMMA 3.7.

(i) $\langle g_i, f_N \rangle = 0$ $(1 \le i \le N)$.

(ii) *For i for which $\partial_i\varphi(x_o) := \partial\varphi(x_o)/\partial x_i \neq 0$,*

$$g_i = -\sum_{k\neq i} \frac{\partial_k\varphi(x_o)}{\partial_i\varphi(x_o)} g_k.$$

(iii) $\mathrm{span}(g_1, \ldots, g_N) = \mathrm{span}(f_1, \ldots, f_{N-1})$.

(iv) $D(D\varphi(x_o)/|D\varphi(x_o)|)(\mathbf{R}^N) \subset \mathrm{span}(g_1, \ldots, g_N)$. *As a mapping on* $\mathrm{span}(g_1, \ldots, g_N)$, *eigenvalues and eigenvectors of* $(g_1 \cdots g_N)(D^2\varphi(x_o)/|D\varphi(x_o)|)(g_1 \cdots g_N)$ *are the same as those of* $D(D\varphi(x_o)/|D\varphi(x_o)|)$. *In particular, all eigenvalues of* $D(D\varphi(x_o)/|D\varphi(x_o)|)$ *are real.*

(v) *If eigenvalues* $\lambda_1 \le \cdots \le \lambda_{N-1}$ *of* $-D(D\varphi(x_o)/|D\varphi(x_o)|)$ *as a mapping on* $\mathrm{span}(g_1, \ldots, g_N)$ *are nonnegative, then*

$$\prod_{i=1}^{N-1} \lambda_i = \frac{G(D\varphi(x_o), D^2\varphi(x_o))}{|D\varphi(x_o)|}. \tag{3.29}$$

Proof. It is easy to see that (i) and (ii) hold. Take i for which $\partial_i\varphi(x_o) \neq 0$. Then, by (i) and (ii), we have only to show, to prove (iii), that $\{g_j\}_{j\neq i}$ is independent. Suppose that for $j = 1, \ldots, N$,

$$\sum_{k\neq i} \lambda_k \left(\delta_{kj} - \frac{\partial_k\varphi(x_o)\partial_j\varphi(x_o)}{|D\varphi(x_o)|^2} \right) = 0. \tag{3.30}$$

Putting $j = i$ in (3.30), we obtain

$$\sum_{k\neq i} \lambda_k \frac{\partial_k\varphi(x_o)\partial_i\varphi(x_o)}{|D\varphi(x_o)|^2} = 0,$$

from which

$$\sum_{k\neq i} \lambda_k \partial_k\varphi(x_o) = 0. \tag{3.31}$$

Putting $j \neq i$ in (3.30), we obtain

$$\lambda_j - \partial_j\varphi(x_o) \sum_{k\neq i} \lambda_k \frac{\partial_k\varphi(x_o)}{|D\varphi(x_o)|^2} = 0,$$

from which $\lambda_j = 0$ for $j \neq i$, by (3.31).

We prove (iv). It is easy to see that

$$D\left(\frac{D\varphi(x_o)}{|D\varphi(x_o)|}\right) = (g_1 \cdots g_N)\frac{D^2\varphi(x_o)}{|D\varphi(x_o)|}. \tag{3.32}$$

Hence

$$D\left(\frac{D\varphi(x_o)}{|D\varphi(x_o)|}\right)\left(\sum_{i=1}^N x_i g_i\right) = \lambda \sum_{i=1}^N x_i g_i$$

if and only if

$$(g_1 \cdots g_N)\frac{D^2\varphi(x_o)}{|D\varphi(x_o)|}(g_1 \cdots g_N)\left(\sum_{i=1}^N x_i g_i\right) = \lambda \sum_{i=1}^N x_i g_i,$$

since

$$(g_1 \cdots g_N)^2 = (g_1 \cdots g_N). \tag{3.33}$$

Put $P := (f_1 \cdots f_N)$ and $Q := (f_1 \cdots f_{N-1})$, and let Q^* and o_{N-1} denote the transposed matrix of Q and the $(N-1)$-dimensional zero vector, respectively. The proof of (v) is divided into the following.

Step I. The eigenvalues of

$$-(I_N - f_N \otimes f_N)\frac{D^2\varphi(x_o)}{|D\varphi(x_o)|}(I_N - f_N \otimes f_N) + f_N \otimes f_N$$

are those of

$$\begin{pmatrix} -Q^* D(\frac{D\varphi(x_o)}{|D\varphi(x_o)|})Q & o_{N-1} \\ o_{N-1}^* & 1 \end{pmatrix}.$$

Step II. The eigenvalues of $Q^* D(D\varphi(x_o)/|D\varphi(x_o)|)Q$ are those of $D(D\varphi(x_o)/|D\varphi(x_o)|)$ on $\mathrm{span}(g_1, \ldots, g_N)$.

Proof of Step I. Denote by O_{N-1} the $(N-1)\times(N-1)$-zero matrix. For $\lambda \in \mathbf{R}$,

$$\begin{aligned}
&\det\left(-(I_N - f_N \otimes f_N)\frac{D^2\varphi(x_o)}{|D\varphi(x_o)|}(I_N - f_N \otimes f_N) + f_N \otimes f_N - \lambda I_N\right) \\
&= \det\left(-\begin{pmatrix} I_{N-1} & o_{N-1} \\ o_{N-1}^* & 0 \end{pmatrix} P^* \frac{D^2\varphi(x_o)}{|D\varphi(x_o)|} P \begin{pmatrix} I_{N-1} & o_{N-1} \\ o_{N-1}^* & 0 \end{pmatrix}\right. \\
&\qquad \left. + \begin{pmatrix} O_{N-1} & o_{N-1} \\ o_{N-1}^* & 1 \end{pmatrix} - \lambda I_N\right) \\
&= \det\left(\begin{pmatrix} -Q^* \frac{D^2\varphi(x_o)}{|D\varphi(x_o)|} Q & o_{N-1} \\ o_{N-1}^* & 1 \end{pmatrix} - \lambda I_N\right)
\end{aligned}$$

since

$$P^*P = I_N, \quad P\begin{pmatrix} O_{N-1} & o_{N-1} \\ o_{N-1}^* & 1 \end{pmatrix} P^* = f_N \otimes f_N.$$

Equations (3.28) and (3.32) complete the proof since $\langle f_i, f_N \rangle = 0$ if $i \neq N$.

Proof of Step II. Let $x = (x_i)_{i=1}^{N-1} \in \mathbf{R}^{N-1}$ and $\lambda \in \mathbf{R}$. Suppose that

$$Q^* D\left(\frac{D\varphi(x_o)}{|D\varphi(x_o)|}\right) Qx = \lambda x. \tag{3.34}$$

Then

$$QQ^* D\left(\frac{D\varphi(x_o)}{|D\varphi(x_o)|}\right)\left(\sum_{1\le i\le N-1} x_i f_i\right) = \lambda \sum_{1\le i\le N-1} x_i f_i.$$

Hence, by (3.32),

$$D\left(\frac{D\varphi(x_o)}{|D\varphi(x_o)|}\right) \sum_{1\le i\le N-1} x_i f_i = \lambda \sum_{1\le i\le N-1} x_i f_i \tag{3.35}$$

since, by (iii),

$$QQ^*(I_N - f_N \otimes f_N) = I_N - f_N \otimes f_N.$$

It is easy to see that (3.35) implies (3.34) since $Q^*Q = I_{N-1}$. □

For $i = 1, \dots, N$, put

$$y_i(x) := \left((\delta_{ij} - 1)\frac{\partial_j \varphi(x)}{|D\varphi(x)|} + \delta_{ij}\varphi(x)\right)_{j=1}^{N}.$$

Then we have the following.

LEMMA 3.8. *Suppose that all eigenvalues of $D(D\varphi(x_o)/|D\varphi(x_o)|)$ are nonpositive. Then, for $i = 1, \dots, N$,*

$$\frac{\partial_i \varphi(x_o)}{|D\varphi(x_o)|} G(D\varphi(x_o), D^2\varphi(x_o)) = \det(Dy_i(x_o)). \tag{3.36}$$

Proof. For the sake of simplicity, we assume that $i = N$.

We first consider the case when $\partial_N \varphi(x_o) \neq 0$. By (ii) in Lemma 3.7, it is easy to see that the following holds:

$$\begin{aligned} A &:= \begin{pmatrix} I_{N-1} & o_{N-1} \\ -\frac{D\varphi(x_o)^*}{\partial_N \varphi(x_o)} & \end{pmatrix} Dy_N(x_o) \\ &= D\left(-\frac{D\varphi(x_o)}{|D\varphi(x_o)|}\right) + \begin{pmatrix} O_{N-1} & o_{N-1} \\ -D\varphi(x_o)^* & \end{pmatrix}. \end{aligned} \tag{3.37}$$

By Lemma 3.7 (i) and (iv), the eigenvalues and eigenvectors of $D(-D\varphi(x_o)/|D\varphi(x_o)|)$ on $\operatorname{span}(g_1, \dots, g_N)$ are real and are also those of the matrix A.

Therefore, by (iii) in Lemma 3.7, the matrix A has a real eigenvalue λ and a real eigenvector $x_\lambda \notin \operatorname{span}(g_1, \dots, g_N)$. Indeed, by (3.32),

$$A\mathbf{e}_N \in (-\partial_N \varphi(x_o))\mathbf{e}_N + \operatorname{span}(g_1, \dots, g_N),$$

where $\mathbf{e}_N := (\delta_{jN})_{j=1}^N$. Besides, $\mathbf{e}_N \notin \operatorname{span}(g_1, \dots, g_N)$ since $\partial_N \varphi(x_o) \neq 0$.

We show that $\lambda = -\partial_N \varphi(x_o)$. Put

$$x_\lambda := x_g + a\mathbf{e}_N \quad (x_g \in \operatorname{span}(g_1, \dots, g_N), a \neq 0), \tag{3.38}$$

which is possible since $\mathrm{span}(\mathbf{e}_N, g_1, \dots, g_N) = \mathbf{R}^N$ by (iii) in Lemma 3.7. Then

$$(-\partial_N \varphi(x_o))a\mathbf{e}_N = \lambda a \mathbf{e}_N \tag{3.39}$$

by (i) in Lemma 3.7 since $Ax_\lambda = \lambda x_\lambda$ and since $\mathbf{e}_N$ and $\{g_1, \dots, g_N\}$ are independent. This implies that $\lambda = -\partial_N \varphi(x_o)$.

Suppose that $\partial_N \varphi(x_o) = 0$. Then, by (3.32) and (i) in Lemma 3.7, for $x \in \mathbf{R}^N$,

$$\langle f_N, Dy_N(x_o)x\rangle = \left\langle f_N, D\left(-\frac{D\varphi(x_o)}{|D\varphi(x_o)|}\right)x\right\rangle = 0. \tag{3.40}$$

Hence $Dy_N(x_o)(\mathbf{R}^N)$ is at most $(N-1)$-dimensional and (3.36) holds. □

4. Proofs. In this section we prove the results in section 2.

Proof of Theorem 2.1. By Lemmas 3.1–3.6, there exists a unique (nonrandom) solution $\{D(t)\}_{0\le t<V_0}$ (see (3.6) for notation) of (1.13)–(1.14) on $[0, V_0)$ such that $I_{D(\cdot)} \in C([0, V_0) : \mathcal{S})$ and such that the following holds: for any $T \in [0, V_0)$ and $\gamma > 0$,

$$\lim_{m\to\infty} P\left(\sup_{0\le t\le T} d_{\mathcal{S}}(\overline{Y}_m(t,\cdot), I_{D(t)}(\cdot)) \ge \gamma\right) = 0. \tag{4.1}$$

Therefore

$$\lim_{m\to\infty} P\left(\sup_{0\le t\le T} ||\overline{Y}_m(t,\cdot) - I_{D(t)}(\cdot)||_{L^2([-K,K]^N)} \ge \gamma\right) = 0, \tag{4.2}$$

since, for $m \ge 1$ and $t \in [0, T]$,

$$\begin{aligned} &||\overline{Y}_m(t,\cdot) - I_{D(t)}(\cdot)||^2_{L^2([-K,K]^N)} \\ &= \int_{[-K,K]^N} (\overline{Y}_m(t,x) - 2\overline{Y}_m(t,x)I_{D(t)}(x) + I_{D(t)}(x))dx. \end{aligned}$$

We prove that the following holds:

$$\lim_{m\to\infty} P\left(\sup_{0\le t\le T} ||X_m(t,\cdot) - I_{D(t)}(\cdot)||_{L^2([-K,K]^N)} \ge \gamma\right) = 0. \tag{4.3}$$

For any s and t for which $0 \le s < t \le T$,

$$\begin{aligned} &||X_m(t,\cdot) - I_{D(t)}(\cdot)||_{L^2([-K,K]^N)} \\ &\le ||X_m(t,\cdot) - X_m(s,\cdot)||_{L^2([-K,K]^N)} + ||X_m(s,\cdot) - \overline{Y}_m(s,\cdot)||_{L^2([-K,K]^N)} \\ &\quad + ||\overline{Y}_m(s,\cdot) - I_{D(s)}(\cdot)||_{L^2([-K,K]^N)} + ||I_{D(s)}(\cdot) - I_{D(t)}(\cdot)||_{L^2([-K,K]^N)}. \end{aligned} \tag{4.4}$$

Let $U_{-N^{1/2}/m}(D) := \{x \in D | dist(x, D^c) > N^{1/2}/m\}$. Then

$$\begin{aligned} &||X_m(t,\cdot) - X_m(s,\cdot)||^2_{L^2([-K,K]^N)} = ||X_m(t,\cdot) - X_m(s,\cdot)||_{L^1([-K,K]^N)} \\ &\le 2^N \sum_{z\in D_m} (Y_m(s,z) - Y_m(t,z))\frac{1}{m^N} + \mathrm{Vol}(D\backslash U_{-N^{1/2}/m}(D)) \end{aligned} \tag{4.5}$$

(see (2.2) for notation). Indeed, if $x = (x_i)_{i=1}^N \in U_{-N^{1/2}/m}(D)\backslash(\mathrm{co}\ Y_m(t,\cdot)^{-1}(1))$, then $Y_m(t,z) = 0$ for some $z = (z_i)_{i=1}^N \in \mathbf{Z}^N/m$ for which $|x_i - z_i| \le 1/m$ for all $i = 1, \dots, N$.

In the same way as in (3.5), by (4.5), for any $\gamma > 0$, there exists $\delta > 0$ such that the following holds: for any $s \in [0, T - \delta]$,

$$\lim_{m\to\infty} P\left(\sup_{s\le s_1\le s+\delta} ||X_m(s_1,\cdot) - X_m(s,\cdot)||_{L^2([-K,K]^N)} \ge \gamma\right) = 0. \tag{4.6}$$

Since, for any $t \in [0, V_0)$, any subsequence of $\{C_m(t)\}_{m\ge1}$ has a convergent subsequence (see (3.13)–(3.16)),

$$\lim_{m\to\infty} ||\overline{Y}_m(t,\cdot) - X_m(t,\cdot)||_{L^2([-K,K]^N)} = 0 \tag{4.7}$$

for all $t \in [0, V_0)$, P_1-a.s. (see the discussion after (3.18)). Hence, for any $\gamma > 0$,

$$\lim_{m\to\infty} P(||\overline{Y}_m(s,\cdot) - X_m(s,\cdot)||_{L^2([-K,K]^N)} \ge \gamma) = 0. \tag{4.8}$$

$I_{D(\cdot)} \in C([0, V_0) : L^2([-K, K]^N))$ since

$$||I_{D(s)}(\cdot) - I_{D(t)}(\cdot)||^2_{L^2([-K,K]^N)} = \int_{[-K,K]^N} I_{D(s)}(x)dx - \int_{[-K,K]^N} I_{D(t)}(x)dx$$

and since $t \mapsto \int_{[-K,K]^N} I_{D(t)}(x)dx$ is continuous on $[0, V_0)$.

Equation (4.2) and the discussion after (4.3) show that (4.3) is true.

Recall Lemmas 3.2 and 3.3 and the notation therein. For $T < V_0$, take $x_0 \in D(T)$ and r_0 so that $U_{4r_0}(x_0) \subset D(T)$. For sufficiently large $k \ge 1$,

$$U_{3r_0}(x_0) \subset (\text{co } Y_{m_k}(T,\cdot)^{-1}(1))^o \cap D, \quad P_1\text{-a.s.},$$

since

$$\lim_{k\to\infty} ||X_{m_k}(T,\cdot) - I_{D(T)}(\cdot)||_{L^2([-K,K]^N)} = 0, \quad P_1\text{-a.s.}$$

by Lemma 3.2 and (4.7) (see the discussion below (4.2)). Hence in the same way as in Lemma 3.3,

$$\begin{aligned} V_0 &\le \liminf_{k\to\infty} \sum_{z\in D_{m_k}} (Y_{m_k}(T,z) - Y_{m_k}(\tau_{m_k},z))\frac{1}{m_k^N} \\ &\le \liminf_{k\to\infty} (\tau_{m_k} - T) \quad P_1\text{-a.s.}, \end{aligned} \tag{4.9}$$

which implies that (4.3) holds for $T < 2V_0$. Repeating the same procedure as above and then letting $r_0 \downarrow 0$, (4.3) holds for all $T < T^* := \text{Vol}(D)$.

Put

$$D(t) = \emptyset \quad \text{for } t \ge T^*. \tag{4.10}$$

Then $I_{D(\cdot)} \in C([0,\infty) : L^2([-K, K]^N))$ and $\{D(t)\}_{t\ge0}$ is a unique solution to (1.13)–(1.14) on $[0,\infty)$ by Lemma 3.6, since $t \mapsto I_{D(t)}$ is nonincreasing and since

$$\text{Vol}(D(t)) = \text{Vol}(D(0)) - t \downarrow 0 \quad \text{as } t \uparrow T^* \tag{4.11}$$

by (1.14).

We prove (2.9). By (4.11), for $\gamma > 0$,

$$\text{Vol}(D(t)) \le \left(\frac{\gamma}{4}\right)^2 \quad \text{for } t \ge t_\gamma := T^* - \left(\frac{\gamma}{4}\right)^2. \tag{4.12}$$

Therefore

$$\begin{aligned} &P\left(\sup_{t\ge 0} ||X_m(t,\cdot) - I_{D(t)}(\cdot)||_{L^2([-K,K]^N)} \ge \gamma\right) \\ &\le P\left(\sup_{0\le t\le t_\gamma} ||X_m(t,\cdot) - I_{D(t)}(\cdot)||_{L^2([-K,K]^N)} \ge \gamma\right) \\ &\qquad + P\left(\sup_{t\ge t_\gamma} ||X_m(t,\cdot) - I_{D(t)}(\cdot)||_{L^2([-K,K]^N)} \ge \gamma\right) \\ &\le 2P\left(\sup_{0\le t\le t_\gamma} ||X_m(t,\cdot) - I_{D(t)}(\cdot)||_{L^2([-K,K]^N)} \ge \frac{\gamma}{2}\right) \to 0 \quad (\text{as } m\to\infty) \end{aligned} \tag{4.13}$$

since for $t \ge t_\gamma$,

$$\begin{aligned} &||X_m(t,\cdot) - I_{D(t)}(\cdot)||_{L^2([-K,K]^N)} \\ &\le ||X_m(t_\gamma,\cdot)||_{L^2([-K,K]^N)} + ||I_{D(t_\gamma)}(\cdot)||_{L^2([-K,K]^N)} \\ &\le ||X_m(t_\gamma,\cdot) - I_{D(t_\gamma)}(\cdot)||_{L^2([-K,K]^N)} + 2||I_{D(t_\gamma)}(\cdot)||_{L^2([-K,K]^N)}. \end{aligned}$$ □

Proof of Corollary 2.2. Since D is convex,

$$(\text{co } Y_m(t,\cdot)^{-1}(1))^o \cap D = (\text{co } Y_m(t,\cdot)^{-1}(1))^o =: D_m(t).$$

For $T < T^*(= \text{Vol}(D))$, take $x_0 \in D(T)$ and r_0 so that $U_{4r_0}(x_0) \subset D(T)$ (see (3.6) for notation). Then, for sufficiently large m, $U_{3r_0}(x_0) \subset D_m(0)$.

Consider cones

$$\text{cone}(x) := \text{co } (\{x\} \cup U_0^-) \quad (x \in D^-)$$

(see (3.6) for notation), and for $r > 0$, put

$$V(r) := \inf_{x\in\partial D} \text{Vol}(\text{cone}(x) \cap H(x_0 - x, x + r(x_0 - x))), \tag{4.14}$$

$$V_m(r) := \inf_{x\in\partial D_m(0)} \text{Vol}(\text{cone}(x) \cap H(x_0 - x, x + r(x_0 - x))). \tag{4.15}$$

Then for $\gamma > 0$ and sufficiently large $m \ge 1$, by Theorem 2.1,

$$\begin{aligned} &P\left(\sup_{0\le t\le T} d_H(\partial D_m(t), \partial D(t)) \ge \gamma\right) \\ &\le P\Bigg(||I_{D_m(T)}(\cdot) - I_{D(T)}(\cdot)||^2_{L^2([-K,K])} \ge V_0) \\ &\qquad + P(U_0 \subset D_m(T), \sup_{0\le t\le T} d_H(\partial D_m(t), \partial D(t)) \ge \gamma\Bigg) \\ &\to 0 \quad (\text{as } m\to\infty) \end{aligned} \tag{4.16}$$

(see (3.6) for notation). Indeed,

$$P\left(U_0 \subset D_m(T),\ \sup_{0\le t\le T} d_H(\partial D_m(t), \partial D(t)) \ge \gamma\right)$$
$$\le P\left(\sup_{0\le t\le T} ||I_{D_m(t)}(\cdot) - I_{D(t)}(\cdot)||^2_{L^2([-K,K])} \ge \min(V(\gamma), V_m(\gamma))\right),$$

and $V_m(\gamma) \ge V(\gamma)$ for all $m \ge 1$. □

Proof of Corollary 2.3. For $r \in \mathbf{R}$, let $\{D_r(t)\}_{t\ge 0}$ denote the unique solution of (1.13)–(1.14) with $D_r(0) = h^{-1}((r,\infty))$ on $[0,\infty)$. Notice that

$$D_r(\cdot) = \begin{cases} \mathbf{R}^N & \text{if } r < \inf\{h(x)|x \in \mathbf{R}^N\}, \\ \emptyset & \text{if } r \ge \sup\{h(x)|x \in \mathbf{R}^N\}. \end{cases} \tag{4.17}$$

Put

$$u(t,x) := \sup\{r \in \mathbf{R}|x \in D_r(t)\}. \tag{4.18}$$

Then, for all $t \ge 0$ and $r \in \mathbf{R}$ for which $D_r(t) \ne \emptyset$, $\mathbf{R}^N$,

$$u(t,\cdot)^{-1}((r,\infty)) = D_r(t), \tag{4.19}$$

since $D_r(t) = D_{r+}(t) := \cup_{\tilde r > r} D_{\tilde r}(t)$ by (1.13).

Indeed, $D_r(0) = D_{r+}(0)$ and $D_r(t) \supset D_{\tilde r}(t)$ for $\tilde r > r$ by Lemma 3.6. If $\tilde r - r$ is positive and is sufficiently small, then $D_{\tilde r}(t) \ne \emptyset$ by (b) in Theorem 2.4, and

$$\int_{\mathbf{R}^N} (I_{D_{\tilde r}(t)}(x) - I_{D_r(t)}(x))dx = \int_{\mathbf{R}^N} (I_{D_{\tilde r}(0)}(x) - I_{D_r(0)}(x))dx \uparrow 0 \quad (\text{as } \tilde r \to r). \tag{4.20}$$

By Lemma 3.6 and (4.19), u is continuous.

For $m \ge 1$, put

$$k_{m,1} := [m \sup\{h(y)|y \in \mathbf{R}^N\}],$$
$$k_{m,0} := [m \inf\{h(y)|y \in \mathbf{R}^N\}] - 1.$$

Then

$$\sum_{k_{m,0}\le k\le k_{m,1}} \frac{k}{m}\left(I_{D_{\frac{k}{m}}(t)}(x) - I_{D_{\frac{k+1}{m}}(t)}(x)\right) \tag{4.21}$$
$$= \sum_{k_{m,0}< k\le k_{m,1}} \frac{1}{m} I_{D_{\frac{k}{m}}(t)}(x) - \frac{k_{m,1}+1}{m} I_{D_{\frac{k_{m,1}+1}{m}}(t)}(x) + \frac{k_{m,0}}{m} I_{D_{\frac{k_{m,0}}{m}}(t)}(x).$$

Since $I_{D_{\frac{k_{m,1}+1}{m}}(t)}(x) \equiv 0$ and since $I_{D_{\frac{k_{m,0}}{m}}(t)}(x) \equiv 1$, the following holds: for any $\varphi \in C_o(\mathbf{R}^N)$ and any $t \ge 0$,

$$\int_{\mathbf{R}^N} \varphi(x) \left[\sum_{k_{m,0}\le k\le k_{m,1}} \frac{k}{m}\left(I_{D_{\frac{k}{m}}(0)}(x) - I_{D_{\frac{k+1}{m}}(0)}(x)\right)\right. \tag{4.22}$$
$$\left. - \sum_{k_{m,0}\le k\le k_{m,1}} \frac{k}{m}\left(I_{D_{\frac{k}{m}}(t)}(x) - I_{D_{\frac{k+1}{m}}(t)}(x)\right)\right] dx$$
$$= \int_0^t ds \left[\sum_{k_{m,0}< k\le k_{m,1}} \frac{1}{m} \int_{\mathbf{R}^N} \varphi(x)\omega_1\left(I_{D_{\frac{k}{m}}(s)}(\cdot), dx\right)\right].$$

Letting $m \to \infty$ in (4.22), one can show that u is a solution to (1.15) by Lemma 3.4 since

$$\omega_1\left(I_{D_{\frac{[mr]+1}{m}}(s)}(\cdot), dx\right) \to \omega_1(I_{D_r(s)}(\cdot), dx) \quad \text{weakly as } m \to \infty$$

and since

$$\omega_1(I_{D_r(s)}(\cdot), dx) = \omega_r(u(s,\cdot), dx),$$

except for at most countably many $r \in (\inf\{u(s,y)|y \in \mathbf{R}^N\}, \sup\{u(s,y)|y \in \mathbf{R}^N\})$.

Indeed,

$$\cup_{\tilde{r}>r}\{\text{co } u(s,\cdot)^{-1}((\tilde{r},\infty))\}^- = \text{co } u(s,\cdot)^{-1}((r,\infty)) \tag{4.23}$$

since $D_r(s) = D_{r+}(s)$ (see (4.19)–(4.20)) and since for $\tilde{r} > r$,

$$\begin{aligned}&\text{co } u(s,\cdot)^{-1}((\tilde{r},\infty)) \subset \{\text{co } u(s,\cdot)^{-1}((\tilde{r},\infty))\}^- \\ &\subset \text{co } u(s,\cdot)^{-1}([\tilde{r},\infty)) \subset \text{co } u(s,\cdot)^{-1}((r,\infty)).\end{aligned}$$

Besides, except for at most countably many r,

$$\{\text{co } u(s,\cdot)^{-1}((r,\infty))\}^- = \text{co } u(s,\cdot)^{-1}([r,\infty)) \tag{4.24}$$

since the sets $(\text{co } u(s,\cdot)^{-1}([r,\infty))\backslash\{\text{co } u(s,\cdot)^{-1}((r,\infty))\}^-)$ are disjoint for different r and since $(\text{co } u(s,\cdot)^{-1}([r,\infty))\backslash\{\text{co } u(s,\cdot)^{-1}((r,\infty))\}^-)$ is not empty if and only if it has a positive Lebesgue measure.

Let $v \in C([0,\infty) \times \mathbf{R}^N)$ be a solution to (1.15) with $v(0,\cdot) = h(\cdot)$. Then for $m \geq 1$, $r \in [\inf\{h(y)|y \in \mathbf{R}^N\}, \sup\{h(y)|y \in \mathbf{R}^N\})$, and $\varphi \in C_o(\mathbf{R}^N)$ and $t \geq 0$,

$$\begin{aligned}&\int_{\mathbf{R}^N} \varphi(x)\{\eta_m(v(0,x)-r) - \eta_m(v(t,x)-r)\}dx \\ &= \int_0^t ds \int_{\mathbf{R}} \frac{d\eta_m(\tilde{r}-r)}{d\tilde{r}} d\tilde{r} \int_{\mathbf{R}^N} \varphi(x)\omega_{\tilde{r}}(v(s,\cdot), dx)\end{aligned} \tag{4.25}$$

(see (3.24) for notation).

Let $m \to \infty$ in (4.25). Then we see that $\tilde{D}_r(t) := v(t,\cdot)^{-1}((r,\infty))$ is a solution to (1.14) on $[0,\infty)$ by Lemma 3.4. Indeed, in the same way as in (4.23), one can show that the following holds:

$$\cup_{\tilde{r}>r}\text{co } v(s,\cdot)^{-1}([\tilde{r},\infty)) = \text{co } v(s,\cdot)^{-1}((r,\infty)). \tag{4.26}$$

We prove that $v(t,\cdot)^{-1}((r,\infty))$ satisfies (1.13). For $x \in (\text{co } \tilde{D}_r(t)) \cap \tilde{D}_r(0)$, take $\delta > 0$ so that $U_\delta(x) \subset (\text{co } \tilde{D}_r(t)) \cap \tilde{D}_r(0)$. Then $U_\delta(x) \subset \text{co } \tilde{D}_r(s)$ for all $s \leq t$. Hence, by (1.14), for any $\varphi \in C_o(\mathbf{R}^N)$ such that $\varphi \equiv 0$ in $U_\delta(x)^c$,

$$\int_{\mathbf{R}^N} \varphi(y)\{I_{\tilde{D}_r(0)}(y) - I_{\tilde{D}_r(t)}(y)\}dy = \int_0^t ds \int_{\mathbf{R}^N} \varphi(y)\omega_1(I_{\tilde{D}_r(s)}(\cdot), dy) = 0, \tag{4.27}$$

which implies that $x \in (U_\delta(x) \subset)\tilde{D}_r(t)$. Hence (1.13) holds.

The uniqueness of u follows from that of $D_r(\cdot)$ for all r. □

Theorem 2.4 is an easy consequence of Theorem 2.1 and Lemma 3.6, and we omit the proof.

Proof of Theorem 2.5.

Step I. We first show that $u(t,x) := I_{D(t)}(x)$ is a viscosity supersolution of (1.20) in $(0,\infty)\times\mathbf{R}^N$.

Let $\psi \in \mathcal{A}((0,\infty)\times\mathbf{R}^N)$ and assume that $u-\psi$ attains a local minimum at $(t_0,x_0)\in(0,\infty)\times\mathbf{R}^N$. Without loss of generality, we may assume that $u(t_0,x_0)=\psi(t_0,x_0)$ and that $u(t,x)>\psi(t,x)$ for all $(t,x)\in(0,\infty)\times\mathbf{R}^N\setminus\{(t_0,x_0)\}$ (see [8]).

If $x_0\notin\partial(\text{co } D(t_0))\cap\partial D(t_0)$, then $\partial_t\psi(t_0,x_0)\geq 0$.

Indeed, $t\mapsto u(t,x_0)$ is constant if t_0-t is a sufficiently small positive number, from which $\psi(t_0,x_0)>\psi(t,x_0)$ for such t.

Suppose that $x_0\in\partial(\text{co } D(t_0))\cap\partial D(t_0)$. Then $u(t_0,x_0)=0$, and $D\psi(t_0,x_0)=o$ or $\sigma^+(u,D\psi(t_0,x_0),t_0,x_0)=1$.

Indeed, if $D\psi(t_0,x_0)\neq o$, then for y for which $y+x_0\notin H(D\psi(t_0,x_0),x_0)$ and for $r>0$, by the mean value theorem, there exists $\theta\in(0,1)$ such that

$$u(t_0,x_0+ry)>\psi(t_0,x_0+ry)=\psi(t_0,x_0)+r\langle D\psi(t_0,x_0+\theta ry),y\rangle>0,$$

provided r is sufficiently small, by the continuity of $D\psi$.

Case 1. We first consider the case when $D\psi(t_0,x_0)=o$. We may assume that there exist $f\in\mathcal{F}$ and $\varphi_1\in C^2((0,\infty))$ such that

$$\psi(t,x)=-f(|x-x_0|)-\varphi_1(t) \tag{4.28}$$

(see [21]).

For $A>0$ and $i\geq 2$, put

$$\psi_{i,A}(t,x)=\psi(t,x)-A\{|t-t_0|^2+|x-x_0|^i\}. \tag{4.29}$$

Then

$$\partial_t\psi_{i,A}(t_0,x_0)=\partial_t\psi(t_0,x_0),\quad D\psi_{i,A}(t_0,x_0)=D\psi(t_0,x_0), \tag{4.30}$$

and

$$\begin{aligned} U^+_{i,A,\varepsilon} &:= \{(t,x)\in(0,\infty)\times\mathbf{R}^N \,|\, \psi_{i,A}(t,x)+\varepsilon>u(t,x)\} \\ &\subset U_{(2^{i/2}\varepsilon/A)^{1/i}}((t_0,x_0)) \end{aligned} \tag{4.31}$$

($\varepsilon\in(0,A)$), and the following holds: for $t\geq 0$,

$$\lim_{x\to x_0} G(D\psi_{N,A}(t,x),D^2\psi_{N,A}(t,x))=NA. \tag{4.32}$$

We argue by contradiction. We consider $\psi_{N,A}$ instead of ψ. When it is not confusing, we omit $_{N,A}$ for the sake of simplicity.

Assume that the following holds:

$$\partial_t\psi(t_0,x_0)<0. \tag{4.33}$$

By reselecting $A>0$ sufficiently small and $\varepsilon>0$ sufficiently small compared to A if necessary, we may assume that $U_{(2^{i/2}\varepsilon/A)^{1/i}}((t_0,x_0))\subset(0,\infty)\times\mathbf{R}^N$, that

$$\partial_t\psi(t,x)+R\left(\frac{D\psi(t,x)}{|D\psi(t,x)|}\right)G(D\psi(t,x),D^2\psi(t,x))+\varepsilon<0 \quad \text{on } U^+_\varepsilon, \tag{4.34}$$

and that

$$U_\varepsilon^+ = \cup_{t>0}\{t\} \times (\psi(t,\cdot)^{-1}((-\varepsilon,\infty)) \cap D(t)^c). \tag{4.35}$$

We may also assume that $x \mapsto \psi(t,x)$ is strictly concave on U_ε^+; hence $x \mapsto (\psi(s,x), D\psi(s,x)/|D\psi(s,x)|)$ is one-to-one on some neighborhood of $\partial\psi(s,\cdot)^{-1}((-\varepsilon,\infty)) \cap D(s)^c$, provided $\psi(s,\cdot)^{-1}((-\varepsilon,\infty)) \cap D(s)^c \neq \emptyset$.

Indeed, if $\psi(s,\cdot)^{-1}((-\varepsilon,\infty)) \cap D(s)^c \neq \emptyset$, then $-\varepsilon$ is not the maximum of $\psi(s,\cdot)$ on $\psi(s,\cdot)^{-1}((-\varepsilon,\infty)) \cap D(s)^c$ and hence $D\psi(s,\cdot) \neq o$ on some neighborhood of $\partial\psi(s,\cdot)^{-1}((-\varepsilon,\infty)) \cap D(s)^c$.

For $t \geq 0$ and m and $k \geq 1$,

$$\begin{aligned}\int_{\mathbf{R}^N} &(\zeta_k(\eta_m(\psi(t,x)+\varepsilon) - u(t,x)) \\ &\quad - \zeta_k(\eta_m(\psi(0,x)+\varepsilon) - u(0,x)))dx \\ = \int_{\mathbf{R}^N} dx \int_0^t \Big(&- \zeta_k(\eta_m(\psi(s,x)+\varepsilon) - u(s,x))u(ds,x) \\ &+ \eta_k(\eta_m(\psi(s,x)+\varepsilon) - u(s,x))\frac{d\eta_m(\psi(s,x)+\varepsilon)}{dr}\partial_s\psi(s,x)ds\Big)\end{aligned} \tag{4.36}$$

(see (3.24)–(3.25) for notation).

Letting $k \to \infty$ in (4.36), by (4.31) and (4.35),

$$\begin{aligned}0 \leq \int_0^t ds \Bigg\{ &\int_{\psi(s,\cdot)^{-1}((-\varepsilon,\infty))\cap D(s)^c} \eta_m(\psi(s,x)+\varepsilon)\omega_1(u(s,\cdot),dx) \\ &+ \int_{\psi(s,\cdot)^{-1}((-\varepsilon,-\varepsilon+1/m))\cap D(s)^c} \frac{d\eta_m(\psi(s,x)+\varepsilon)}{dr}\partial_s\psi(s,x)dx\Bigg\}.\end{aligned} \tag{4.37}$$

For s for which $\psi(s,\cdot)^{-1}((-\varepsilon,-\varepsilon+1/m))\cap D(s)^c \neq \emptyset$ and sufficiently large $m \geq 1$, by Lemma 3.8 and (4.34),

$$\begin{aligned}&\int_{\psi(s,\cdot)^{-1}((-\varepsilon,-\varepsilon+1/m))\cap D(s)^c} \frac{d\eta_m(\psi(s,x)+\varepsilon)}{dr}\partial_s\psi(s,x)dx \\ &< -\int_{-\varepsilon}^{-\varepsilon+1/m} \frac{d\eta_m(r+\varepsilon)}{dr}dr \int_{\{\frac{-D\psi(s,x)}{|D\psi(s,x)|}:x\in\partial\psi(s,\cdot)^{-1}((r,\infty))\cap D(s)^c\}} (R(p) \\ &\qquad + \varepsilon\sup\{G(D\psi(s,x), D^2\psi(s,x)) : (s,x) \in U_\varepsilon^+\}^{-1})d\mathcal{H}^{N-1}(p) \\ &\to -\int_{\cup_{r>-\varepsilon}\{\frac{-D\psi(s,x)}{|D\psi(s,x)|}:x\in\partial\psi(s,\cdot)^{-1}((r,\infty))\cap D(s)^c\}} (R(p) + \varepsilon\sup\{G(D\psi(s,x), \\ &\qquad D^2\psi(s,x)) : (s,x) \in U_\varepsilon^+\}^{-1})d\mathcal{H}^{N-1}(p) \quad (\text{as } m \to \infty).\end{aligned} \tag{4.38}$$

Equations (4.37)–(4.38) contradict

$$\begin{aligned}&\{p \in \mathbf{S}^{N-1} : \sigma^+(u,-p,s,x) = 1 \text{ for some } x \in \psi(s,\cdot)^{-1}((-\varepsilon,\infty)) \cap D(s)^c\} \\ &\subset \cup_{r>-\varepsilon}\left\{-\frac{D\psi(s,x)}{|D\psi(s,x)|} : x \in \partial\psi(s,\cdot)^{-1}((r,\infty)) \cap D(s)^c\right\}\end{aligned}$$

since

$$\eta_m(\psi(s,x)+\varepsilon) \to 1 \quad \text{if } x \in \psi(s,\cdot)^{-1}((-\varepsilon,\infty)), \text{ as } m \to \infty.$$

Case 2. Next we consider the case when $\sigma^+(u, D\psi(t_0,x_0), t_0, x_0) = 1$. By (ii)–(iv) in Lemma 3.7, all eigenvalues of $-D(D\psi(t_0,x_0)/|D\psi(t_0,x_0)|)$ are nonnegative since the function $x \mapsto \psi(t_0,x)$ takes a maximum $\psi(t_0,x_0)$ on the set $\{x_0 + y \in \mathbf{R}^N | \langle y, D\psi(t_0,x_0)\rangle = 0\}$.

For $A > 0$, all eigenvalues of $-D(D\psi_{2,A}(t_0,x_0)/|D\psi_{2,A}(t_0,x_0)|)$ as a mapping on the set $\{y \in \mathbf{R}^N | \langle y, D\psi_{2,A}(t_0,x_0)\rangle = 0\}$ are greater than or equal to $2A/|D\psi(t_0,x_0)|$ (see (3.32)–(3.33)) since, in Lemma 3.7, 1 and $f_1, \ldots, f_{N-1}$ are an eigenvalue and eigenvectors of $(g_1 \cdots g_N)$, respectively.

We argue by contradiction. Assume that the following holds:

$$\partial_t\psi(t_0,x_0) + R\left(\frac{D\psi(t_0,x_0)}{|D\psi(t_0,x_0)|}\right)G(D\psi(t_0,x_0), D^2\psi(t_0,x_0)) < 0. \tag{4.39}$$

We consider $\psi_{2,A}$ instead of ψ. When it is not confusing, we omit $_{2,A}$ for the sake of simplicity. By reselecting A, $\varepsilon > 0$ if necessary, we may assume that (4.34)–(4.35) hold.

One can also assume, in $U_{(2\varepsilon/A)^{1/2}}((t_0,x_0))$, that $\partial_i\psi(s,x) \neq 0$ and all eigenvalues of $-D(D\psi(s,x)/|D\psi(s,x)|)$ as a mapping on the set $\{y \in \mathbf{R}^N | \langle y, D\psi(s,x)\rangle = 0\}$ are greater than or equal to $A/|D\psi(t_0,x_0)|$, and $x \mapsto y_i(s,x)$ is one-to-one for some $i \in \{1, \ldots, N\}$ by the inverse function theorem, (v) in Lemma 3.7, and Lemma 3.8.

In the same way as in (4.36)–(4.38), we obtain a contradiction.

Step II. We show that $u^-(t,x) = I_{D(t)^-}(x)$ is a viscosity subsolution of (1.20).

Let $\psi \in \mathcal{A}((0,\infty)\times\mathbf{R}^d)$ and assume that $u^- - \psi$ attains a maximum at $(t_0,x_0) \in (0,\infty)\times\mathbf{R}^d$. We may assume as well that $u^-(t_0,x_0) = \psi(t_0,x_0)$, so that $u^-(t,x) < \psi(t,x)$ for all $(t,x) \in (0,\infty)\times\mathbf{R}^d \setminus \{(t_0,x_0)\}$ (see [8]).

Since $t \mapsto u^-(t,x)$ is nonincreasing, $\partial_t\psi(t_0,x_0) \leq 0$.

Hence we have only to consider the case when the following holds: $D\psi(t_0,x_0) \neq o$, and

$$\sigma^-(u^-, D\psi(t_0,x_0), t_0, x_0) = 1, \quad R\left(\frac{D\psi(t_0,x_0)}{|D\psi(t_0,x_0)|}\right)G(D\psi(t_0,x_0), D^2\psi(t_0,x_0)) > 0.$$

In particular, $u^-(t_0,x_0) = 1$. By adding to ψ the function $(t,x) \mapsto A\{|t-s|^2+|x-y|^2\}$, with a sufficiently small $A > 0$, if necessary, we may assume that

$$U_\varepsilon^- := \{(t,x) \in (0,\infty)\times\mathbf{R}^d | \psi(t,x) - \varepsilon < u^-(t,x)\} \quad (\varepsilon > 0) \tag{4.40}$$

is contained in the set $U_{(\varepsilon/A)^{1/2}}((t_0,x_0))$.

We argue by contradiction. Assume that the following holds:

$$\partial_t\psi(t_0,x_0) + R\left(\frac{D\psi(t_0,x_0)}{|D\psi(t_0,x_0)|}\right)G(D\psi(t_0,x_0), D^2\psi(t_0,x_0)) > 0. \tag{4.41}$$

By reselecting $\varepsilon > 0$ if necessary, we may assume that

$$\partial_t\psi(t,x) + R\left(\frac{D\psi(t,x)}{|D\psi(t,x)|}\right)G(D\psi(t,x), D^2\psi(t,x)) - \varepsilon > 0, \tag{4.42}$$

and $u^-(t,x) = 1$ on U_ε^- by the continuity of ψ.

Put $\tilde{\eta}_m(r) = \eta_m(r + 1/m)$ for $r \in \mathbf{R}$ and $m \geq 1$. In the same way as in Step I, considering $u^-(t,x) - \tilde{\eta}_m(\psi(t,x) - 1 - \varepsilon)$ instead of $\eta_m(\psi(t,x) + \varepsilon) - u(t,x)$, we obtain a contradiction. □

Proof of Corollary 2.6. We first show that u is a viscosity supersolution of (1.20) in $(0,\infty)\times\mathbf{R}^N$. Let $\varphi\in\mathcal{A}((0,\infty)\times\mathbf{R}^N)$ and assume that $u-\varphi$ attains a minimum at $(t_0,x_0)\in(0,\infty)\times\mathbf{R}^N$. We may assume that $u(t_0,x_0)=\varphi(t_0,x_0)$, so that $u(t,x)>\varphi(t,x)$ for all $(t,x)\in(0,\infty)\times\mathbf{R}^N\setminus\{(t_0,x_0)\}$ (see [8]). By subtracting a constant, we may assume that $\varphi\le u<0$.

Put $r_0:=\varphi(t_0,x_0)$ and

$$u_r(t,x):=I_{u^{-1}(t,\cdot)((r,0))}(x)\quad(r<0). \tag{4.43}$$

Then

$$u_{r_0}(t,x)\ge\frac{\varphi(t,x)}{|r_0|}+1\quad\text{for all }(t,x)\in(0,\infty)\times\mathbf{R}^N, \tag{4.44}$$

where the equality holds if and only if $(t,x)=(t_0,x_0)$.

Since u_r is a viscosity supersolution of (1.20) in $(0,\infty)\times\mathbf{R}^N$ by Corollary 2.3 and Theorem 2.5, and since

$$\sigma^+(u_{r_0},D(\varphi(t_0,x_0)/|r_0|+1),t_0,x_0)=\sigma^+(u,D\varphi(t_0,x_0),t_0,x_0),$$

(1.25) holds.

Next we show that u is a viscosity subsolution of (1.20) in $(0,\infty)\times\mathbf{R}^N$. Let $\varphi\in\mathcal{A}((0,\infty)\times\mathbf{R}^d)$ and assume that $u-\varphi$ attains a maximum at $(t_1,x_1)\in(0,\infty)\times\mathbf{R}^d$. We may assume as well that $u(t_1,x_1)=\varphi(t_1,x_1)$, so that $u(t,x)<\varphi(t,x)$ for all $(t,x)\in(0,\infty)\times\mathbf{R}^d\setminus\{(t_1,x_1)\}$ (see [8]).

By adding a constant, we may assume that $\varphi\ge u>0$.

Put $r_1:=\varphi(t_1,x_1)$ and

$$u_r^-(t,x):=I_{u^{-1}(t,\cdot)([r,\infty))}(x)\quad(r<0). \tag{4.45}$$

Then

$$u_{r_1}^-(t,x)\le\frac{\varphi(t,x)}{r_1}\quad\text{for all }(t,x)\in(0,\infty)\times\mathbf{R}^N, \tag{4.46}$$

where the equality holds if and only if $(t,x)=(t_1,x_1)$.

Since u_r^- is a viscosity subsolution of (1.20) in $(0,\infty)\times\mathbf{R}^N$ by Corollary 2.3 and Theorem 2.5, and since

$$\sigma^-(u_{r_1}^-,D(\varphi(t_1,x_1)/r_1),t_1,x_1)=\sigma^-(u,D\varphi(t_1,x_1),t_1,x_1),$$

(1.27) holds. □

REFERENCES

[1] D. Adalsteinsson, L. C. Evans, and H. Ishii, *The level set method for etching and deposition*, Math. Models Methods Appl. Sci., 7 (1997), pp. 1153–1186.

[2] B. Andrews, *Gauss curvature flow: The fate of the rolling stones*, Invent. Math., 138 (1999), pp. 151–161.

[3] B. Andrews, *Motion of hypersurfaces by Gauss curvature*, Pacific J. Math., 195 (2000), pp. 1–34.

[4] I. J. Bakelman, *Convex Analysis and Nonlinear Geometric Elliptic Equations*, Springer-Verlag, Berlin, Heidelberg, New York, 1994.

[5] Y.-G. Chen, Y. Giga, and S. Goto, *Uniqueness and existence of viscosity solutions of generalized mean curvature flow equations*, J. Differential Geom., 33 (1991), pp. 749–786.

[6] D. Chopp, L. C. Evans, and H. Ishii, *Waiting time effects for Gauss curvature flows*, Indiana Univ. Math. J., 48 (1999), pp. 311–334.
[7] B. Chow, *Deforming convex hypersurfaces by the nth root of the Gaussian curvature*, J. Differential Geom., 22 (1985), pp. 117–138.
[8] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.
[9] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*, John Wiley and Sons, New York, 1986.
[10] L. C. Evans and J. Spruck, *Motion of level sets by mean curvature* I, J. Differential Geom., 33 (1991), pp. 635–681.
[11] W. J. Firey, *Shapes of worn stones*, Mathematika, 21 (1974), pp. 1–11.
[12] M.-H. Giga and Y. Giga, *Crystalline and level set flow-convergence of a crystalline algorithm for a general anisotropic curvature flow in the plain*, in Free Boundary Problems: Theory and Applications, I (Chiba, 1999), GAKUTO Internat. Ser. Math. Sci. Appl. 13, Gakkōtosho, Tokyo, 2000, pp. 64–79.
[13] P. M. Girão, *Convergence of a crystalline algorithm for the motion of a simple closed convex curve by weighted curvature*, SIAM J. Numer. Anal., 32 (1995), pp. 886–899.
[14] R. Hamilton, *Worn stones with flat sides*, in A Tribute to Ilya Bakelman (College Station, TX, 1993), Discourses Math. Appl. 3, Texas A & M University, College Station, TX, 1994, pp. 69–78.
[15] N. Ikeda and S. Watanabe, *Stochastic Differential Equations and Diffusion Processes*, North-Holland/Kodansha, Tokyo, 1981.
[16] H. Ishii, *Gauss curvature flow and its approximation*, in Free Boundary Problems: Theory and Applications, II (Chiba, 1999), GAKUTO Internat. Ser. Math. Sci. Appl. 14, Gakkōtosho, Tokyo, 2000, pp. 198–206.
[17] H. Ishii and T. Mikami, *A mathematical model of the wearing process of a nonconvex stone*, SIAM J. Math. Anal., 33 (2001), pp. 860–876.
[18] H. Ishii and T. Mikami, *A two dimensional random crystalline algorithm for Gauss curvature flow*, Adv. in Appl. Probab., 34 (2002), pp. 491–504.
[19] H. Ishii and T. Mikami, *A level set approach to the wearing process of a nonconvex stone*, Calc. Var. Partial Differential Equations, 19 (2003), pp. 53–93.
[20] H. Ishii and T. Mikami, *Motion of a graph by R-curvature*, Arch. Ration. Mech. Anal., 171 (2004), pp. 1–23.
[21] H. Ishii and P. E. Souganidis, *Generalized motion of noncompact hypersurfaces with velocity having arbitrary growth on the curvature tensor*, Tohoku Math. J. (2), 47 (1995), pp. 227–250.
[22] S. Osher and J. Sethian, *Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations*, J. Comput. Phys., 79 (1988), pp. 12–49.
[23] K. Tso, *Deforming a hypersurface by its Gauss-Kronecker curvature*, Comm. Pure Appl. Math., 38 (1985), pp. 867–882.

SIAM J. MATH. ANAL.
Vol. 36, No. 2, pp. 580–617

PROPAGATION OF VISCOUS SHOCK WAVES AWAY FROM THE BOUNDARY*

CHIU-YA LAN†, HUEY-ER LIN†, TAI-PING LIU‡, AND SHIH-HSIEN YU§

Abstract. We study the propagation of shock waves away from the boundary for viscous conservation law. Our main purpose is to obtain pointwise description of the perturbation of the shock profile. We show that there are different convergence rates for the region between the boundary and the shock and the region ahead of the shock. The dependence of these rates on the shock strength, viscosity, and initial perturbation is studied. There are two mechanisms which govern the solution behavior: the compressibility of the shock and the presence of the boundary. We introduce an iteration scheme to decouple these two effects. Thus near the boundary we use the Green's function for the initial-boundary value problem of the equation linearized around the boundary value; away from the boundary we use the Green's function for the initial-value problem of the equation linearized around the shock profile. To focus on our main ideas, we study the Burgers equation, for which the Green's functions have explicit forms. Our new approach should be applicable to more general situations such as the system of viscous conservation laws.

Key words. time asymptotic, pointwise approach, shock location

AMS subject classifications. 35L65, 76L05, 76N10

DOI. 10.1137/S0036141003428159

1. Introduction.

The purpose of this paper is to study the effect of the boundary on the time asymptotic behavior of the propagation of shock waves for the viscous conservation law

$$u_t + f(u)_x = u_{xx}.$$

We are interested in the interplay of the effects of boundary, nonlinearity through the strength of the shock, and the initial data. We will consider the case when the flux is strongly nonlinear: $f''(u) \neq 0$. For simplicity we consider the Burgers equation. The situation is simpler when the shock is propagating toward the boundary and becomes the boundary layer. We consider the case when the shock is propagating away from the boundary. Thus we will study the stability of a viscous shock profile for the initial-boundary value problem

$$u_t + uu_x = u_{xx}, \tag{1.1}$$

$$u(-L-t,t) = u_-, \qquad u(\infty,t) = -u_- \equiv u_+, \tag{1.2}$$

$$u(x,0) = \phi(x) + \bar{u}(x), \qquad \bar{u}(-L) = u_- - \phi(-L), \tag{1.3}$$

$$\phi(x) \equiv -u_- \tanh\frac{(u_- x)}{2},$$

*Received by the editors May 19, 2003; accepted for publication (in revised form) December 5, 2003; published electronically August 6, 2004.
http://www.siam.org/journals/sima/36-2/42815.html

†Institute of Mathematics, Academia Sinica, Taipei 11529, Taiwan (cylan@math.sinica.edu.tw, helin@math.sinica.edu.tw). The research of the first two authors was supported by the Institute of Mathematics, Academia Sinica, and NSC 91-2811-M-001-013, NSC 91-2811-M-001-015, NSC 91007p, and NSC 91008p.

‡Institute of Mathematics, Academia Sinica, Taipei 11529, Taiwan and Department of Mathematics, Stanford University, Stanford, CA 94305-2125 (tpliu@math.sinica.edu.tw and liu@math.stanford.edu). This author's research was supported in part by NSC 91-2115-M-001-004.

§Department of Mathematics, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong (mashyu@cityu.edu.hk). This author's research was supported in part by CityU Strategic Research Grant 7001426.

where $\phi(x)$ is a stationary shock wave solution of the Burgers equation and $L > 0$ is a given constant.

To highlight the effect of the boundary, we will carry out the detailed analysis for the case when the initial perturbation is small and has algebraic decay. Since we are interested in the time asymptotic behavior, we assume that the shock is initially not located near the boundary; i.e. L is large:

$$\bar{u}(x) = O(1)e^{-\frac{L}{d}}(1+x+L)^{-\alpha}, \tag{1.4}$$

$$\bar{u}_x(x) = O(1)e^{-\frac{L}{d}}(1+x+L)^{-\alpha-1}, \quad \alpha > 1,\ d \geq 1. \tag{1.5}$$

In this case, the boundary effect and the presence of the shock give rise to convergence at an exponential rate for the region between the shock and the boundary, and at an algebraic rate for the region around and ahead of the shock. The choice of algebraic decaying initial data is also motivated by the study of the system of conservation laws, for which the wave interactions of distinct characteristic families give rise to waves of algebraic rates [1].

There have been works on the boundary effects using the energy method [3], [2], [5]. Our pointwise approach extends that of [4], which studies the propagation of stationary shocks. For the pointwise approach, one needs to use the exact solution representation, that is, make use of the Green's functions. In this regard, the present study introduces an important new analytical device for separating the nonlinearity effect of the shock from that of the boundary. Specifically, for the region around the boundary we use the Green's function for the initial-boundary value problem but ignore the presence of the shock. For the region away from the boundary we use the Green's function for the equation linearized around the shock, ignoring the presence of the boundary. The advantage is the simplicity of the forms of these Green's functions and the clarity of the different roles played by the boundary and the shock on the asymptotic behavior of the solutions. This avoids the potentially complicated construction of the Green's function with the presence of both the boundary and the shock. This new methodology was raised first in the unpublished work [6], and is applicable in principle to systems, a task we will pursue in the future.

There are three parts in this paper: first, we obtain preliminary pointwise estimates, particularly the boundary estimate of the solution of (1.1), (1.2), and (1.3) around a viscous shock wave. The boundary estimate thus obtained allows us to locate, time-asymptotically, the shock profile. With the shock location determined, we then study the time asymptotic behavior of the perturbation. We show that the perturbation decays exponentially for the region between the boundary and the shock, and algebraically for the region ahead of the shock. Finally, in the last section we remark on the different roles played by the nonlinearity, viscosity, boundary, and initial data in the time asymptotic behavior of the solutions.

2. Boundary estimate. In this section we will obtain the boundary estimate. Since the shift of the shock location due to the perturbation and the boundary is not known a priori, we therefore will not try to properly shift the shock. Thus the interior estimate given in this section is rough and there is no time decay. However, the boundary estimate given here is optimal and will be crucial in the next section in determining the shock shift. For simplicity in the analysis, we will take the shock to be of finite strength; in fact, we assume that $u_- = 1$ and the initial data to be

$$\bar{u}(x) = O(1)e^{-L}(1+x+L)^{-\alpha}, \qquad \bar{u}_x(x) = O(1)e^{-L}(1+x+L)^{-\alpha-1}.$$

Set the perturbation with zero boundary value to be $v(x,t)$:

$$v(x,t) \equiv u(x,t) - \phi(x) - \Psi(x,t),$$
$$\Psi(x,t) = (1-\phi(-L-t))(1+x+L+t)^{-\alpha}.$$

Then $v(x,t)$ satisfies

$$v_t + (\phi v)_x - v_{xx} = -\left(\frac{v^2}{2} + \Psi v\right)_x - \Psi_t - \left(\phi\Psi + \frac{\Psi^2}{2}\right)_x + \Psi_{xx}, \tag{2.1}$$

$$v(-L-t,t) = 0, \qquad v(\infty,t) = 0, \tag{2.2}$$

$$\begin{cases} v(x,0) = \bar{u}(x) - \Psi(x,0) = O(1)e^{-L}(1+x+L)^{-\alpha}, & v(-L,0) = 0, \\ v_x(x,0) = O(1)e^{-L}(1+x+L)^{-\alpha-1}. \end{cases} \tag{2.3}$$

As mentioned before, we separate the space-time domain into two parts: the region near the boundary, $x \in [-L-t, -(L+t)/2)$, and that away from the boundary, $x \geq -(L+t)/2$.

Region I. For $x \in [-L-t, -(L+t)/2)$, we use the Green's function $K^B(x,t;y,\sigma)$ for

$$w_t + w_x = w_{xx}, \tag{2.4}$$

$$w(-L-t,t) = 0, \qquad w(\infty,t) = 0. \tag{2.5}$$

The following expression of the Green's function is easily obtained by converting, through reflection, the initial-boundary problem into the initial value problem for which the Green's function is the translated heat kernel:

$$k(x,t) \equiv e^{-\frac{x^2}{4t}}/\sqrt{4\pi t}.$$

LEMMA 2.1. *The Green's function $K^B(x,t;y,\sigma)$ for* (2.4) *and* (2.5) *is*

$$K^B(x,t;y,\sigma) = k(x-y-(t-\sigma),t-\sigma) - k(x+y-t+3\sigma+2L,t-\sigma)e^{-2(y+L+\sigma)}.$$

We represent the solution $v(x,t)$ of (2.1), (2.2), and (2.3) using this Green's function. Multiply K^B with (2.1) and integrate the equation to yield

$$\begin{aligned} v(x,t) = &\int_{-L}^{\infty} K^B(x,t;y,0)v(y,0)dy \\ &+ \int_0^t \int_{-L-\sigma}^{\infty} K_y^B(x,t;y,\sigma)\left[(\phi(y)-1+\Psi(y,\sigma))v(y,\sigma) + \frac{v(y,\sigma)^2}{2}\right]dyd\sigma \\ &+ \int_0^t \int_{-L-\sigma}^{\infty} K^B(x,t;y,\sigma)\left\{-\Psi_\sigma - \left(\phi\Psi + \frac{\Psi^2}{2}\right)_y + \Psi_{yy}\right\}(y,\sigma)dyd\sigma. \end{aligned} \tag{2.6}$$

Region II. For $x \geq -(L+t)/2$, we focus on the shock profile and consider the Green's function $G(x,t;y,\sigma)$ for the initial value problem

$$w_t + (\phi(x)w)_x - w_{xx} = 0, \quad -\infty < x < \infty. \tag{2.7}$$

The Green's function for this is easily obtained from Hopf–Cole transformation. It is also easy to see that the Green's function is the weighted heat kernels, as stated in the corollary below; cf. [1].

LEMMA 2.2. *The Green's function $G(x,t;y,\sigma)$ for (2.7) is*

$$
\begin{aligned}
G(x,t;y,\sigma) &= -\int_{-\infty}^{y} g_x(x,t;\xi,\sigma)d\xi \\
&= g(x,t;y,\sigma) + \frac{\int_{-\infty}^{y} \sinh(\frac{x-\xi}{2})k(x-\xi,t-\sigma)e^{-\frac{t-\sigma}{4}}d\xi}{2\cosh^2\frac{x}{2}},
\end{aligned}
\tag{2.8}
$$

where

$$
g(x,t;y,\sigma) \equiv \frac{\cosh(\frac{y}{2})}{\cosh(\frac{x}{2})}k(x-y,t-\sigma)e^{-\frac{t-\sigma}{4}} \tag{2.9}
$$

is the Green's function for

$$
w_t + \phi(x)w_x - w_{xx} = 0.
$$

COROLLARY 2.3. *The Green's function $g(x,t;y,\sigma)$ of $w_t+\phi(x)w_x-w_{xx}=0$ can be viewed as a weighted heat kernel:*

$$
\begin{aligned}
&\text{for } x>0,\ y>0, \quad g(x,t;y,\sigma) = O(1)k(x-y+(t-\sigma),t-\sigma),\\
&\text{for } x>0,\ y<0, \quad g(x,t;y,\sigma) = O(1)e^{-|x|}k(x-y-(t-\sigma),t-\sigma),\\
&\text{for } x<0,\ y<0, \quad g(x,t;y,\sigma) = O(1)k(x-y-(t-\sigma),t-\sigma),\\
&\text{for } x<0,\ y>0, \quad g(x,t;y,\sigma) = O(1)e^{-|x|}k(x-y+(t-\sigma),t-\sigma).
\end{aligned}
$$

More precisely, $g(x,t;y,\sigma)$ can be written as the following two equivalent expressions:

$$
g(x,t;y,\sigma) = \begin{cases} \dfrac{1+e^{y}}{1+e^{x}}k(x-y-(t-\sigma),t-\sigma), \\ \dfrac{1+e^{-y}}{1+e^{-x}}k(x-y+(t-\sigma),t-\sigma). \end{cases}
$$

With the second Green's function, we obtain a representation for the solution that is different from (2.6):

$$
\begin{aligned}
v(x,t) = &\int_{-L}^{\infty} G(x,t;y,0)v(y,0)dy \\
&-\int_0^t G(x,t;-L-\sigma,\sigma)v_y(-L-\sigma,\sigma)d\sigma \\
&+\int_0^t\int_{-L-\sigma}^{\infty} G_y(x,t;y,\sigma)\left(\Psi(y,\sigma)v(y,\sigma)+\frac{v(y,\sigma)^2}{2}\right)dyd\sigma \\
&+\int_0^t\int_{-L-\sigma}^{\infty} G(x,t;y,\sigma)\left\{-\Psi_\sigma-\left(\phi\Psi+\frac{\Psi^2}{2}\right)_y+\Psi_{yy}\right\}(y,\sigma)dyd\sigma.
\end{aligned}
\tag{2.10}
$$

We study the solution $v(x,t)$ for (2.1), (2.2), and (2.3) by the following iterations: for $x\in[-L-t,-(L+t)/2)$,

$$
\begin{aligned}
v^0(x,t) = &\int_{-L}^{\infty} K^B(x,t;y,0)v(y,0)dy \\
&+\int_0^t\int_{-L-\sigma}^{\infty} K^B(x,t;y,\sigma)\left\{-\Psi_\sigma-\left(\phi\Psi+\frac{\Psi^2}{2}\right)_y+\Psi_{yy}\right\}(y,\sigma)dyd\sigma;
\end{aligned}
\tag{2.11}
$$

for $x \geq -(L+t)/2$,

$$
\begin{aligned}
v^0(x,t) = &\int_{-L}^{\infty} G(x,t;y,0)v(y,0)dy \\
&- \int_0^t G(x,t;-L-\sigma,\sigma)v_y^0(-L-\sigma,\sigma)d\sigma \\
&+ \int_0^t \int_{-L-\sigma}^{\infty} G(x,t;y,\sigma)\left\{-\Psi_\sigma - \left(\phi\Psi + \frac{\Psi^2}{2}\right)_y + \Psi_{yy}\right\}(y,\sigma)dyd\sigma;
\end{aligned} \tag{2.12}
$$

for $x \in [-L-t, -(L+t)/2)$, $n \geq 1$,

$$
\begin{aligned}
v^n(x,t) = &\int_{-L}^{\infty} K^B(x,t;y,0)v(y,0)dy \\
&+ \int_0^t \int_{-L-\sigma}^{\infty} K_y^B(x,t;y,\sigma) \\
&\qquad \cdot \left[(\phi(y) - 1 + \Psi(y,\sigma))v^{n-1}(y,\sigma) + \frac{v^{n-1}(y,\sigma)^2}{2}\right] dyd\sigma \\
&+ \int_0^t \int_{-L-\sigma}^{\infty} K^B(x,t;y,\sigma)\left\{-\Psi_\sigma - \left(\phi\Psi + \frac{\Psi^2}{2}\right)_y + \Psi_{yy}\right\}(y,\sigma)dyd\sigma;
\end{aligned} \tag{2.13}
$$

for $x \geq -(L+t)/2$, $n \geq 1$,

$$
\begin{aligned}
v^n(x,t) = &\int_{-L}^{\infty} G(x,t;y,0)v(y,0)dy \\
&- \int_0^t G(x,t;-L-\sigma,\sigma)v_y^n(-L-\sigma,\sigma)d\sigma \\
&+ \int_0^t \int_{-L-\sigma}^{\infty} G_y(x,t;y,\sigma)\left(\Psi(y,\sigma)v^{n-1}(y,\sigma) + \frac{v^{n-1}(y,\sigma)^2}{2}\right) dyd\sigma \\
&+ \int_0^t \int_{-L-\sigma}^{\infty} G(x,t;y,\sigma)\left\{-\Psi_\sigma - \left(\phi\Psi + \frac{\Psi^2}{2}\right)_y + \Psi_{yy}\right\}(y,\sigma)dyd\sigma.
\end{aligned} \tag{2.14}
$$

From the expression of $\Psi(x,t)$ at the beginning of this section, we have

$$
\Psi_t + \left(\phi\Psi + \frac{\Psi^2}{2}\right)_x - \Psi_{xx} = O(1)e^{-L-t}(1+x+L+t)^{-\alpha}.
$$

The first step is to estimate v^0, for which we need the following lemmas. Lemma 2.4 follows from straightforward computations. In Lemma 2.5 both algebraic and exponential rates are described, and we need to consider various regions in the (x,t) space in its proof.

LEMMA 2.4. *For $x \in [-L-t, 0)$,*

$$
\int_{-L}^{\infty} k(x-y-t,t)e^{-L}(1+y+L)^{-\alpha}dy \leq Ce^{-\frac{L}{2}}e^{-\frac{|x|}{2}}e^{-\frac{t}{4}}, \tag{2.15}
$$

$$
\begin{aligned}
&\int_0^t \int_{-L-\sigma}^{\infty} k(x-y-(t-\sigma),t-\sigma)e^{-L-\sigma}(1+y+L+\sigma)^{-\alpha}dyd\sigma \\
&\qquad \leq e^{-\frac{L}{2}}e^{-\frac{|x|}{2}}e^{-\frac{t}{4}}.
\end{aligned} \tag{2.16}
$$

LEMMA 2.5. *For $x \geq -(L+t)/2$ and $r > 1$,*

$$\int_0^t G(x,t;-L-\sigma,\sigma)e^{-\frac{L}{2r}}e^{-\frac{L+\sigma}{2}}d\sigma \leq Ce^{-\frac{L}{2r}}e^{-\frac{|x|}{2}}. \tag{2.17}$$

$$\begin{aligned}&\int_0^t\int_{-L-\sigma}^{\infty} G(x,t;y,\sigma)e^{-L-\sigma}(1+y+L+\sigma)^{-\alpha}dyd\sigma\\ &\qquad\leq C\begin{cases} e^{-L}e^{-|x|}+e^{-\frac{L}{2}}e^{-\frac{|x|}{2}}(1+t)^{-\alpha} & \text{if } x\in[-(L+t)/2,0),\\ e^{-\frac{L}{2}}e^{-\frac{|x|}{2}}+e^{-L}(1+x+L+t)^{-\alpha} & \text{if } x\geq 0.\end{cases}\end{aligned} \tag{2.18}$$

$$\begin{aligned}&\int_{-L}^{\infty} G(x,t;y,0)e^{-L}(1+y+L)^{-\alpha}dy\\ &\qquad\leq C\begin{cases} e^{-\frac{L}{2}}e^{-\frac{|x|}{2}} & \text{if } x\in[-(L+t)/2,0),\\ e^{-L}e^{-|x|}+e^{-L}(1+x+L+t)^{-\alpha} & \text{if } x\geq 0.\end{cases}\end{aligned} \tag{2.19}$$

Proof. From (2.8)

$$G(x,t;y,\sigma) = O(1)\left\{e^{-\frac{|x|}{2}}e^{\frac{|y|}{2}}k(x-y,t-\sigma)e^{-\frac{t-\sigma}{4}}+e^{-|x|}\right\}$$

and so we have (2.17)

$$\begin{aligned}&\int_0^t G(x,t;-L-\sigma,\sigma)e^{-\frac{L}{2r}}e^{-\frac{L+\sigma}{2}}d\sigma\\ &\quad\leq C\left\{\int_0^t e^{-\frac{|x|}{2}}k(x+L+\sigma,t-\sigma)e^{-\frac{t-\sigma}{4}}e^{-\frac{L}{2r}}d\sigma+\int_0^t e^{-|x|}e^{-\frac{L}{2r}}e^{-\frac{L+\sigma}{2}}d\sigma\right\}\\ &\quad\leq Ce^{-\frac{L}{2r}}e^{-\frac{|x|}{2}}.\end{aligned}$$

Similarly,

$$\begin{aligned}&\int_0^t\int_{-L-\sigma}^{\infty} G(x,t;y,\sigma)e^{-L-\sigma}(1+y+L+\sigma)^{-\alpha}dyd\sigma\\ &\qquad=\int_0^t\int_{-L-\sigma}^{\infty} g(x,t;y,\sigma)e^{-L-\sigma}(1+y+L+\sigma)^{-\alpha}dyd\sigma\\ &\qquad\quad+\int_0^t\int_{-L-\sigma}^{\infty} O(1)e^{-|x|}e^{-L-\sigma}(1+y+L+\sigma)^{-\alpha}dyd\sigma\equiv g_1+g_2,\end{aligned}$$

where g_2 is obtained by straightforward computation,

$$g_2=\int_0^t\int_{-L-\sigma}^{\infty} O(1)e^{-|x|}e^{-L-\sigma}(1+y+L+\sigma)^{-\alpha}dyd\sigma\leq Ce^{-|x|}e^{-L},$$

and from Corollary 2.3, when $x\in[-(L+t)/2,-t/2)$,

$$\begin{aligned}g_1&=\int_0^t\int_{-L-\sigma}^{\infty} g(x,t;y,\sigma)e^{-L-\sigma}(1+y+L+\sigma)^{-\alpha}dyd\sigma\\ &\leq C\left\{\int_0^t\int_{-L-\sigma}^{0} k(x-y-(t-\sigma),t-\sigma)e^{-L-\sigma}(1+y+L+\sigma)^{-\alpha}dyd\sigma\right.\\ &\qquad\left.+\int_0^t\int_0^{\infty} e^{-|x|}k(x-y+(t-\sigma),t-\sigma)e^{-L-\sigma}(1+y+L+\sigma)^{-\alpha}dyd\sigma\right\}\end{aligned}$$

$$\le C\left\{\int_0^t\int_{-L-\sigma}^0 \frac{1}{\sqrt{4\pi(t-\sigma)}}e^{-\frac{(x-y)^2}{4(t-\sigma)}+\frac{(x-y)}{2}-\frac{t-\sigma}{4}}e^{-L-\sigma}(1+y+L+\sigma)^{-\alpha}dyd\sigma \right.$$
$$\left. +\int_0^t e^{-|x|}e^{-L-\sigma}d\sigma\right\}$$
$$\le C\left\{e^{-\frac{L}{2}}e^{-\frac{|x|}{2}}e^{-\frac{t}{4}}+e^{-L}e^{-|x|}\right\}$$
$$\le Ce^{-\frac{L}{2}}e^{-\frac{|x|}{2}}e^{-\frac{t}{4}};$$

when $x\in[-t/2,0)$,

$$g_1\le C\left\{e^{-\frac{L}{2}}e^{-\frac{|x|}{2}}e^{-\frac{t}{4}}\right.$$
$$+\int_0^{\frac{t}{4}}\left(\int_0^{\frac{t-\sigma}{8}}+\int_{\frac{t-\sigma}{8}}^{\infty}\right)$$
$$e^{-|x|}k(x-y+(t-\sigma),t-\sigma)e^{-L-\sigma}(1+y+L+\sigma)^{-\alpha}dyd\sigma$$
$$\left.+\int_{\frac{t}{4}}^t\int_0^{\infty}e^{-|x|}k(x-y+(t-\sigma),t-\sigma)e^{-L-\sigma}(1+y+L+\sigma)^{-\alpha}dyd\sigma\right\}$$
$$\le C\left\{e^{-\frac{L}{2}}e^{-\frac{|x|}{2}}e^{-\frac{t}{4}}\right.$$
$$+\int_0^{\frac{t}{4}}e^{-|x|}e^{-\frac{(x-\frac{(t-\sigma)}{8}+(t-\sigma))^2}{4(t-\sigma)}}e^{-L-\sigma}(1+L+\sigma)^{-\alpha}d\sigma$$
$$+\int_0^{\frac{t}{4}}e^{-|x|}e^{-L-\sigma}\left(1+\frac{t}{8}+L+\frac{7\sigma}{8}\right)^{-\alpha}d\sigma$$
$$\left.+\int_{\frac{t}{4}}^t e^{-|x|}e^{-L-\sigma}(1+L+\sigma)^{-\alpha}d\sigma\right\}$$
$$\le C\left\{e^{-\frac{L}{2}}e^{-\frac{|x|}{2}}e^{-\frac{t}{4}}+e^{-L}e^{-\frac{|x|}{2}}(1+t)^{-\alpha}\right\};$$

and when $x\ge 0$,

$$g_1\le C\left\{\int_0^t\int_{-L-\sigma}^0 e^{-|x|}k(x-y-(t-\sigma),t-\sigma)e^{-L-\sigma}(1+y+L+\sigma)^{-\alpha}dyd\sigma\right.$$
$$\left.+\int_0^t\int_0^{\infty}k(x-y+(t-\sigma),t-\sigma)e^{-L-\sigma}(1+y+L+\sigma)^{-\alpha}dyd\sigma\right\}$$
$$\le C\left\{e^{-\frac{L}{2}}e^{-\frac{|x|}{2}}e^{-\frac{t}{4}}\right.$$
$$+\int_0^t\left(\int_0^{\frac{x}{2}}+\int_{\frac{x}{2}}^{x+\frac{t-\sigma}{2}}+\int_{x+\frac{t-\sigma}{2}}^{\infty}\right)$$
$$\left.k(x-y+(t-\sigma),t-\sigma)e^{-L-\sigma}(1+y+L+\sigma)^{-\alpha}dyd\sigma\right\}$$

$$
\begin{aligned}
&\le C\left\{e^{-\frac{L}{2}}e^{-\frac{|x|}{2}}e^{-\frac{t}{4}}\right.\\
&\quad+\int_0^t e^{-\frac{(x-(x/2)+(t-\sigma))^2}{4(t-\sigma)}}e^{-L-\sigma}d\sigma+\int_0^t e^{-\frac{((t-\sigma)/2)^2}{4(t-\sigma)}}e^{-L-\sigma}\left(1+\frac{x}{2}+L+\sigma\right)^{-\alpha}d\sigma\\
&\quad\left.+\int_0^t e^{-L-\sigma}\left(1+x+\frac{t}{2}+L+\frac{\sigma}{2}\right)^{-\alpha}d\sigma\right\}\\
&\le C\left\{e^{-\frac{L}{2}}e^{-\frac{|x|}{2}}e^{-\frac{t}{4}}+e^{-L}e^{-\frac{x}{4}}e^{-\frac{t}{4}}+e^{-L}e^{-\frac{t}{16}}(1+x+L)^{-\alpha}\right.\\
&\quad\left.+e^{-L}(1+x+L+t)^{-\alpha}\right\}\\
&\le C\left\{e^{-\frac{L}{2}}e^{-\frac{|x|}{2}}e^{-\frac{t}{4}}+e^{-L}(1+x+L+t)^{-\alpha}\right\}.
\end{aligned}
$$

By adding g_1 and g_2 we obtain (2.18).

The estimate for (2.19) is similar to that for (2.18) so we omit the detail. □

LEMMA 2.6. *For $0\le x+L+t\le 1$, there exists a constant $C>0$ such that*

$$
\begin{aligned}
|K^B(x,t;y,\sigma)|\le C&\frac{|x+L+t|}{\sqrt{t-\sigma}}\\
&\cdot\int_{-1}^1 k\left(\frac{y+L+\sigma+\theta(x+L+t)}{1.5},t-\sigma\right)d\theta e^{-(y+L+\sigma)-(t-\sigma)},
\end{aligned}
$$

and

$$
\begin{aligned}
|K_y^B(x,t;y,\sigma)|\le C&\frac{|x+L+t|}{\sqrt{t-\sigma}}\left(1+\frac{1}{\sqrt{t-\sigma}}\right)\\
&\cdot\int_{-1}^1 k\left(\frac{y+L+\sigma+\theta(x+L+t)}{1.5},t-\sigma\right)d\theta e^{-(y+L+\sigma)-(t-\sigma)}.
\end{aligned}
$$

Proof. Let $X=x+L+t$ and $Y=y+L+\sigma$, $0\le X\le 1$. Then the estimates of K^B and K_y^B follow from the following expressions:

$$
\begin{aligned}
K^B&=k(x-y-(t-\sigma),t-\sigma)-k(x+y-t+3\sigma+2L,t-\sigma)e^{-2(y+L+\sigma)}\\
&=[k(Y-X,t-\sigma)-k(Y+X,t-\sigma)]e^{(X-Y)-(t-\sigma)}\\
&=-X\int_{-1}^1 k_Y(Y+\theta X,t-\sigma)d\theta e^{(X-Y)-(t-\sigma)}\\
&=O(1)\frac{X}{\sqrt{t-\sigma}}\int_{-1}^1 k\left(\frac{Y+\theta X}{1.5},t-\sigma\right)d\theta e^{-Y-(t-\sigma)},\\
K_y^B&=[k_Y(Y-X,t-\sigma)-k_Y(Y+X,t-\sigma)]e^{(X-Y)-(t-\sigma)}-K^B\\
&=-X\int_{-1}^1 k_{YY}(Y+\theta X,t-\sigma)d\theta e^{(X-Y)-(t-\sigma)}-K^B\\
&=O(1)\frac{X}{\sqrt{t-\sigma}}\left(1+\frac{1}{\sqrt{t-\sigma}}\right)\int_{-1}^1 k\left(\frac{Y+\theta X}{1.5},t-\sigma\right)d\theta e^{-Y-(t-\sigma)}.
\end{aligned}
$$

□

With the above lemmas we are now ready to estimate the leading term v^0 of iterations (2.11)–(2.14).

PROPOSITION 2.7. *There exist constants $C_0 > 0$ and $r > 1$ such that*

$$v^0(x,t) \le C_0 \begin{cases} e^{-\frac{L}{2r}} e^{-\frac{|x|}{2}} |x+L+t| & \text{for } x \in [-L-t, -L-t+1], \\ e^{-\frac{L}{2r}} e^{-\frac{|x|}{2}} & \text{for } x \in [-L-t, 0), \\ e^{-\frac{L}{2r}} e^{-\frac{|x|}{2}} + e^{-\frac{L}{2}} (1+x+L+t)^{-\alpha} & \text{for } x \ge 0, \end{cases}$$

$$v^0_x(-L-t,t) \le C_0 e^{-\frac{L}{2r}} e^{-\frac{L+t}{2}}.$$

Proof. From the property of K^B in Lemma 2.6, we have, for $x \in [-L-t, -L-t+1]$,

$$\begin{aligned} &\int_0^t \int_{-L-\sigma}^{\infty} K^B(x,t;y,\sigma) e^{-L-\sigma} (1+y+L+\sigma)^{-\alpha} dy d\sigma \\ &\qquad \le C|x+L+t| e^{-s(L+t)}, \quad 0 < s < 1, \end{aligned} \tag{2.20}$$

and by Fubini's theorem and integration by parts, with $X = x+L+t$, $Y = y+L$,

$$\begin{aligned} &\left| \int_{-L}^{\infty} K^B(x,t;y,0) v(y,0) dy \right| \\ &\quad = \left| \int_0^{\infty} [k(Y-X,t) - k(Y+X,t)] e^{(X-Y)-t} v(Y,0) dY \right| \\ &\quad = \left| \int_0^{\infty} \int_{-1}^{1} (-X) \partial_Y k(Y+\theta X, t) d\theta e^{(X-Y)-t} v(Y,0) dY \right| \\ &\quad = \left| (-X) \int_{-1}^{1} \int_0^{\infty} \partial_Y k(Y+\theta X, t) e^{(X-Y)-t} v(Y,0) dY d\theta \right| \\ &\quad = \left| X \int_{-1}^{1} \int_0^{\infty} k(Y+\theta X, t) [-e^{(X-Y)-t} v(Y,0) + e^{(X-Y)-t} v_Y(Y,0)] dY d\theta \right| \\ &\quad \le CXe^{-L} e^{-t}. \end{aligned} \tag{2.21}$$

Here we have used $|v(Y,0)| + |v_Y(Y,0)| = O(1)e^{-L}$ from (2.3). In particular we have from (2.20) and (2.21) that $v^0(-L-t,t) = 0$ and

(2.22)

$$\begin{aligned} &v^0_x(-L-t,t) \\ &= \lim_{x+L+t \to 0^+} \left\{ \int_{-L}^{\infty} \frac{K^B(x,t;y,0)}{|x+L+t|} v(y,0) dy \right. \\ &\qquad \left. + \int_0^t \int_{-L-\sigma}^{\infty} \frac{K^B(x,t;y,\sigma)}{|x+L+t|} \left\{ -\Psi_\sigma - \left(\phi\Psi + \frac{\Psi^2}{2} \right)_y + \Psi_{yy} \right\} dy d\sigma \right\} \\ &\le Ce^{-L-t} \\ &\quad + \int_0^t \int_{-L-\sigma}^{\infty} \frac{C}{\sqrt{t-\sigma}} \int_{-1}^{1} k\left(\frac{(x+L+t)\theta + (y+L+\sigma)}{1.5}, t-\sigma \right) e^{-(y+L+\sigma)-(t-\sigma)} d\theta \\ &\qquad \cdot \left\{ -\Psi_\sigma - \left(\phi\Psi + \frac{\Psi^2}{2} \right)_y + \Psi_{yy} \right\} dy d\sigma \\ &\le Ce^{-s(L+t)}, \quad 0 < s < 1. \end{aligned}$$

Since $K^B(x,t;y,\sigma) \le k(x-y-(t-\sigma), t-\sigma)$, we may apply Lemmas 2.4 and 2.5 and (2.20)–(2.22) by letting $s = \frac{1}{2} + \frac{1}{2r}$ for $r > 1$. This proves the lemma. ∎

Remark 2.8. For general initial data (1.4) with $d > 1$, v^0 is estimated as

$$v^0(x,t) \le O(1) \begin{cases} |x+L+t| e^{-\frac{|x|}{2d}} e^{-\frac{L+t}{2d}} & \text{for } x \in [-L-t, -L-t+1], \\ e^{-\frac{|x|}{2d}} e^{-\frac{L}{2d}} & \text{for } x \in [-L-t, 0), \\ e^{-\frac{|x|}{2}} \max\{e^{-\frac{L}{d}}, e^{-\frac{L}{2}}\} + (1+x+L+t)^{-\alpha} e^{-\frac{L}{d}} & \text{for } x \ge 0, \end{cases}$$

$$v^0_x(-L-t,t) \le O(1) e^{-\frac{L}{2d}} e^{-\frac{(L+t)}{2d}}.$$

Our main theorem of this section, Theorem 2.13 on the solution $v(x,t)$, will be proved by induction. The general procedure is to estimate v^n for all $n \in N$. The ansatz is more complicated than that of v^0 because of nonlinearity in (2.1) and the singularity in $K^B_y(x,t;y,\sigma)$ and $G_y(x,t;y,\sigma)$. To deal with the singularity, we need the following lemmas.

LEMMA 2.9. *For $|x+L+t| < 1$ and L large, there exists a positive constant C such that*

$$\int_0^t \int_{-L-\sigma}^{-L-\sigma+4} |K^B_y(x,t;y,\sigma)|(y+L+\sigma) e^{-|y|} dy d\sigma \le C|x+L+t| e^{-\frac{|x|}{2}} e^{-\frac{L}{2}}, \tag{2.23}$$

$$\int_0^t \int_{-L-\sigma+4}^{0} |K^B_y(x,t;y,\sigma)| e^{-|y|} dy d\sigma \le C|x+L+t| e^{-\frac{|x|}{2}} e^{-\frac{L}{2}}, \tag{2.24}$$

$$\int_0^t \int_{0}^{\infty} |K^B_y(x,t;y,\sigma)| dy d\sigma \le C|x+L+t| e^{-\frac{|x|}{2}} e^{-\frac{L}{2}}. \tag{2.25}$$

Proof. If $t > 1$,

$$\begin{aligned} &\int_0^t \int_{-L-\sigma}^{-L-\sigma+4} |K^B_y(x,t;y,\sigma)|(y+L+\sigma) e^{-|y|} dy d\sigma \\ &\quad = \left(\int_0^{t-1} + \int_{t-1}^t \right) \int_{-L-\sigma}^{-L-\sigma+4} |K^B_y(x,t;y,\sigma)|(y+L+\sigma) e^{-|y|} dy d\sigma \\ &\quad \equiv J_1 + J_2. \end{aligned}$$

If $X = x+L+t$, $Y = y+L+\sigma$, and $L > 0$ is large, then from the property of K^B_y in Lemma 2.6 we have

$$\begin{aligned} J_1 &= \int_0^{t-1} \int_{-L-\sigma}^{-L-\sigma+4} |K^B_y(x,t;y,\sigma)|(y+L+\sigma) e^{-|y|} dy d\sigma \\ &\le \int_0^{t-1} \int_0^4 C \frac{X}{\sqrt{t-\sigma}} \left(1 + \frac{1}{\sqrt{t-\sigma}}\right) \int_{-1}^1 k\left(\frac{Y+\theta X}{1.5}, t-\sigma\right) d\theta \\ &\qquad e^{-Y-(t-\sigma)} Y e^{(Y-L-\sigma)} dY d\sigma \\ &\le \int_0^{t-1} CX d\sigma e^{-(L+t)} \quad (t-\sigma > 1,\ 0 < Y < 4) \\ &\le CX(t-1) e^{-(L+t)} \le CX e^{x/2} e^{-\frac{L}{2}}, \\ J_2 &= \int_{t-1}^t \int_{-L-\sigma}^{-L-\sigma+4} |K^B_y(x,t;y,\sigma)|(y+L+\sigma) e^{-|y|} dy d\sigma \\ &\le \int_{t-1}^t \int_{-L-\sigma}^{-L-\sigma+4} K^B(x,t;y,\sigma)(y+L+\sigma) e^{-|y|} dy d\sigma \end{aligned}$$

$$+\int_{t-1}^{t}\int_{0}^{2X}|k_Y(Y-X,t-\sigma)-k_Y(Y+X,t-\sigma)|e^{-Y-(t-\sigma)}Ye^{(Y-L-\sigma)}dYd\sigma$$
$$+\int_{t-1}^{t}\int_{2X}^{4}|k_Y(Y-X,t-\sigma)-k_Y(Y+X,t-\sigma)|e^{-Y-(t-\sigma)}Ye^{(Y-L-\sigma)}dYd\sigma$$
$$\equiv j_1+j_2+j_3.$$

The terms j_1, j_2, and j_3 are estimated using Lemma 2.6:

$$j_1=\int_{t-1}^{t}\int_{-L-\sigma}^{-L-\sigma+4}|K^B(x,t;y,\sigma)|(y+L+\sigma)e^{-|y|}dyd\sigma$$
$$\le C\int_{t-1}^{t}\int_{0}^{4}\frac{X}{\sqrt{t-\sigma}}\int_{-1}^{1}k\left(\frac{Y+\theta X}{1.5},t-\sigma\right)d\theta e^{-Y-(t-\sigma)}Ye^{(Y-L-\sigma)}dYd\sigma$$
$$\le C\int_{t-1}^{t}\frac{X}{\sqrt{t-\sigma}}d\sigma e^{-(L+t)}$$
$$\le C|x+L+t|e^{x/2}e^{-\frac{L}{2}},$$
$$j_2=\int_{t-1}^{t}\int_{0}^{2X}|k_Y(Y-X,t-\sigma)-k_Y(Y+X,t-\sigma)|e^{-Y-(t-\sigma)}Ye^{(Y-L-\sigma)}dYd\sigma$$
$$\le C\int_{t-1}^{t}\int_{0}^{2X}\frac{2X}{\sqrt{t-\sigma}}k\left(\frac{Y-X}{2},t-\sigma\right)e^{-Y-(t-\sigma)}e^{(Y-L-\sigma)}dYd\sigma$$
$$\le C\int_{t-1}^{t}\frac{2X}{\sqrt{t-\sigma}}d\sigma e^{-(L+t)}$$
$$\le C|x+L+t|e^{x/2}e^{-\frac{L}{2}},$$
$$j_3=\int_{t-1}^{t}\int_{2X}^{4}|k_Y(Y-X,t-\sigma)-k_Y(Y+X,t-\sigma)|e^{-Y-(t-\sigma)}Ye^{(Y-L-\sigma)}dYd\sigma$$
$$=\int_{t-1}^{t}\int_{2X}^{4}\left|-X\int_{-1}^{1}k_{YY}(Y+\theta X,t-\sigma)d\theta\right|e^{-Y-(t-\sigma)}Ye^{(Y-L-\sigma)}dYd\sigma$$
$$\le C\int_{t-1}^{t}\int_{2X}^{4}\int_{-1}^{1}\frac{X}{t-\sigma}k\left(\frac{Y+\theta X}{1.5},t-\sigma\right)d\theta e^{-Y-(t-\sigma)}Ye^{(Y-L-\sigma)}dYd\sigma$$
$$\le C\int_{t-1}^{t}\int_{2X}^{4}\frac{XY}{t-\sigma}k\left(\frac{Y}{3},t-\sigma\right)e^{-(L+t)}dYd\sigma$$
$$\le C\int_{t-1}^{t}\frac{X}{\sqrt{t-\sigma}}d\sigma e^{-(L+t)}$$
$$\le C|x+L+t|e^{x/2}e^{-\frac{L}{2}}.$$

Thus we have obtained (2.23) for $t>1$. For $t\le 1$ the calculation is the same as for J_2 and is omitted.

Next, for $0\le X\le 1$ and $Y\ge 4$, K_y^B in Lemma 2.6 satisfies

$$|K_y^B(x,t;y,\sigma)|\le C\frac{X}{\sqrt{t-\sigma}}\left(1+\frac{1}{\sqrt{t-\sigma}}\right)k\left(\frac{Y-1}{1.5},t-\sigma\right)e^{-Y-(t-\sigma)}$$
$$\le CXk\left(\frac{Y-1}{2},t-\sigma\right)e^{-Y-(t-\sigma)},$$

and therefore

$$\begin{aligned}
&\int_0^t \int_{-L-\sigma+4}^0 |K_y^B(x,t;y,\sigma)| e^{-|y|} dy d\sigma \\
&\quad \le \int_0^t \int_4^{L+\sigma} CXk\left(\frac{Y-1}{2}, t-\sigma\right) e^{-Y-(t-\sigma)} e^{(Y-L-\sigma)} dY d\sigma \\
&\quad \le CX \int_0^t d\sigma e^{-(L+t)} = CXte^{-(L+t)} \\
&\quad \le C|x+L+t|e^{x/2} e^{-\frac{L}{2}},
\end{aligned}$$

and

$$\begin{aligned}
&\int_0^t \int_0^\infty |K_y^B(x,t;y,\sigma)| dy d\sigma \\
&\quad \le \int_0^t \int_0^\infty C|x+L+t| k\left(\frac{y+L+\sigma-1}{2}, t-\sigma\right) e^{-(y+L+\sigma)-(t-\sigma)} dy d\sigma \\
&\quad \le C|x+L+t|te^{-(L+t)} \le C|x+L+t|e^{-\frac{|x|}{2}} e^{-\frac{L}{2}}.
\end{aligned}$$

This establishes (2.24), (2.25) and completes the proof of the lemma. □

LEMMA 2.10. *For $x \in [-L-t, -(L+t)/2)$, $D > 1$,*

$$\int_0^t \int_{-L-\sigma}^0 k(x-y-(t-\sigma), t-\sigma) e^y dy d\sigma = O(1)te^x, \tag{2.26}$$

$$\int_0^t \int_0^\infty k(x-y-(t-\sigma), t-\sigma) e^{-y/2} dy d\sigma = O(1)\sqrt{t}e^x, \tag{2.27}$$

$$\int_0^t \int_0^\infty k(x-y-(t-\sigma), t-\sigma)(1+y+L+\sigma)^{-\alpha} dy d\sigma = O(1)\sqrt{t}e^x, \tag{2.28}$$

$$\int_0^t \int_{-L-\sigma}^0 \frac{1}{\sqrt{t-\sigma}} k\left(\frac{x-y+(t-\sigma)}{D}, t-\sigma\right) e^{x-y} e^{-|y|} dy d\sigma = O(1)\sqrt{t}e^x, \tag{2.29}$$

$$\int_0^t \int_0^\infty \frac{1}{\sqrt{t-\sigma}} k\left(\frac{x-y+(t-\sigma)}{D}, t-\sigma\right) e^{x-y} e^{-\frac{|y|}{2}} dy d\sigma = O(1)\sqrt{t}e^x, \tag{2.30}$$

$$\begin{aligned}
&\int_0^t \int_0^\infty \frac{1}{\sqrt{t-\sigma}} k\left(\frac{x-y+(t-\sigma)}{D}, t-\sigma\right) e^{x-y} (1+y+L+\sigma)^{-\alpha} dy d\sigma \\
&\quad = O(1)\sqrt{t}e^x.
\end{aligned} \tag{2.31}$$

Proof. The last three estimates, (2.29), (2.30), and (2.31), are obvious from

$$\int_{-\infty}^\infty k\left(\frac{x-y+(t-\sigma)}{D}, t-\sigma\right) dy = O(1).$$

The first three estimates, (2.26), (2.27), and (2.28), are obtained from the following equations, which hold for $x < 0$:

$$\begin{aligned}
&\int_{-L-\sigma}^0 k(x-y-(t-\sigma), t-\sigma) e^y dy \\
&\quad = \int_{-L-\sigma}^0 k(x-y+(t-\sigma), t-\sigma) e^x dy \le e^x,
\end{aligned}$$

$$\int_0^\infty k(x-y-(t-\sigma),t-\sigma)e^{-y/2}dy$$
$$= \int_0^\infty \frac{1}{\sqrt{4\pi(t-\sigma)}} e^{-\frac{x^2}{4(t-\sigma)}+\frac{x}{2}-y-\frac{t-\sigma}{4}}dy$$
$$= \frac{1}{\sqrt{4\pi(t-\sigma)}} e^{-\frac{(x-(t-\sigma))^2}{4(t-\sigma)}} \le \frac{1}{\sqrt{4\pi(t-\sigma)}} e^x,$$
$$\int_0^\infty k(x-y-(t-\sigma),t-\sigma)(1+y+L+\sigma)^{-\alpha}dy$$
$$\le \int_0^\infty k(x-(t-\sigma),t-\sigma)e^{-y/2}(1+y+L+\sigma)^{-\alpha}dy$$
$$\le Ck(x-(t-\sigma),t-\sigma) \le \frac{C}{\sqrt{t-\sigma}} e^x. \qquad \square$$

LEMMA 2.11. *For $x \ge -(L+t)/2$,*

$$\int_0^t \int_{-L-\sigma}^0 |g_x(x,t;y,\sigma)|e^{y/2}dyd\sigma = O(1)e^{-\frac{|x|}{2}}, \tag{2.32}$$

$$\int_0^t \int_0^\infty |g_x(x,t;y,\sigma)|e^{-y/2}dyd\sigma = O(1)e^{-\frac{|x|}{2}}. \tag{2.33}$$

Proof. These estimates follow immediately from the inequality

$$|g_x(x,t;y,\sigma)| \le \frac{4}{\sqrt{t-\sigma}}\left(1+\frac{1}{\sqrt{t-\sigma}}\right) e^{-\frac{1}{2}(|x|-|y|)} e^{-\frac{t-\sigma}{4}} e^{-\frac{(x-y)^2}{8(t-\sigma)}}. \qquad \square$$

LEMMA 2.12. *For $x \ge -(L+t)/2$,*

$$\int_0^t \int_0^\infty |g_x(x,t;y,\sigma)|(1+y+L+\sigma)^{-2\alpha}dyd\sigma$$
$$= O(1)\begin{cases} e^{-\frac{|x|}{2}}(1+t)^{-2\alpha+1} & \text{if } x \in [-(L+t)/2, 0), \\ (1+x+L+t)^{-\alpha} & \text{if } x \ge 0. \end{cases} \tag{2.34}$$

Proof. First we note that

$$g_x(x,t;y,\sigma)$$
$$= -\frac{1}{2}\tanh\frac{x}{2} g(x,t;y,\sigma) - \frac{(x-y)}{2(t-\sigma)} g(x,t;y,\sigma)$$
$$= O(1)\begin{cases} \left(1+\frac{1}{\sqrt{t-\sigma}}\right)e^{-|x|}k\left(\frac{x-y+(t-\sigma)}{D}, t-\sigma\right) & \text{for } x<0,\ y>0, \\ \left(1+\frac{1}{\sqrt{t-\sigma}}\right)k\left(\frac{x-y+(t-\sigma)}{D}, t-\sigma\right) & \text{for } x>0,\ y>0, \end{cases} \quad \text{for } D>1.$$

This yields, for $x \in [-(L+t)/2, -t/2]$,

$$\begin{aligned}
&\int_0^t \int_0^\infty |g_x(x,t;y,\sigma)|(1+y+L+\sigma)^{-2\alpha} dyd\sigma \\
&\quad \le C \int_0^t \int_0^\infty \left(1+\frac{1}{\sqrt{t-\sigma}}\right) e^{-|x|} k\left(\frac{x-y+(t-\sigma)}{D}, t-\sigma\right) \\
&\qquad\qquad \cdot (1+y+L+\sigma)^{-2\alpha} dyd\sigma \\
&\quad \le Ce^{-|x|} \int_0^t \left(1+\frac{1}{\sqrt{t-\sigma}}\right)(1+L+\sigma)^{-2\alpha} d\sigma \\
&\quad \le Ce^{-|x|} = O(1)e^{-\frac{|x|}{2}} e^{-\frac{t}{4}};
\end{aligned}$$

for $x \in (-t/2, 0)$,

$$\begin{aligned}
&\int_0^t \int_0^\infty |g_x(x,t;y,\sigma)|(1+y+L+\sigma)^{-2\alpha} dyd\sigma \\
&\quad \le C \left\{ \int_0^{\frac{t}{4}} \left(\int_0^{\frac{t-\sigma}{8}} + \int_{\frac{t-\sigma}{8}}^\infty \right) \left(1+\frac{1}{\sqrt{t-\sigma}}\right) e^{-|x|} k\left(\frac{x-y+(t-\sigma)}{D}, t-\sigma\right) \right. \\
&\qquad\qquad \cdot (1+y+L+\sigma)^{-2\alpha} dyd\sigma \\
&\qquad + \int_{\frac{t}{4}}^t \int_0^\infty \left(1+\frac{1}{\sqrt{t-\sigma}}\right) e^{-|x|} k\left(\frac{x-y+(t-\sigma)}{D}, t-\sigma\right) \\
&\qquad\qquad \left. \cdot (1+y+L+\sigma)^{-2\alpha} dyd\sigma \right\} \\
&\quad \le Ce^{-\frac{|x|}{2}} (1+t)^{-2\alpha+1};
\end{aligned}$$

and, for $x \ge 0$,

$$\begin{aligned}
&\int_0^t \int_0^\infty |g_x(x,t;y,\sigma)|(1+y+L+\sigma)^{-2\alpha} dyd\sigma \\
&\quad \le C \int_0^t \left(\int_0^{\frac{x}{2}} + \int_{\frac{x}{2}}^{x+\frac{t-\sigma}{2}} + \int_{x+\frac{t-\sigma}{2}}^\infty \right) \left(1+\frac{1}{\sqrt{t-\sigma}}\right) \\
&\qquad\qquad \cdot k\left(\frac{x-y+(t-\sigma)}{D}, t-\sigma\right)(1+y+L+\sigma)^{-2\alpha} dyd\sigma \\
&\quad \equiv A_1 + A_2 + A_3.
\end{aligned}$$

As in the proof of Lemma 2.5, we have broken the above integral into the terms A_1, A_2, and A_3 according to various regions. The estimates for these terms are straightforward:

$$\begin{aligned}
A_1 &= \int_0^t \int_0^{\frac{x}{2}} \left(1+\frac{1}{\sqrt{t-\sigma}}\right) k\left(\frac{x-y+(t-\sigma)}{D}, t-\sigma\right)(1+y+L+\sigma)^{-2\alpha} dyd\sigma \\
&\le C \int_0^t \left(1+\frac{1}{\sqrt{t-\sigma}}\right) e^{-\frac{(\frac{x}{2}+t-\sigma)^2}{4D(t-\sigma)}} (1+L+\sigma)^{-2\alpha} d\sigma \\
&\le Ce^{-\frac{x}{4D}} \left(\int_0^{\frac{t}{2}} + \int_{\frac{t}{2}}^t \right) e^{-\frac{t-\sigma}{4D}} \left(1+\frac{1}{\sqrt{t-\sigma}}\right)(1+L+\sigma)^{-2\alpha} d\sigma \\
&\le Ce^{-\frac{x}{4D}} (e^{-\frac{t}{8D}} + (1+L+t)^{-2\alpha})
\end{aligned}$$

$$\le C(1+x+L+t)^{-2\alpha},$$

$$A_2 = \int_0^t \int_{\frac{x}{2}}^{x+\frac{t-\sigma}{2}} \left(1+\frac{1}{\sqrt{t-\sigma}}\right) k\left(\frac{x-y+(t-\sigma)}{D}, t-\sigma\right)(1+y+L+\sigma)^{-2\alpha} dy d\sigma$$

$$\le C\int_0^t \left(1+\frac{1}{\sqrt{t-\sigma}}\right) e^{-\frac{(\frac{t-\sigma}{2})^2}{4D(t-\sigma)}} \left(1+\frac{x}{2}+L+\sigma\right)^{-2\alpha} d\sigma$$

$$= C\left(\int_0^{\frac{t}{2}} + \int_{\frac{t}{2}}^t\right)\left(1+\frac{1}{\sqrt{t-\sigma}}\right) e^{-\frac{t-\sigma}{16D}}(1+x+L+\sigma)^{-2\alpha} d\sigma$$

$$\le C\left[e^{-\frac{t}{32D}}(1+x+L)^{-2\alpha}\int_0^{\frac{t}{2}}\left(1+\frac{1}{\sqrt{t-\sigma}}\right)d\sigma \right.$$

$$\left. + (1+x+L+t)^{-2\alpha}\int_{\frac{t}{2}}^t \left(1+\frac{1}{\sqrt{t-\sigma}}\right) e^{-\frac{t-\sigma}{16D}} d\sigma\right]$$

$$\le C(1+x+L+t)^{-2\alpha},$$

$$A_3 = \int_0^t \int_{x+\frac{t-\sigma}{2}}^{\infty} \left(1+\frac{1}{\sqrt{t-\sigma}}\right) k\left(\frac{x-y+(t-\sigma)}{D}, t-\sigma\right)(1+y+L+\sigma)^{-2\alpha} dy d\sigma$$

$$\le C\int_0^t \left(1+\frac{1}{\sqrt{t-\sigma}}\right)\left(1+x+L+\frac{t}{2}+\frac{\sigma}{2}\right)^{-2\alpha} d\sigma$$

$$\le C(1+x+L+t)^{-2\alpha+1}.$$

By adding A_1, A_2, and A_3 we obtain (2.34) in the case $x \ge 0$. □

Finally, we can establish the basic structure of the global solution of the problem (2.1), (2.2), and (2.3).

THEOREM 2.13. *Suppose that L is sufficiently large. Then there exists a global solution $v(x,t)$ of* (2.1), (2.2), *and* (2.3) *satisfying the following estimate:*

$$v(x,t) = O(1)e^{-\frac{L}{2r}} \begin{cases} e^{-\frac{|x|}{2}}|x+L+t| & \text{for } x \in [-L-t, -L-t+1], \\ e^{-\frac{|x|}{2}} & \text{for } x \in [-L-t, 0), \\ e^{-\frac{|x|}{2}} + (1+x+L+t)^{-\alpha} & \text{for } x \ge 0. \end{cases} \tag{2.35}$$

Moreover, $v_x(-L-t,t)$ exists and has the estimate

$$v_x(-L-t,t) = O(1)e^{-\frac{L}{2r}}e^{-\frac{(L+t)}{2}}, \quad r>1. \tag{2.36}$$

Proof. From Proposition 2.7. we have proved that v^0 satisfies estimates (2.35) and (2.36). Now, suppose that $v^n(x,t)$, for any $n \le k$, satisfies estimate (2.35). We define a weighted super norm $|||\cdot|||$ as follows:

$$|||y||| \equiv \sup_{\substack{t\ge 0 \\ -L-t\le x<0}} \frac{|y(x,t)|}{e^{-\frac{|x|}{2}}} + \sup_{\substack{t\ge 0 \\ x\ge 0}} \frac{|y(x,t)|}{e^{-\frac{|x|}{2}} + (1+x+L+t)^{-\alpha}}.$$

Let

$$\delta^n(x,t) \equiv v^n(x,t) - v^{n-1}(x,t) \quad \text{for } n \ge 1;$$

then it suffices to show that δ^n is a geometric sequence by induction. We have the following representation for δ^n's:

for $x \in [-L-t, -(L+t)/2)$,

$$\delta^1(x,t) = \int_0^t \int_{-L-\sigma}^{\infty} K_y^B(x,t;y,\sigma)\left[(\phi - 1 + \Psi)v^0 + \frac{(v^0)^2}{2}\right](y,\sigma)dyd\sigma; \tag{2.37}$$

for $x \geq -(L+t)/2$,

$$\begin{aligned}\delta^1(x,t) = &-\int_0^t G(x,t;-L-\sigma,\sigma)\delta_y^1(-L-\sigma,\sigma)d\sigma \\ &+\int_0^t \int_{-L-\sigma}^{\infty} G_y(x,t;y,\sigma)\left[\frac{(v^0)^2}{2} + \Psi v^0\right](y,\sigma)dyd\sigma;\end{aligned} \tag{2.38}$$

for $x \in [-L-t, -(L+t)/2)$, $n \geq 1$,

$$\begin{aligned}\delta^{n+1}(x,t) = &\int_0^t \int_{-L-\sigma}^{\infty} K_y^B(x,t;y,\sigma) \\ &\cdot \left[(\phi - 1 + \Psi)\delta^n + \frac{\delta^n(v^n + v^{n-1})}{2}\right](y,\sigma)dyd\sigma;\end{aligned} \tag{2.39}$$

and, for $x \geq -(L+t)/2$, $n \geq 1$,

$$\begin{aligned}\delta^{n+1}(x,t) = &-\int_0^t G(x,t;-L-\sigma,\sigma)\delta_y^{n+1}(-L-\sigma,\sigma)d\sigma \\ &+\int_0^t \int_{-L-\sigma}^{\infty} G_y(x,t;y,\sigma)\left[\Psi\delta^n + \frac{\delta^n(v^n + v^{n-1})}{2}\right](y,\sigma)dyd\sigma.\end{aligned} \tag{2.40}$$

Since $\Psi(x,t) \leq 1 - \phi(x)$, either less than e^x if $-L-t \leq x \leq 0$ or 2 if $x > 0$, and

$$\begin{aligned}|K_y^B(x,t;y,\sigma)| &\leq K^B + (|k_Y(Y-X,t-\sigma)| + |k_Y(Y+X,t-\sigma)|)e^{(X-Y)-(t-\sigma)} \\ &\leq C\left\{k(x-y-(t-\sigma),t-\sigma)\right. \\ &\qquad \left. + \frac{1}{\sqrt{t-\sigma}}k\left(\frac{x-y+(t-\sigma)}{D}, t-\sigma\right)e^{x-y}\right\} \quad \text{for all } D > 1,\end{aligned}$$

we have from Lemma 2.10 and (2.37) that, for $x \in [-L-t, -(L+t)/2)$,

$$\begin{aligned}|\delta^1(x,t)| &\leq C_0 \int_0^t \int_{-L-\sigma}^{0} |K_y^B(x,t;y,\sigma)| \\ &\qquad \cdot \left[|||v^0|||e^{-|y|/2}(2(1-\phi(y))) + \frac{1}{2}|||v^0|||e^{-\frac{L}{2r}}e^{-|y|}\right]dyd\sigma \\ &\quad + C_0 \int_0^t \int_0^{\infty} |K_y^B(x,t;y,\sigma)| \\ &\qquad \cdot \left[|||v^0|||(e^{-|y|/2} + (1+y+L+\sigma)^{-\alpha})2(1-\phi(y))\right. \\ &\qquad \left. + \frac{1}{2}|||v^0|||e^{-\frac{L}{2r}}\left(e^{-|y|/2} + (1+y+L+\sigma)^{-\alpha}\right)^2\right]dyd\sigma \\ &\leq C|||v^0|||\int_0^t \int_{-L-\sigma}^{0} |K_y^B(x,t;y,\sigma)| \cdot e^{-\frac{|y|}{2}}\left(e^{-|y|} + e^{-\frac{L}{2r}}e^{-|y|/2}\right)dyd\sigma\end{aligned} \tag{2.41}$$

$$
\begin{aligned}
&+ C|||v^0||| \int_0^t \int_0^\infty |K_y^B(x,t;y,\sigma)| \\
&\qquad \cdot \left[\left(e^{-|y|/2} + (1+y+L+\sigma)^{-\alpha} \right) \right. \\
&\qquad \left. + e^{-\frac{L}{2r}} \left(e^{-|y|} + (1+y+L+\sigma)^{-2\alpha} \right) \right] dyd\sigma \\
&\le C e^{-\frac{L}{2r}} e^{-\frac{|x|}{2}} |||v^0|||.
\end{aligned}
$$

From Lemma 2.9 and the fact that $\phi(x) - 1 + \Psi(x,t) = O(1)|x+L+t|e^{-|x|}$, for $-L-t \le x < 0$, we obtain, for $x \in [-L-t, -L-t+1)$,

$$
\begin{aligned}
|\delta^1(x,t)| &\le C \int_0^t \int_{-L-\sigma}^{-L-\sigma+4} |K_y^B(x,t;y,\sigma)|(y+L+\sigma) \\
&\qquad \cdot |||v^0|||e^{-|y|/2} \left[e^{-|y|} + \frac{1}{2} e^{-\frac{L}{2r}} e^{-\frac{|y|}{2}} \right] dyd\sigma \\
&+ C \int_0^t \int_{-L-\sigma+4}^0 |K_y^B(x,t;y,\sigma)| \\
&\qquad \cdot \left[|||v^0|||e^{-|y|/2}(2(1-\phi(y))) + \frac{1}{2}|||v^0|||e^{-\frac{L}{2r}} e^{-|y|} \right] dyd\sigma \\
&+ C \int_0^t \int_0^\infty |K_y^B(x,t;y,\sigma)| \\
&\qquad \cdot \left[|||v^0||| \left(e^{-|y|/2} + (1+y+L+\sigma)^{-\alpha} \right) 2(1-\phi(y)) \right. \\
&\qquad \left. + \frac{1}{2}|||v^0|||e^{-\frac{L}{2r}} \left(e^{-|y|/2} + (1+y+L+\sigma)^{-\alpha} \right)^2 \right] dyd\sigma \\
&\le C|x+L+t|e^{-\frac{L}{2r}} e^{-\frac{|x|}{2}} |||v^0|||.
\end{aligned}
\tag{2.42}
$$

This yields $\delta^1(-L-t,t) = 0$. Furthermore, similar to (2.22) and the proof of Lemma 2.9, with $X = x+L+t$, $Y = y+L+\sigma$, we have that

$$
\begin{aligned}
&|\delta_x^1(-L-t,t)| \\
&\quad = \left| \lim_{x+L+t \to 0^+} \int_0^t \int_{-L-\sigma}^\infty \frac{-K^B(x,t;y,\sigma)}{|x+L+t|} \left[(\phi - 1 + \Psi)v^0 + \frac{(v^0)^2}{2} \right](y,\sigma) dyd\sigma \right. \\
&\qquad + \lim_{X \to 0^+} \int_0^t \int_{-L-\sigma}^\infty \frac{(k_Y(Y-X,t-\sigma) - k_Y(Y+X,t-\sigma))}{X} e^{X-Y-(t-\sigma)} \\
&\qquad\qquad \left. \cdot \left[(\phi - 1 + \Psi)v^0 + \frac{(v^0)^2}{2} \right](y,\sigma) dyd\sigma \right| \\
&\quad \le C e^{-\frac{L}{2r}} e^{-\frac{L+t}{2}} |||v^0|||.
\end{aligned}
\tag{2.43}
$$

Also, from Lemmas 2.5, 2.11, and 2.12, equations (2.38), (2.43), and the fact that

$$
G_y(x,t;y,\sigma) = -g_x(x,t;y,\sigma),
$$

we have, for $x \ge -(L+t)/2$,

$$
\begin{aligned}
|\delta^1(x,t)| \le C\Bigg\{ & \int_0^t |G(x,t;-L-\sigma,\sigma)|\,|||v^0|||e^{-\frac{L}{2r}}e^{-(L+\sigma)/2}d\sigma \\
& + \int_0^t\int_{-L-\sigma}^0 |G_y(x,t;y,\sigma)| \\
& \qquad \cdot\left[|||v^0|||e^{-\frac{|y|}{2}}e^{-L-\sigma}(1+y+L+\sigma)^{-\alpha} + \frac{1}{2}|||v^0|||e^{-\frac{L}{2r}}e^{-|y|}\right]dyd\sigma \\
& + \int_0^t\int_0^\infty |G_y(x,t;y,\sigma)| \\
& \qquad \cdot\Big[|||v^0|||\left(e^{-\frac{|y|}{2}} + (1+y+L+\sigma)^{-\alpha}\right)e^{-L-\sigma}(1+y+L+\sigma)^{-\alpha} \\
& \qquad\quad + |||v^0|||e^{-\frac{L}{2r}}(e^{-|y|} + (1+y+L+\sigma)^{-2\alpha})\Big]dyd\sigma\Bigg\} \\
\le Ce^{-\frac{L}{2r}}|||v^0|||\Bigg\{ & \begin{array}{ll} e^{-\frac{|x|}{2}} & \text{for } x\in[-(L+t)/2,0), \\ e^{-\frac{|x|}{2}} + (1+x+L+t)^{-\alpha} & \text{for } x\ge 0. \end{array}
\end{aligned}
\tag{2.44}
$$

Inequalities (2.41) and (2.44) imply that

$$
|||\delta^1||| \le Ce^{-\frac{L}{2r}}|||v^0|||. \tag{2.45}
$$

Similarly, we have from the induction hypothesis that, for $x\in[-L-t,-L-t+1)$,

(2.46)

$$
\begin{aligned}
|\delta^{n+1}(x,t)| \le & C|||\delta^n||| \\
& \cdot\Bigg\{\int_0^t\int_{-L-\sigma}^{-L-\sigma+4} |K_y^B(x,t;y,\sigma)|(y+L+\sigma)e^{-\frac{|y|}{2}}\left(e^{-|y|}+e^{-\frac{L}{2r}}e^{-\frac{|y|}{2}}\right)dyd\sigma \\
& + \int_0^t\int_{-L-\sigma+4}^0 |K_y^B(x,t;y,\sigma)|e^{-\frac{|y|}{2}}\left(e^{-|y|}+e^{-\frac{L}{2r}}e^{-\frac{|y|}{2}}\right)dyd\sigma \\
& + \int_0^t\int_0^\infty |K_y^B(x,t;y,\sigma)|\left(e^{-\frac{|y|}{2}} + (1+y+L+\sigma)^{-\alpha}\right) \\
& \qquad \cdot\left[4 + e^{-\frac{L}{2r}}\left(e^{-\frac{|y|}{2}} + (1+y+L+\sigma)^{-\alpha}\right)\right]dyd\sigma\Bigg\} \\
\le & C|x+L+t|||\delta^n|||e^{-\frac{L}{2r}}e^{-\frac{|x|}{2}}.
\end{aligned}
$$

Thus

$$
|\delta_x^{n+1}(-L-t,t)| \le C|||\delta^n|||e^{-\frac{L}{2r}}e^{-\frac{L+t}{2}}; \tag{2.47}
$$

and, for $x\in[-L-t,-(L+t)/2)$,

$$
\begin{aligned}
|\delta^{n+1}(x,t)| \le & C|||\delta^n||| \\
& \cdot\Bigg\{\int_0^t\int_{-L-\sigma}^0 |K_y^B(x,t;y,\sigma)|e^{-\frac{|y|}{2}}\left(e^{-|y|}+e^{-\frac{L}{2r}}e^{-\frac{|y|}{2}}\right)dyd\sigma \\
& + \int_0^t\int_0^\infty |K_y^B(x,t;y,\sigma)|\left(e^{-\frac{|y|}{2}} + (1+y+L+\sigma)^{-\alpha}\right) \\
& \qquad \cdot\left[4 + e^{-\frac{L}{2r}}\left(e^{-\frac{|y|}{2}} + (1+y+L+\sigma)^{-\alpha}\right)\right]dyd\sigma\Bigg\} \\
\le & C|||\delta^n|||e^{-\frac{L}{2r}}e^{-\frac{|x|}{2}};
\end{aligned}
\tag{2.48}
$$

for $x \geq -(L+t)/2$,

$$
\begin{aligned}
|\delta^{n+1}(x,t)| \leq C|||\delta^n|||
&\cdot \Bigg\{ \int_0^t |G(x,t;-L-\sigma,\sigma)| e^{-\frac{L}{2r}} e^{-(L+\sigma)/2} d\sigma \\
&+ \int_0^t \int_{-L-\sigma}^0 |G_y(x,t;y,\sigma)| e^{-\frac{|y|}{2}} \\
&\qquad \cdot \left[e^{-L-\sigma}(1+y+L+\sigma)^{-\alpha} + e^{-\frac{L}{2r}} e^{-\frac{|y|}{2}} \right] dy d\sigma \\
&+ \int_0^t \int_0^\infty |G_y(x,t;y,\sigma)| \left(e^{-\frac{|y|}{2}} + (1+y+L+\sigma)^{-\alpha} \right) \\
&\qquad \cdot \Big[e^{-L-\sigma}(1+y+L+\sigma)^{-\alpha} \\
&\qquad\quad + e^{-\frac{L}{2r}} \left(e^{-\frac{|y|}{2}} + (1+y+L+\sigma)^{-\alpha} \right) \Big] dy d\sigma \Bigg\} \\
&\leq C e^{-\frac{L}{2r}} |||\delta^n||| \begin{cases} e^{-\frac{|x|}{2}} & \text{for } x \in [-(L+t)/2, 0), \\ e^{-\frac{|x|}{2}} + (1+x+L+t)^{-\alpha} & \text{for } x \geq 0. \end{cases}
\end{aligned} \tag{2.49}
$$

From the inequalities (2.47), (2.48), and (2.49) we conclude that

$$
|||\delta^{n+1}||| \leq C|||\delta^n|| e^{-\frac{L}{2r}}, \tag{2.50}
$$

$$
|\delta_x^{n+1}(-L-t,t)| \leq C|||\delta^n||| e^{-\frac{L}{2r}} e^{-\frac{L+t}{2}}, \quad 1 \leq n \leq k. \tag{2.51}
$$

Thus, when L is sufficiently large, $|||\delta^n|||$ and $|\delta_x^{n+1}(-L-t,t)|$ are geometric sequences and

$$
|||v^{k+1} - v^0||| \leq \sum_{n=0}^{k} |||\delta^{n+1}||| \leq \sum_{n=0}^{\infty} \left(C e^{-\frac{L}{2r}} \right)^{n+1} |||v^0||| \leq \frac{1}{2} |||v^0|||, \tag{2.52}
$$

which means, by induction, that for all n, v^n satisfy the estimates (2.35) and (2.36), and there exists

$$
v(x,t) \equiv \lim_{n \to \infty} v^n(x,t)
$$

satisfying (2.35), (2.36), (2.6), and (2.10). ∎

From the above estimate of $v(x,t)$ and $v_x(-L-t,t)$, we obtain the estimate of the solution $u(x,t)$ for (1.1), (1.2), and (1.3).

THEOREM 2.14. *Suppose that L is sufficiently large. Then, there exists a global solution $u(x,t)$ of* (1.1), (1.2), *and* (1.3) *such that*

$$
u(x,t) = \phi(x) + O(1) e^{-\frac{L}{2r}} \begin{cases} e^{-\frac{|x|}{2}} & \text{for } x \in [-L-t, 0), \\ e^{-\frac{|x|}{2}} + (1+x+L+t)^{-\alpha} & \text{for } x \geq 0. \end{cases} \tag{2.53}
$$

Moreover, $u_x(-L-t,t)$ exists and has the estimate

$$
u_x(-L-t,t) = O(1) e^{-\frac{L}{2r}} e^{-\frac{(L+t)}{2}}, \quad r > 1. \tag{2.54}
$$

Remark 2.15. For general initial data (1.4) with $d > 1$, we also obtain the structure of global solution in time as follows:

$$u(x,t) = \phi(x) + O(1)e^{-\frac{L}{2d}} \begin{cases} e^{-\frac{|x|}{2d}} & \text{for } x \in [-L-t, 0), \\ e^{-\frac{|x|}{2}} + (1+x+L+t)^{-\alpha} & \text{for } x \geq 0, \end{cases}$$

$$u_x(-L-t,t) = O(1)e^{-\frac{L}{2d}} e^{-\frac{(L+t)}{2d}}.$$

The computation is similar to those in section 3 and is therefore omitted.

3. Time asymptotic stability. In this section, we will study the time asymptotic behavior of the solution for (1.1), (1.2), and (1.3). For that, we need to locate the time asymptotic position of the shock profile through the conservation law.

As a consequence of the boundary estimate (2.54), we have

$$\begin{aligned} \frac{d}{dt}\int_{-L-t}^{\infty} u(x,t) - \phi(x)dx &= (1 - \phi(-L-t)) - u_x(-L-t,t) \\ &\leq O(1)e^{-\frac{L}{6}} e^{-\frac{L+t}{2}}, \end{aligned}$$

and therefore

$$\lim_{t\to\infty} \int_{-L-t}^{\infty} u(x,t) - \phi(x)dx$$

exists. We thus obtain the time asymptotic shift x_0 of the shock location

$$x_0 \equiv \frac{\lim_{t\to\infty} \int_{-L-t}^{\infty} u(x,t) - \phi(x)dx}{u_+ - u_-} = O(1)e^{-\frac{L}{6}} e^{-\frac{L}{2}}, \tag{3.1}$$

where u_-, u_+ are the end states of the stationary Burgers shock and here are taken to be $u_- = -u_+ = 1$.

Now, consider the new variables

$$\mathbf{v}(x,t) \equiv u(x,t) - \phi(x + x_0), \tag{3.2}$$

$$w(x,t) \equiv -\int_x^{\infty} \mathbf{v}(r,t)dr. \tag{3.3}$$

Then

$$\begin{aligned} w(-L-t,t) &= -\int_{-L-t}^{\infty} u(x,t) - \phi(x+x_0)dx \\ &= -\int_{-L-t}^{\infty} u(x,t) - \phi(x)dx + \int_{-\infty}^{\infty} \phi(x+x_0) - \phi(x)dx \\ &\quad - \int_{-\infty}^{-L-t} \phi(x+x_0) - \phi(x)dx \\ &= -\int_{-L-t}^{\infty} u(x,t) - \phi(x)dx + O(1)e^{-L-t} + x_0(u_+ - u_-) \\ &\leq O(1)e^{-\frac{L}{6}} e^{-\frac{L+t}{2}}, \end{aligned}$$

$$\begin{aligned} w_x(-L-t,t) &= u(-L-t,t) - \phi(-L-t+x_0) \\ &= 1 - \phi(-L-t+x_0) = O(1)e^{-L-t} \leq O(1)e^{-\frac{L}{6}} e^{-\frac{L+t}{2}}, \end{aligned}$$

$$\begin{aligned} w_{xx}(-L-t,t) &= u_x(-L-t,t) - \phi'(-L-t+x_0) \\ &= O(1)e^{-\frac{L}{6}} e^{-\frac{L+t}{2}}. \end{aligned}$$

And the initial value for $w(x,t)$ satisfies

$$\begin{aligned} w(x,0) &= -\int_x^\infty u(r,0) - \phi(r+x_0)dr \\ &\le \frac{C}{\alpha-1}e^{-L}(1+x+L)^{-\alpha+1} + \int_x^\infty \phi(r) - \phi(r+x_0)dr \\ &\le Ce^{-L}(1+x+L)^{-\alpha+1} + C\begin{cases} |x_0|, & -L \le x \le 0, \\ |x_0|e^{-|x|}, & x > 0, \end{cases} \\ w_x(x,0) &= u(x,0) - \phi(x+x_0) \\ &\le O(1)(e^{-L}(1+x+L)^{-\alpha} + |x_0|e^{-|x|}). \end{aligned}$$

In summary, $w(x,t)$ satisfies

$$w_t + \phi(x+x_0)w_x - w_{xx} = \frac{-\mathbf{v}^2(x,t)}{2}, \tag{3.4}$$

$$w(x,0) \le O(1)\begin{cases} e^{\frac{-L}{2}-\frac{L}{6}}, & -L \le x \le 0, \\ e^{\frac{-L}{2}-\frac{L}{6d}}(x+L+1)^{-\alpha+1}, & x > 0,\ d \gtrsim 1, \end{cases}$$

$$w_x(x,0) \le O(1)\begin{cases} e^{\frac{-L}{2}-\frac{L}{6}}, & -L \le x \le 0, \\ e^{\frac{-L}{2}-\frac{L}{6d}}(x+L+1)^{-\alpha}, & x > 0,\ d \gtrsim 1, \end{cases}$$

$$w(-L-t,t) \le O(1)e^{\frac{-(L+t)}{2}}e^{\frac{-L}{6}},$$

$$w_x(-L-t,t) \le O(1)e^{\frac{-(L+t)}{2}}e^{\frac{-L}{6}},$$

$$w_{xx}(-L-t,t) \le O(1)e^{\frac{-(L+t)}{2}}e^{\frac{-L}{6}},$$

$$|x_0| \le O(1)e^{\frac{-L}{2}}e^{\frac{-L}{6}},$$

where $d \gtrsim 1$ means that $d > 1$ and is close to 1.

To simplify the boundary condition, we set

$$\begin{aligned} &\bar{w}(x,t) \equiv w(x,t) - R(x,t), \\ &\text{where} \quad R(x,t) = w(-L-t,t)e^{-(x+L+t)}. \end{aligned} \tag{3.5}$$

Then it follows from (3.4) and previous estimates that

$$\begin{aligned} &\bar{w}_t + \phi(x+x_0)\bar{w}_x - \bar{w}_{xx} + \frac{1}{2}\bar{w}_x^2 + \bar{w}_x R_x \\ &\quad + R_t + \phi(x+x_0)R_x - R_{xx} + \frac{1}{2}R_x^2 = 0, \end{aligned} \tag{3.6}$$

$$\bar{w}(-L-t,t) = 0, \tag{3.7}$$

$$\bar{w}(x,0) \le O(1)\begin{cases} e^{\frac{-L}{2}-\frac{L}{6}}, & -L-t \le x \le 0, \\ e^{\frac{-L}{2}-\frac{L}{6d}}(x+L+1)^{-\alpha+1}, & x > 0,\ d \gtrsim 1, \end{cases} \tag{3.8}$$

$$\bar{w}_x(x,0) \le O(1)\begin{cases} e^{\frac{-L}{2}-\frac{L}{6}}, & -L-t \le x \le 0, \\ e^{\frac{-L}{2}-\frac{L}{6d}}(x+L+1)^{-\alpha}, & x > 0,\ d \gtrsim 1. \end{cases} \tag{3.9}$$

As before, we have two different representations for the solution in two different regions.

Region I: $x \in [-L-t, -(L+t)/2)$. We have from (3.6) that

$$\begin{aligned}\bar{w}(x,t) = &\int_{-L}^{\infty} K^B(x,t;y,0)\bar{w}(y,0)dy\\ &+\int_0^t\int_{-L-\sigma}^{\infty} K^B(1-\phi(y+x_0))\bar{w}_y dyd\sigma\\ &-\int_0^t\int_{-L-\sigma}^{\infty} K^B\frac{(\bar{w}_y)^2}{2}dyd\sigma\\ &-\int_0^t\int_{-L-\sigma}^{\infty} K^B\bar{w}_y R_y dyd\sigma\\ &-\int_0^t\int_{-L-\sigma}^{\infty} K^B\left[R_\sigma+\phi(y+x_0)R_y-R_{yy}+\frac{1}{2}R_y^2\right]dyd\sigma.\end{aligned} \tag{3.10}$$

Region II: $x \geq \frac{-L-t}{2}$. We have from (3.6) that

$$\begin{aligned}\bar{w}(x,t) = &\int_{-L}^{\infty} \bar{g}(x,t;y,0)\bar{w}(y,0)dy\\ &-\int_0^t \bar{g}(x,t;-L-\sigma,\sigma)\bar{w}_y(-L-\sigma,\sigma)d\sigma\\ &-\int_0^t\int_{-L-\sigma}^{\infty} \bar{g}\frac{(\bar{w}_y)^2}{2}dyd\sigma\\ &-\int_0^t\int_{-L-\sigma}^{\infty} \bar{g}\bar{w}_y R_y dyd\sigma\\ &-\int_0^t\int_{-L-\sigma}^{\infty} \bar{g}\left[R_\sigma+\phi(y+x_0)R_y-R_{yy}+\frac{1}{2}R_y^2\right]dyd\sigma,\end{aligned} \tag{3.11}$$

where $\bar{g}(x,t;y,\sigma) \equiv g(x+x_0,t;y+x_0,\sigma)$.

And, as before, we use the following iterations:
for $(-L-t) \leq x < (-L-t)/2$,

$$\begin{aligned}\bar{w}^0(x,t) = &\int_{-L}^{\infty} K^B(x,t;y,0)\bar{w}(y,0)dy\\ &-\int_0^t\int_{-L-\sigma}^{\infty} K^B\left[R_\sigma+\phi(y+x_0)R_y-R_{yy}+\frac{1}{2}R_y^2\right]dyd\sigma;\end{aligned}$$

for $x \geq (-L-t)/2$,

$$\begin{aligned}\bar{w}^0(x,t) = &\int_{-L}^{\infty} \bar{g}(x,t;y,0)\bar{w}(y,0)dy\\ &-\int_0^t \bar{g}(x,t;-L-\sigma,\sigma)\bar{w}^0_y(-L-\sigma,\sigma)d\sigma\\ &-\int_0^t\int_{-L-\sigma}^{\infty} \bar{g}\left[R_\sigma+\phi(y+x_0)R_y-R_{yy}+\frac{1}{2}R_y^2\right]dyd\sigma;\end{aligned}$$

for $(-L-t) \leq x < (-L-t)/2$, $n \geq 0$,

$$\bar{w}^{n+1}(x,t) = \int_{-L}^{\infty} K^B(x,t;y,0)\bar{w}^n(y,0)dy$$

$$
\begin{aligned}
&+\int_0^t\int_{-L-\sigma}^{\infty} K^B(1-\phi(y+x_0))\bar{w}_y^n dyd\sigma\\
&-\int_0^t\int_{-L-\sigma}^{\infty} K^B\frac{(\bar{w}_y^n)^2}{2}dyd\sigma\\
&-\int_0^t\int_{-L-\sigma}^{\infty} K^B\bar{w}_y^n R_y dyd\sigma\\
&-\int_0^t\int_{-L-\sigma}^{\infty} K^B\left[R_\sigma+\phi(y+x_0)R_y-R_{yy}+\frac{1}{2}R_y^2\right]dyd\sigma;
\end{aligned}
$$

for $x \geq (-L-t)/2$, $n \geq 0$,

$$
\begin{aligned}
\bar{w}^{n+1}(x,t) = &\int_{-L}^{\infty} \bar{g}(x,t;y,0)\bar{w}^n(y,0)dy\\
&-\int_0^t \bar{g}(x,t;-L-\sigma,\sigma)\bar{w}_y^{n+1}(-L-\sigma,\sigma)d\sigma\\
&-\int_0^t\int_{-L-\sigma}^{\infty} \bar{g}\frac{(\bar{w}_y^n)^2}{2}dyd\sigma\\
&-\int_0^t\int_{-L-\sigma}^{\infty} \bar{g}\bar{w}_y^n R_y dyd\sigma\\
&-\int_0^t\int_{-L-\sigma}^{\infty} \bar{g}\left[R_\sigma+\phi(y+x_0)R_y-R_{yy}+\frac{1}{2}R_y^2\right]dyd\sigma.
\end{aligned}
$$

It is noted that

$$
\bar{w}^m(x,0) = \bar{w}(x,0) \quad \text{for all } m \in \mathbf{N}, \tag{3.12}
$$

$$
R_x(x,t) = -R(x,t) \quad \text{and} \quad |R_t(x,t)| \leq O(1)e^{\frac{-(L+t)}{2}}e^{\frac{-L}{6}}e^{-(x+L+t)} \tag{3.13}
$$

by (3.5). We will show that the boundary value is unchanged through the iteration. The proof of Lemma 3.1 contains the essential estimate for that.

LEMMA 3.1.

$$
\bar{w}^0(-L-t,t) = 0. \tag{3.14}
$$

Proof. Let $X = x+L+t\,(\geq 0)$ and $Y = y+L+\sigma\,(\geq 0)$. $K^B(x,t;y,\sigma)$ can be represented as

$$
K^B(x,t;y,\sigma) = [k(Y-X,t-\sigma)-k(Y+X,t-\sigma)]e^{(X-Y)-(t-\sigma)}. \tag{3.15}
$$

Then by (3.15) and Fubini's theorem,

$$
\begin{aligned}
&\left|\int_{-L}^{\infty} K^B(x,t;y,0)\bar{w}(y,0)dy\right|\\
&\quad=\left|(-X)\int_{-L}^{\infty}\int_{-1}^{1}\partial_y k(Y-\theta X,t)d\theta e^{(X-Y)-t}\bar{w}(y,0)dy\right|\\
&\quad=\left|(-X)\int_{-1}^{1}\int_{-L}^{\infty}\partial_y k(Y-\theta X,t)e^{(X-Y)-t}\bar{w}(y,0)dyd\theta\right|\\
&\quad=\left|X\int_{-1}^{1}\int_{-L}^{\infty} k(Y-\theta X,t)\left[-e^{(X-Y)-t}\bar{w}(y,0)+e^{(X-Y)-t}\bar{w}_y(y,0)\right]dyd\theta\right|
\end{aligned}
$$

$$
\begin{aligned}
&= \left| X \int_{-1}^{1} \int_{-L}^{\infty} \frac{1}{\sqrt{4\pi t}} e^{\frac{-(Y-X)^2}{4t}} e^{\frac{-2XY(1-\theta)}{4t}} e^{\frac{X^2(1-\theta^2)}{4t}} \right. \\
&\qquad \left. \cdot \left[-e^{(X-Y)-t} \bar{w}(y,0) + e^{(X-Y)-t} \bar{w}_y(y,0) \right] dy d\theta \right| \\
&\le X e^{\frac{X^2}{4t}} \int_{-1}^{1} e^{\frac{-X^2\theta^2}{4t}} \int_{-L}^{\infty} k(x-y-t,t)[|\bar{w}(y,0)| + |\bar{w}_y(y,0)|] dy d\theta, \\
&\le O(1) X e^{\frac{X^2}{4t}} e^{\frac{x}{2}} e^{\frac{-t}{4}} e^{\frac{-L}{6}},
\end{aligned}
$$

and

$$
\begin{aligned}
&\left| \int_0^t \int_{-L-\sigma}^{\infty} K^B(x,t;y,\sigma) R_\sigma dy d\sigma \right| \\
&\quad = \left| (-X) \int_0^t \int_{-L-\sigma}^{\infty} \int_{-1}^{1} \partial_y k(Y-\theta X, t-\sigma) d\theta e^{(X-Y)-(t-\sigma)} R_\sigma dy d\sigma \right| \\
&\quad = \left| (-X) \int_0^t \int_{-1}^{1} \int_{-L-\sigma}^{\infty} \partial_y k(Y-\theta X, t-\sigma) e^{(X-Y)-(t-\sigma)} R_\sigma dy d\theta d\sigma \right| \\
&\quad \le O(1) X \int_0^t \int_{-1}^{1} \int_{-L-\sigma}^{\infty} \frac{1}{(t-\sigma)} e^{\frac{-(Y-\theta X)^2}{4\bar{d}(t-\sigma)}} e^{(x-y)} \\
&\qquad\qquad \cdot e^{\frac{-(L+\sigma)}{2}} e^{\frac{-L}{6}} e^{-(y+L+\sigma)} dy d\theta d\sigma \\
&\quad \le O(1) X e^x \int_0^t \int_{-1}^{1} \frac{e^{\sigma/2}}{\sqrt{t-\sigma}} d\theta d\sigma e^{\frac{L}{2} - \frac{L}{6}} \\
&\quad \le O(1) X e^x \sqrt{t} e^{\frac{t}{2}} e^{\frac{L}{2} - \frac{L}{6}}, \quad \text{where } \bar{d} > 1,
\end{aligned}
$$

which imply (3.14). □

The leading ansatz for the solution $\bar{w}(x,t)$ is provided by the estimate for $\bar{w}^0(x,t)$, for which we need the following lemmas. They are analogous to Lemmas 2.4 and 2.5, with additional attention to the property of time decay.

LEMMA 3.2. *For $x \ge (-L-t)/2$,*

$$
\int_0^t \bar{g}(x,t;-L-\sigma,\sigma) e^{\frac{-\sigma}{2}} e^{\frac{-L}{2} - \frac{L}{6}} d\sigma \le O(1) e^{\frac{-|x|}{4}} e^{\frac{-3t}{16}} e^{\frac{-L}{4} - \frac{L}{6}}, \tag{3.16}
$$

$$
\left| \int_0^t \bar{g}_x(x,t;-L-\sigma,\sigma) e^{\frac{-\sigma}{2}} e^{\frac{-L}{2} - \frac{L}{6}} d\sigma \right| \le O(1) e^{\frac{-|x|}{4}} e^{\frac{-3t}{16}} e^{\frac{-L}{4} - \frac{L}{6}}. \tag{3.17}
$$

Proof. First we notice that, for $y < 0$,

$$
\begin{aligned}
\bar{g}_x(x,t;y,\sigma) &= \frac{1}{\sqrt{4\pi(t-\sigma)}} \frac{-(1+e^{y+x_0}) e^{x+x_0}}{(1+e^{x+x_0})^2} e^{\frac{-(x-y-(t-\sigma))^2}{4(t-\sigma)}} \\
&\quad + \frac{1}{\sqrt{4\pi(t-\sigma)}} \frac{1+e^{y+x_0}}{1+e^{x+x_0}} \frac{-2(x-y-(t-\sigma))}{4(t-\sigma)} e^{\frac{-(x-y-(t-\sigma))^2}{4(t-\sigma)}} \\
&\equiv \bar{g}_{1x} + \bar{g}_{2x}.
\end{aligned} \tag{3.18}
$$

Since

$$
\int_0^t \bar{g}(x,t;-L-\sigma,\sigma) e^{\frac{-\sigma}{2}} e^{\frac{-L}{2} - \frac{L}{6}} d\sigma
$$

$$= O(1)\frac{1}{1+e^x}\int_0^t \frac{1}{\sqrt{4\pi(t-\sigma)}} e^{\frac{-(x+L+\sigma-(t-\sigma))^2}{4(t-\sigma)}} \cdot e^{\frac{-\sigma}{2}} e^{\frac{-L}{2}-\frac{L}{6}} d\sigma$$

$$= O(1)\frac{e^{\frac{x}{4}}}{1+e^x}\int_0^t \frac{1}{\sqrt{t-\sigma}} e^{\frac{-[x+L+\sigma-(1/2)(t-\sigma)]^2}{4(t-\sigma)}} e^{\frac{-3}{16}(t-\sigma)} e^{\frac{-\sigma}{4}} d\sigma e^{\frac{-L}{4}-\frac{L}{6}}$$

$$\le O(1)\frac{e^{\frac{x}{4}}}{1+e^x} e^{\frac{-3t}{16}} e^{\frac{-L}{4}-\frac{L}{6}},$$

the first estimate is proved. Set

$$\eta = \frac{x+L+\sigma-\frac{1}{2}(t-\sigma)}{\sqrt{4(t-\sigma)}},$$

$$\Rightarrow d\eta = \left(\frac{3/2}{\sqrt{4(t-\sigma)}} + 2\frac{x+L+\sigma-(1/2)(t-\sigma)}{(4(t-\sigma))^{\frac{3}{2}}}\right) d\sigma.$$

Then

$$\left|\int_0^t \bar{g}_x(x,t;-L-\sigma,\sigma) e^{\frac{-\sigma}{2}} e^{\frac{-L}{2}-\frac{L}{6}} d\sigma\right|$$

$$= \left| O(1)\frac{1}{1+e^x}\int_0^t \left(1+\frac{-2(x+L+\sigma-(t-\sigma))}{4(t-\sigma)}\right)\frac{e^{\frac{-(x+L+\sigma-(t-\sigma))^2}{4(t-\sigma)}}}{\sqrt{4(t-\sigma)}} \right.$$

$$\left. \cdot e^{\frac{-\sigma}{2}} e^{\frac{-L}{2}-\frac{L}{6}} d\sigma \right|$$

$$= \left| O(1)\frac{e^{\frac{x}{4}}}{1+e^x}\int_0^t \left(-\frac{3/2}{\sqrt{4(t-\sigma)}} - 2\frac{x+L+\sigma-(1/2)(t-\sigma)}{(4(t-\sigma))^{\frac{3}{2}}}\right) \right.$$

$$\cdot e^{\frac{-[x+L+\sigma-(1/2)(t-\sigma)]^2}{4(t-\sigma)}} e^{(-3/16)(t-\sigma)} e^{\frac{-\sigma}{4}} d\sigma e^{\frac{-L}{4}-\frac{L}{6}}$$

$$\left. + O(1)\frac{e^{\frac{x}{4}}}{1+e^x}\int_0^t \frac{1}{\sqrt{t-\sigma}} e^{\frac{-[x+L+\sigma-(1/2)(t-\sigma)]^2}{4(t-\sigma)}} e^{(-3/16)(t-\sigma)} e^{\frac{-\sigma}{4}} d\sigma e^{\frac{-L}{4}-\frac{L}{6}} \right|$$

$$\le O(1)\frac{e^{\frac{x}{4}}}{1+e^x} e^{\frac{-3t}{16}} e^{\frac{-L}{4}-\frac{L}{6}}. \qquad \square$$

LEMMA 3.3. *For* $x \ge \frac{(-L-t)}{2}$,

$$\int_{-L}^{\infty} \bar{g}(x,t;y,0)\bar{w}(y,0)dy$$

$$\le O(1)\begin{cases} e^{\frac{x}{4}} e^{\frac{-3t}{16}} e^{\frac{-L}{4}-\frac{L}{6}} + e^{\frac{2x}{3}}(x+t+L+1)^{-\alpha+1} e^{\frac{-L}{2}}, & \frac{-L-t}{2} \le x \le 0, \\ (x+t+L+1)^{-\alpha+1} e^{\frac{-L}{4}-\frac{L}{6d}}, & x > 0. \end{cases}$$

Proof. We have

$$\int_{-L}^{\infty} \bar{g}(x,t;y,0)\bar{w}(y,0)dy$$

$$
\begin{aligned}
&= \int_{-L}^{0} \frac{1}{\sqrt{4\pi t}} \frac{1+e^{y+x_0}}{1+e^{x+x_0}} e^{\frac{-(x-t-y)^2}{4t}} \bar{w}(y,0)dy \\
&\quad + \int_{0}^{\infty} \frac{1}{\sqrt{4\pi t}} \frac{1+e^{-(y+x_0)}}{1+e^{-(x+x_0)}} e^{\frac{-(x+t-y)^2}{4t}} \bar{w}(y,0)dy \\
&\equiv I_1 + I_2.
\end{aligned}
$$

The estimate of I_1 is straightforward by (3.8):

$$
\begin{aligned}
|I_1| &= O(1)\frac{1}{1+e^x} e^{-\frac{L}{2}-\frac{L}{6}} \int_{-L}^{0} \frac{e^{\frac{-(x-t-y)^2}{4t}}}{\sqrt{4\pi t}} dy \\
&\le O(1)\frac{1}{1+e^x} e^{-\frac{L}{2}-\frac{L}{6}} \cdot e^{\frac{x}{4}} e^{\frac{L}{4}} e^{\frac{-3t}{16}} \int_{-L}^{0} \frac{e^{\frac{-(x-t/2-y)^2}{4t}}}{\sqrt{4\pi t}} dy \\
&\le O(1) e^{\frac{-|x|}{4}} e^{\frac{-3t}{16}} e^{\frac{-L}{4}-\frac{L}{6}}.
\end{aligned}
$$

For the estimate of I_2, we have the following cases:

Case 1. For $x \ge 0$,

$$
\begin{aligned}
|I_2| &\le O(1) e^{\frac{-L}{2}-\frac{L}{6d}} \\
&\quad \cdot \left(\int_0^{\frac{(x+t)}{3}} + \int_{\frac{(x+t)}{3}}^{\infty} \right) \frac{1}{\sqrt{4\pi t}} e^{\frac{-(x+t-y)^2}{4t}} (y+L+1)^{-\alpha+1} dy.
\end{aligned}
$$

Since

$$
\begin{aligned}
&\int_0^{\frac{(x+t)}{3}} \frac{1}{\sqrt{4\pi t}} e^{\frac{-(x+t-y)^2}{4t}} (y+L+1)^{-\alpha+1} dy \\
&\qquad \le O(1) \int_{\frac{x+t}{3\sqrt{t}}}^{\frac{x+t}{\sqrt{4t}}} e^{-\eta^2} d\eta \le O(1) e^{\frac{-2x}{9}} e^{\frac{-t}{9}}, \\
&\int_{\frac{(x+t)}{3}}^{\infty} \frac{1}{\sqrt{4\pi t}} e^{\frac{-(x+t-y)^2}{4t}} (y+L+1)^{-\alpha+1} dy \\
&\qquad \le O(1)(x+t+L+1)^{-\alpha+1},
\end{aligned}
$$

we have

$$
|I_2| \le O(1)(x+t+L+1)^{-\alpha+1} e^{\frac{-L}{2}}.
$$

Case 2. For $-t < x < 0$,

$$
\begin{aligned}
|I_2| &\le O(1) e^x e^{\frac{-L}{2}} e^{\frac{-L}{6d}} \cdot \left(\int_0^{\frac{(x+t)}{3}} + \int_{\frac{(x+t)}{3}}^{\infty} \right) \frac{1}{\sqrt{4\pi t}} e^{\frac{-(x+t-y)^2}{4t}} (y+L+1)^{-\alpha+1} dy. \\
&\le O(1) \left(e^{x-\frac{2x}{9}} e^{\frac{-t}{9}} e^{\frac{-L}{2}-\frac{L}{6d}} + e^x (x+t+L+1)^{-\alpha+1} e^{\frac{-L}{2}-\frac{L}{6d}} \right) \\
&\le O(1) e^{\frac{-2|x|}{3}} (x+t+L+1)^{-\alpha+1} e^{\frac{-L}{2}}.
\end{aligned}
$$

Case 3. For $x \le -t$,

$$
|I_2| \le O(1) e^x e^{\frac{-L}{2}-\frac{L}{6d}} \int_0^{\infty} \frac{1}{\sqrt{4\pi t}} e^{\frac{-(x+t-y)^2}{4t}} (y+L+1)^{-\alpha+1} dy
$$

$$\le O(1)e^{x}e^{\frac{-L}{2}-\frac{L}{6d}}\int_{-\infty}^{\frac{x+t}{\sqrt{4t}}}e^{-\eta^2}d\eta$$
$$\le O(1)e^{\frac{x}{2}}e^{\frac{-t}{4}}e^{\frac{-L}{2}-\frac{L}{6d}}.$$

Combining the estimates for I_1 and I_2, the lemma is proved. □

LEMMA 3.4. *For* $x \ge \frac{(-L-t)}{2}$,

$$\int_0^t\int_{-L-\sigma}^{\infty}\bar{g}(x,t;y,\sigma)e^{\frac{-L}{6}}e^{\frac{-(L+\sigma)}{2}}e^{-(y+L+\sigma)}dyd\sigma \le O(1)e^{\frac{-|x|}{4}}e^{-\frac{3t}{16}}e^{\frac{-L}{4}-\frac{L}{6}}.$$

Proof.

$$\begin{aligned}&\int_0^t\int_{-L-\sigma}^{\infty}\bar{g}(x,t;y,\sigma)e^{\frac{-L}{6}}e^{\frac{-(L+\sigma)}{2}}e^{-(y+L+\sigma)}dyd\sigma\\ &\le O(1)\left(\frac{e^{\frac{x}{4}}}{1+e^{x}}\int_0^t\int_{-L-\sigma}^{0}\frac{1}{\sqrt{4\pi(t-\sigma)}}e^{\frac{-\left(x-y-\frac{(t-\sigma)}{2}\right)^2}{4(t-\sigma)}}e^{\frac{-y}{4}}e^{\frac{-3}{16}(t-\sigma)}\right.\\ &\qquad\cdot e^{\frac{-L}{6}}e^{\frac{-(L+\sigma)}{2}}e^{-(y+L+\sigma)}dyd\sigma\\ &\quad+\frac{e^{\frac{-x}{2}}}{1+e^{-x}}\int_0^t\int_0^{\infty}\frac{1}{\sqrt{4\pi(t-\sigma)}}e^{\frac{-(x-y)^2}{4(t-\sigma)}}e^{\frac{y}{2}}e^{\frac{-(t-\sigma)}{4}}\\ &\qquad\left.\cdot e^{\frac{-L}{6}}e^{\frac{-(L+\sigma)}{2}}e^{-(y+L+\sigma)}dyd\sigma\right)\\ &\le O(1)e^{\frac{-|x|}{4}}e^{-\frac{3t}{16}}e^{\frac{-L}{4}-\frac{L}{6}}.\end{aligned}$$ □

LEMMA 3.5. *For* $-L-t \le x \le \frac{-L-t}{2}$ *and* $1<D<2$,

$$\int_0^t\int_{-L-\sigma}^{\infty}\frac{1}{(t-\sigma)}e^{\frac{-(Y-X)^2}{4D(t-\sigma)}}e^{x-y}\cdot e^{\frac{-L}{6}}e^{\frac{-(L+\sigma)}{2}}e^{-(y+L+\sigma)}dyd\sigma \le O(1)e^{\frac{x}{2}}e^{\frac{-L}{6}}.$$

Proof. The calculation is divided into two parts: For $-L-\sigma \le y < 0$,

$$\begin{aligned}&\int_0^t\int_{-L-\sigma}^{0}\frac{1}{(t-\sigma)}e^{\frac{-(Y-X)^2}{4D(t-\sigma)}}e^{x-y}\cdot e^{\frac{-L}{6}}e^{\frac{-(L+\sigma)}{2}}e^{-(y+L+\sigma)}dyd\sigma\\ &\le\int_0^t\int_{-L-\sigma}^{0}\frac{1}{(t-\sigma)}e^{\frac{-(x-y+(t-\sigma))^2}{4D(t-\sigma)}}e^{x-y}\cdot e^{\frac{-L}{6}}e^{\frac{y}{2}}dyd\sigma\\ &=\int_0^t\int_{-L-\sigma}^{0}\frac{1}{(t-\sigma)}e^{\frac{-[x-y+(1-D)(t-\sigma)]^2}{4D(t-\sigma)}}e^{\frac{x}{2}}e^{\frac{-(2-D)}{4}(t-\sigma)}dyd\sigma e^{\frac{-L}{6}}\\ &\le O(1)e^{\frac{x}{2}}e^{\frac{-L}{6}},\end{aligned}$$

and for $y \ge 0$,

$$\int_0^t\int_0^{\infty}\frac{1}{(t-\sigma)}e^{\frac{-(Y-X)^2}{4D(t-\sigma)}}e^{x-y}\cdot e^{\frac{-L}{6}}e^{\frac{-(L+\sigma)}{2}}e^{-(y+L+\sigma)}dyd\sigma$$

$$= \int_0^t \int_0^\infty \frac{1}{(t-\sigma)} e^{\frac{-(x-y)^2}{4D(t-\sigma)} - \frac{x}{2D} + \frac{y}{2D} - \frac{(t-\sigma)}{4D}} e^{x-y} \cdot e^{\frac{-L}{6}} e^{\frac{-(L+\sigma)}{2}} e^{-(y+L+\sigma)} dyd\sigma$$
$$\le O(1) e^{(1-\frac{1}{2D})x} e^{\frac{-t}{4D}} e^{-L-\frac{L}{2}-\frac{L}{6}}. \qquad \square$$

Now, we are ready to study the solution through iterations. We first estimate $\bar{w}^0(x,t)$. From the representation of $\bar{w}^0(x,t)$,

$$|\bar{w}^0(x,t)| \le O(1) \left(\int_{-L}^\infty k(x-y-t,t)|\bar{w}(y,0)|dy \right.$$
$$\left. + \int_0^t \int_{-L-\sigma}^\infty k(x-y-(t-\sigma);t-\sigma) e^{\frac{-(L+\sigma)}{2}} e^{\frac{-L}{6}} e^{-(y+L+\sigma)} dyd\sigma \right)$$
$$\le O(1) e^{\frac{x}{2}} e^{\frac{-L}{6}} \quad \text{for } -L-t \le x < \frac{-L-t}{2}.$$

Based on the relation

$$K_x^B(x,t;y,\sigma) + K_y^B(x,t;y,\sigma)$$
$$= 2K^B(x,t;y,\sigma) - 2\partial_y k(Y-X,t-\sigma) e^{(X-Y)-(t-\sigma)} \tag{3.19}$$

and the facts that $K^B(x,t;y=\infty,\sigma) = 0$ and $\bar{w}(-L,0) = 0$, by integration by parts, we have that

$$\left| \int_{-L}^\infty K_x^B(x,t;y,0)\bar{w}(y,0)dy \right|$$
$$\le O(1) \int_{-L}^\infty k(x-y-t,t)|(-\bar{w}(y,0) + \bar{w}_y(y,0))|dy$$
$$\le O(1) e^{\frac{x}{2}} e^{\frac{-t}{4}} e^{\frac{-L}{6}}.$$

From

$$K_x^B(x,t;y,\sigma) = K^B(x,t;y,\sigma)$$
$$+ \left(\frac{(Y-X)}{2(t-\sigma)} \frac{e^{\frac{-(Y-X)^2}{4(t-\sigma)}}}{\sqrt{4\pi(t-\sigma)}} + \frac{(Y+X)}{2(t-\sigma)} \frac{e^{\frac{-(Y+X)^2}{4(t-\sigma)}}}{\sqrt{4\pi(t-\sigma)}} \right) e^{x-y} \tag{3.20}$$
$$\equiv K^B(x,t;y,\sigma) + K^\star(x,t;y,\sigma)$$

and

$$|K^\star(x,t;y,\sigma)| \le O(1) \frac{1}{(t-\sigma)} e^{\frac{-(Y-X)^2}{4D(t-\sigma)}} e^{x-y}, \tag{3.21}$$

where $D > 1$ and is close to 1, we have, by Lemma 3.5,

$$\left| \int_0^t \int_{-L-\sigma}^\infty K_x^B \left[R_\sigma + \phi(y+x_0)R_y - R_{yy} + \frac{1}{2}R_y^2 \right] dyd\sigma \right|$$
$$\le O(1) \left(\int_0^t \int_{-L-\sigma}^\infty k(x-y-(t-\sigma),t-\sigma) e^{\frac{-L}{6}} e^{\frac{-(L+\sigma)}{2}} e^{-(y+L+\sigma)} dyd\sigma \right.$$
$$\left. + \int_0^t \int_{-L-\sigma}^\infty \frac{1}{t-\sigma} e^{\frac{-(Y-X)^2}{4D(t-\sigma)}} e^{x-y} e^{\frac{-L}{6}} e^{\frac{-(L+\sigma)}{2}} e^{-(y+L+\sigma)} dyd\sigma \right)$$
$$\le O(1) e^{\frac{x}{2}} e^{\frac{-L}{6}}.$$

Therefore,

$$|\bar{w}_x^0(x,t)| \le O(1)e^{\frac{x}{2}}e^{\frac{-L}{6}} \quad \text{for } -L-t \le x < \frac{-L-t}{2},$$

and, in particular,

$$|\bar{w}_x^0(-L-t,t)| \le O(1)e^{\frac{-t}{2}}e^{\frac{-L}{2}-\frac{L}{6}}.$$

Then by Lemmas 3.2, 3.3, and 3.4,

$$|\bar{w}^0(x,t)| \le O(1)\begin{cases} e^{\frac{x}{4}}e^{\frac{-3t}{16}}e^{\frac{-L}{4}-\frac{L}{6}} + e^{\frac{2x}{3}}(x+t+L+1)^{-\alpha+1}e^{\frac{-L}{2}} & \text{for } \dfrac{-L-t}{2} \le x \le 0, \\ (x+t+L+1)^{-\alpha+1}e^{\frac{-L}{4}-\frac{L}{6d}} & \text{for } x>0. \end{cases}$$

Also note that

$$\begin{aligned} \bar{g}_x + \bar{g}_y = {} & \frac{-1}{4}\frac{1}{\cosh^2((x+x_0)/2)}\frac{1}{\sqrt{4\pi(t-\sigma)}}e^{\frac{-(x-y-(t-\sigma))^2}{4(t-\sigma)}} \\ & + \frac{1}{4}\frac{1}{\cosh^2((x+x_0)/2)}\frac{1}{\sqrt{4\pi(t-\sigma)}}e^{\frac{-(x-y+(t-\sigma))^2}{4(t-\sigma)}}, \end{aligned} \tag{3.22}$$

$\bar{g}(x,t,\infty,\sigma) = 0$ and $\bar{w}(-L-t,t) = 0$. Then by Lemmas 3.2 and 3.4, (3.13), and (3.22),

$$\begin{aligned} & |\bar{w}_x^0(x,t)| \\ & \le O(1)\Bigg(\int_{-L}^{\infty} \bar{g}(x,t;y,0)|\bar{w}_y(y,0)|dy \\ & \quad + e^{-|x|}\int_{-L}^{\infty}\frac{1}{\sqrt{4\pi t}}\left(e^{\frac{-(x-y-t)^2}{4t}} + e^{\frac{-(x-y+t)^2}{4t}}\right)|\bar{w}(y,0)|dy \\ & \quad + \left|\int_0^t \bar{g}_x(x,t;-L-\sigma,\sigma)\bar{w}_y^0(-L-\sigma,\sigma)d\sigma\right| \\ & \quad + \int_0^t \bar{g}(x,t;-L-\sigma,\sigma)e^{\frac{-(L+\sigma)}{2}}e^{\frac{-L}{6}}d\sigma \\ & \quad + \int_0^t\int_{-L-\sigma}^{\infty}\bar{g}(x,t;y,\sigma)e^{\frac{-L}{6}}e^{\frac{-(L+\sigma)}{2}}e^{-(y+L+\sigma)}dyd\sigma \\ & \quad + \int_0^t\int_{-L-\sigma}^{\infty}\frac{e^{-|x|}}{\sqrt{t-\sigma}}\left(e^{\frac{-(x-y-(t-\sigma))^2}{4(t-\sigma)}} + e^{\frac{-(x-y+(t-\sigma))^2}{4(t-\sigma)}}\right) \\ & \qquad\qquad \cdot e^{\frac{-L}{6}}e^{\frac{-(L+\sigma)}{2}}e^{-(y+L+\sigma)}dyd\sigma\Bigg) \\ & \le O(1)\begin{cases} e^{\frac{-L}{4}-\frac{L}{6d}}\left[e^{\frac{-|x|}{4}}e^{\frac{-3t}{16}} + e^{\frac{-2|x|}{3}}(x+t+L+1)^{-\alpha+1}\right] & \text{for } \dfrac{-(L+t)}{2} < x \le 0, \\ e^{\frac{-L}{4}-\frac{L}{6d}}[(x+t+L+1)^{-\alpha} + e^{-|x|}(x+t+L+1)^{-\alpha+1}] & \text{for } x>0. \end{cases} \end{aligned}$$

With these, we can now formulate our main theorem.

THEOREM 3.6. *Suppose that L is sufficiently large and $\alpha > 1$. Then the solution $\bar{w}(x,t)$ for (3.6)–(3.9) satisfies*

(3.23)
$$|\bar{w}(x,t)| \le C_0 \cdot \begin{cases} e^{\frac{-|x|}{4}} e^{\frac{-t}{8}} e^{\frac{-L}{4}} + e^{\frac{-|x|}{2}} (x+t+L+1)^{-\alpha+1} e^{\frac{-L}{4}}, & -L-t \le x \le 0, \\ (x+t+L+1)^{-\alpha+1} e^{\frac{-L}{4}}, & x > 0, \end{cases}$$

(3.24)
$$|\bar{w}_x(x,t)| \le C_0 \cdot \begin{cases} e^{\frac{-|x|}{4}} e^{\frac{-t}{8}} e^{\frac{-L}{4}} + e^{\frac{-|x|}{2}} (x+t+L+1)^{-\alpha+1} e^{\frac{-L}{4}}, & -L-t \le x \le 0, \\ (x+t+L+1)^{-\alpha} e^{\frac{-L}{4}} + e^{\frac{-|x|}{3}} (x+t+L+1)^{-\alpha+1} e^{\frac{-L}{4}}, & x > 0, \end{cases}$$

where $C_0 > 0$ is a constant.

Remark 3.7. The theorem induces that the decay rate of $v(x,t) = u(x,t) - \phi(x + x_0)$ is also of the form (3.24).

Before proving the theorem, we need the following lemmas to study the nonlinear terms in the iterations.

LEMMA 3.8. *For $x \ge (-L-t)/2$,*

$$\int_0^t \int_{-L-\sigma}^0 \bar{g} e^{-|y|} (y+\sigma+L+1)^{-2\alpha+2} dy d\sigma \le O(1) \begin{cases} e^{\frac{-|x|}{2}} e^{\frac{-t}{8}} t (L+1)^{-2\alpha+2} + e^{\frac{-|x|}{2}} (x+t+L+1)^{-\alpha+1} & \text{for } \dfrac{-L-t}{2} \le x \le 0, \\ e^{\frac{-|x|}{3}} (x+t+L+1)^{-\alpha+1} & \text{for } x > 0, \end{cases}$$

$$\left| \int_0^t \int_{-L-\sigma}^0 \bar{g}_x e^{-|y|} (y+\sigma+L+1)^{-2\alpha+2} dy d\sigma \right| \le O(1) \begin{cases} e^{\frac{-|x|}{2}} e^{\frac{-t}{8}} (t+\sqrt{t}) (L+1)^{-2\alpha+2} + e^{\frac{-|x|}{2}} (x+t+L+1)^{-\alpha+1} \\ \qquad \text{for } \dfrac{-L-t}{2} \le x \le 0, \\ e^{\frac{-|x|}{3}} (x+t+L+1)^{-\alpha+1} \quad \text{for } x > 0. \end{cases}$$

Proof. Recall that $\bar{g}$ is a simple translation of g. From (2.9)

$$\begin{aligned} &\int_0^t \int_{-L-\sigma}^0 \bar{g} e^{-|y|} (y+\sigma+L+1)^{-2\alpha+2} dy d\sigma \\ &\le O(1) e^{\frac{-|x|}{2}} \int_0^t \int_{-L-\sigma}^0 \frac{e^{\frac{-(x-y)^2}{4(t-\sigma)}}}{\sqrt{4\pi(t-\sigma)}} e^{\frac{-|y|}{2}} e^{\frac{-(t-\sigma)}{4}} (y+\sigma+L+1)^{-2\alpha+2} dy d\sigma \\ &= O(1) e^{\frac{-|x|}{2}} \int_0^t \left(\int_{-L-\sigma}^{\frac{-L-\sigma}{2}} + \int_{\frac{-L-\sigma}{2}}^0 \right) \frac{e^{\frac{-(x-y)^2}{4(t-\sigma)}}}{\sqrt{4\pi(t-\sigma)}} e^{\frac{-|y|}{2}} e^{\frac{-(t-\sigma)}{4}} (y+\sigma+L+1)^{-2\alpha+2} dy d\sigma \\ &\le O(1) e^{\frac{-|x|}{2}} \int_0^t e^{\frac{-(t-\sigma)}{4}} e^{\frac{-L-\sigma}{4}} d\sigma \\ &\quad + O(1) e^{\frac{-|x|}{2}} \int_0^t e^{\frac{-(t-\sigma)}{4}} \left(\frac{L+\sigma}{2} + 1 \right)^{-2\alpha+2} d\sigma. \end{aligned}$$

This proves the first estimate. From (3.18)

$$\left|\int_0^t\int_{-L-\sigma}^0 \bar{g}_x e^{-|y|}(y+\sigma+L+1)^{-2\alpha+2}dyd\sigma\right|$$
$$\le O(1)e^{\frac{-|x|}{2}}\int_0^t\int_{-L-\sigma}^0\left(1+\frac{1}{\sqrt{t-\sigma}}\right)\frac{e^{\frac{-(x-y)^2}{4r(t-\sigma)}}}{\sqrt{t-\sigma}}e^{\frac{-|y|}{2}}e^{\frac{-(t-\sigma)}{4}}(y+\sigma+L+1)^{-2\alpha+2}dyd\sigma,$$

where $r>1$ and is close to 1. From this the second estimate follows easily. □

LEMMA 3.9. *For $x\ge(-L-t)/2$,*

$$\int_0^t\int_0^\infty \bar{g}e^{\frac{-|y|}{3}}(y+\sigma+L+1)^{-2\alpha+1}dyd\sigma$$
$$\le O(1)\begin{cases} e^{\frac{-2|x|}{3}}e^{\frac{-t}{9}}t(L+1)^{-2\alpha+1}+e^{\frac{-2|x|}{3}}(x+t+L+1)^{-\alpha} & \text{for } \dfrac{-L-t}{2}\le x\le 0,\\ e^{\frac{-|x|}{3}}(x+t+L+1)^{-\alpha} & \text{for } x>0,\end{cases}$$

$$\int_0^t\int_0^\infty \bar{g}_x e^{\frac{-|y|}{3}}(y+\sigma+L+1)^{-2\alpha+1}dyd\sigma$$
$$\le O(1)\begin{cases} e^{\frac{-2|x|}{3}}e^{\frac{-t}{9}}(t+\sqrt{t})(L+1)^{-2\alpha+1}+e^{\frac{-2|x|}{3}}(x+t+L+1)^{-\alpha}\\ \qquad\qquad \text{for } \dfrac{-L-t}{2}\le x\le 0,\\ e^{\frac{-|x|}{3}}(x+t+L+1)^{-\alpha} \quad \text{for } x>0.\end{cases}$$

Proof. Since for $x\le 0$,

$$\int_0^t\int_0^\infty \bar{g}e^{\frac{-|y|}{3}}(y+\sigma+L+1)^{-2\alpha+1}dyd\sigma$$
$$\le O(1)e^{\frac{-2|x|}{3}}\int_0^t\int_0^\infty\frac{1}{\sqrt{4\pi(t-\sigma)}}e^{\frac{-[x-y+\frac{1}{3}(t-\sigma)]^2}{4(t-\sigma)}}e^{\frac{-2(t-\sigma)}{9}}(y+\sigma+L+1)^{-2\alpha+1}dyd\sigma$$
$$\le O(1)e^{\frac{-2|x|}{3}}\left(e^{\frac{-t}{9}}t(L+1)^{-2\alpha+1}+(x+t+L+1)^{-\alpha}\right),$$

and for $x>0$,

$$\int_0^t\int_0^\infty \bar{g}e^{\frac{-|y|}{3}}(y+\sigma+L+1)^{-2\alpha+1}dyd\sigma$$
$$\le O(1)e^{\frac{-|x|}{3}}\int_0^t\int_0^\infty\frac{1}{\sqrt{4\pi(t-\sigma)}}e^{\frac{-[x-y+\frac{1}{3}(t-\sigma)]^2}{4(t-\sigma)}}e^{\frac{-2(t-\sigma)}{9}}(y+\sigma+L+1)^{-2\alpha+1}dyd\sigma$$
$$= O(1)e^{\frac{-|x|}{3}}\int_0^t\left(\int_0^{\frac{x+(t-\sigma)/3}{2}}+\int_{\frac{x+(t-\sigma)/3}{2}}^\infty\right)\frac{e^{\frac{-[x-y+\frac{1}{3}(t-\sigma)]^2}{4(t-\sigma)}}}{\sqrt{4\pi(t-\sigma)}}e^{\frac{-2(t-\sigma)}{9}}$$
$$\cdot(y+\sigma+L+1)^{-2\alpha+1}dyd\sigma$$
$$\le O(1)e^{\frac{-|x|}{3}}\left(\int_0^t e^{\frac{-[x+\frac{1}{3}(t-\sigma)]^2}{16(t-\sigma)}}e^{\frac{-2(t-\sigma)}{9}}(\sigma+L+1)^{-2\alpha+1}d\sigma\right.$$
$$\left.+\int_0^t e^{\frac{-2(t-\sigma)}{9}}(x+t+\sigma+L+1)^{-2\alpha+1}d\sigma\right)$$
$$\le O(1)e^{\frac{-|x|}{3}}(x+t+L+1)^{-\alpha},$$

the first estimate is proved. Due to (3.18),

$$\left|\int_0^t\int_0^\infty \bar g_x e^{\frac{-|y|}{3}}(y+\sigma+L+1)^{-2\alpha+1}dyd\sigma\right|$$
$$\le O(1)e^{\frac{-|x|}{3}}\int_0^t\int_0^\infty\left(1+\frac{1}{\sqrt{t-\sigma}}\right)\frac{e^{\frac{-[x-y+\frac{1}{3}(t-\sigma)]^2}{4r(t-\sigma)}}}{\sqrt{t-\sigma}}e^{\frac{-2(t-\sigma)}{9}}(y+\sigma+L+1)^{-2\alpha+1}dyd\sigma,$$

where $r>1$ and is close to 1. The second estimate then follows easily as the first. ∎

LEMMA 3.10. *For $x\ge(-L-t)/2$,*

$$\int_0^t\int_0^\infty \bar g(y+\sigma+L+1)^{-2\alpha}dyd\sigma\, e^{\frac{-L}{6}}$$
$$\le O(1)\begin{cases} e^{\frac{-|x|}{2}}e^{\frac{-t}{2}}e^{\frac{-L}{6}}+e^{\frac{-|x|}{2}}(x+t+L+1)^{-\alpha}e^{\frac{-L}{6}} & \text{for } \dfrac{-L-t}{2}\le x\le 0,\\ (x+t+L+1)^{-\alpha}e^{\frac{-L}{6}} & \text{for } x>0.\end{cases}$$
$$\int_0^t\int_0^\infty \bar g_x(y+\sigma+L+1)^{-2\alpha}dyd\sigma\, e^{\frac{-L}{6}}$$
$$\le O(1)\begin{cases} e^{\frac{-|x|}{2}}e^{\frac{-t}{2}}\sqrt{t}\,e^{\frac{-L}{6}}+e^{\frac{-|x|}{2}}(x+t+L+1)^{-\alpha}e^{\frac{-L}{6}} & \text{for } \dfrac{-L-t}{2}\le x\le 0,\\ (x+t+L+1)^{-\alpha}e^{\frac{-L}{6}} & \text{for } x>0.\end{cases}$$

Proof. For convenience, let

$$P_3\equiv\int_0^t\int_0^\infty \bar g(y+\sigma+L+1)^{-2\alpha}dyd\sigma\, e^{\frac{-L}{6}}.$$

When $-t<x\le 0$

$$P_3\le O(1)e^x\cdot\left(\int_0^{\frac{x+t}{r_1}}+\int_{\frac{x+t}{r_1}}^t\right)\int_0^\infty\frac{1}{\sqrt{4\pi(t-\sigma)}}e^{\frac{-[x+(t-\sigma)-y]^2}{4(t-\sigma)}}(y+\sigma+L+1)^{-2\alpha}dyd\sigma\, e^{\frac{-L}{6}}$$
$$\equiv O(1)(P_{31}^-+P_{32}^-),$$

where $r_1>1$, with

$$P_{31}^-=O(1)e^x\int_0^{\frac{x+t}{r_1}}\left(\int_0^{\frac{1}{r_2}(1-\frac{1}{r_1})(x+t)}+\int_{\frac{1}{r_2}(1-\frac{1}{r_1})(x+t)}^\infty\right)\frac{e^{\frac{-[x+(t-\sigma)-y]^2}{4(t-\sigma)}}}{\sqrt{4\pi(t-\sigma)}}$$
$$\cdot(y+\sigma+L+1)^{-2\alpha}dyd\sigma e^{\frac{-L}{6}}$$
$$\le O(1)e^{\frac{-|x|}{2}}(x+t+L+1)^{-\alpha}e^{\frac{-L}{6}},$$

where $r_2>1$, and

$$P_{32}^-=O(1)e^x\int_{\frac{x+t}{r_1}}^t(\sigma+L+1)^{-2\alpha}d\sigma\cdot e^{\frac{-L}{6}}$$
$$\le O(1)e^x(x+t+L+1)^{-\alpha}e^{\frac{-L}{6}}.$$

When $x \le -t$, it is obvious that

$$P_3 \le O(1)e^x e^{\frac{-L}{6}} \le O(1)e^{\frac{x}{2}} e^{\frac{-t}{2}} e^{\frac{-L}{6}}.$$

When $x > 0$, we have

$$\begin{aligned} P_3 &\le O(1)\int_0^t \left(\int_0^{\frac{x+(t-\sigma)}{2}} + \int_{\frac{x+(t-\sigma)}{2}}^{\infty} \right) \frac{e^{\frac{-[x+(t-\sigma)-y]^2}{4(t-\sigma)}}}{\sqrt{4\pi(t-\sigma)}} (y+\sigma+L+1)^{-2\alpha} dy d\sigma e^{\frac{-L}{6}} \\ &\le O(1)(x+t+L+1)^{-\alpha} e^{\frac{-L}{6}}. \end{aligned}$$

The first estimate is proved. As in the previous two lemmas, the second estimate follows due to (3.18). □

Proof of Theorem 3.6. We will prove the theorem by induction. Suppose that $\bar{w}^m(x,t)$ and $\bar{w}^m_x(x,t)$ satisfy the estimates (3.23) and (3.24) for all $m \le n$. We will show that the sequence

$$\bar{\delta}^m(x,t) \equiv \bar{w}^m(x,t) - \bar{w}^{m-1}(x,t), \quad m \ge 1,$$

is geometric in the weighted norm

$$\begin{aligned} |||h||| \equiv & \sup_{\substack{-L-t\le x\le 0 \\ t>0}} \frac{|h(x,t)|}{e^{\frac{-|x|}{4}} e^{\frac{-t}{8}} + e^{\frac{-|x|}{2}} (x+t+L+1)^{-\alpha+1}} \\ & + \sup_{\substack{x>0 \\ t>0}} \frac{|h(x,t)|}{(x+t+L+1)^{-\alpha+1}} \\ & + \sup_{\substack{-L-t\le x\le 0 \\ t>0}} \frac{|h_x(x,t)|}{e^{\frac{-|x|}{4}} e^{\frac{-t}{8}} + e^{\frac{-|x|}{2}} (x+t+L+1)^{-\alpha+1}} \\ & + \sup_{\substack{x>0 \\ t>0}} \frac{|h_x(x,t)|}{(x+t+L+1)^{-\alpha} + e^{\frac{-|x|}{3}} (x+t+L+1)^{-\alpha+1}}. \end{aligned}$$

From the estimates on $\bar{w}^0$ preceding (3.22), we have, by (3.12), for $(-L-t) \le x < (-L-t)/2$,

$$\begin{aligned} \bar{\delta}^1(x,t) = & \int_0^t \int_{-L-\sigma}^{\infty} K^B (1-\phi(y+x_0)) \bar{w}^0_y dy d\sigma \\ & - \int_0^t \int_{-L-\sigma}^{\infty} K^B \frac{(\bar{w}^0_y)^2}{2} dy d\sigma \\ & - \int_0^t \int_{-L-\sigma}^{\infty} K^B \bar{w}^0_y R_y dy d\sigma; \end{aligned}$$

for $x \ge (-L-t)/2$,

$$\begin{aligned} \bar{\delta}^1(x,t) = & - \int_0^t \bar{g}(x,t;-L-\sigma,\sigma) \bar{\delta}^1_y(-L-\sigma,\sigma) d\sigma \\ & - \int_0^t \int_{-L-\sigma}^{\infty} \bar{g} \frac{(\bar{w}^0_y)^2}{2} dy d\sigma \\ & - \int_0^t \int_{-L-\sigma}^{\infty} \bar{g} \bar{w}^0_y R_y dy d\sigma; \end{aligned}$$

for $(-L-t) \le x < (-L-t)/2$, $m \ge 1$,

$$\begin{aligned}\bar{\delta}^{m+1}(x,t) = &\int_0^t \int_{-L-\sigma}^{\infty} K^B(1-\phi(y+x_0))\bar{\delta}_y^m dyd\sigma \\ &- \int_0^t \int_{-L-\sigma}^{\infty} K^B \bar{\delta}_y^m \frac{(\bar{w}_y^m + \bar{w}_y^{m-1})}{2} dyd\sigma \\ &- \int_0^t \int_{-L-\sigma}^{\infty} K^B \bar{\delta}_y^m R_y dyd\sigma;\end{aligned}$$

for $x \ge (-L-t)/2$, $m \ge 1$,

$$\begin{aligned}\bar{\delta}^{m+1}(x,t) = &- \int_0^t \bar{g}(x,t;-L-\sigma,\sigma)\bar{\delta}_y^{m+1}(-L-\sigma,\sigma)d\sigma \\ &- \int_0^t \int_{-L-\sigma}^{\infty} \bar{g}\bar{\delta}_y^m \frac{(\bar{w}_y^m + \bar{w}_y^{m-1})}{2} dyd\sigma \\ &- \int_0^t \int_{-L-\sigma}^{\infty} \bar{g}\bar{\delta}_y^m R_y dyd\sigma.\end{aligned}$$

Region I: $\{-L-t \le x < (-L-t)/2\}$. Since

$$\left| \int_0^t \int_{-L-\sigma}^{\infty} K^B(1-\phi(y+x_0))\bar{\delta}_y^n dyd\sigma \right| \le O(1) e^{\frac{x}{2}} |||\bar{\delta}^n|||,$$

$$\left| \int_0^t \int_{-L-\sigma}^{\infty} K^B \bar{\delta}_y^n R_y dyd\sigma \right| \le O(1) e^{\frac{x}{2}} e^{\frac{-t}{8}} e^{\frac{-L}{6}} |||\bar{\delta}^n|||$$

and

$$\begin{aligned}&\left| \int_0^t \int_{-L-\sigma}^{\infty} K^B \bar{\delta}_y^n \frac{(\bar{w}_y^n + \bar{w}_y^{n-1})}{2} dyd\sigma \right| \\ &\le O(1) \left(\int_0^t \int_{-L-\sigma}^{0} K^B \left[e^{\frac{-|y|}{4}} e^{\frac{-\sigma}{8}} + e^{\frac{-|y|}{2}} (y+\sigma+L+1)^{-\alpha+1} \right]^2 e^{\frac{-L}{4}} dyd\sigma \right. \\ &\qquad \left. + \int_0^t \int_0^{\infty} K^B \left[(y+\sigma+L+1)^{-\alpha} + e^{\frac{-|y|}{3}} (y+\sigma+L+1)^{-\alpha+1} \right]^2 e^{\frac{-L}{4}} dyd\sigma \right) \\ &\cdot |||\bar{\delta}^n||| \le O(1) e^{\frac{x}{2}} e^{\frac{-L}{4}} \cdot |||\bar{\delta}^n|||,\end{aligned}$$

we have, for $-L-t \le x < (-L-t)/2$,

$$|\bar{\delta}^{n+1}(x,t)| \le O(1) e^{\frac{x}{2}} |||\bar{\delta}^n|||. \tag{3.25}$$

Moreover, similar to Lemma 3.1, we have, from (3.15) and Fubini's theorem, that

$$\bar{\delta}^m(-L-t,t) = 0 \tag{3.26}$$

for all $1 \le m \le (n+1)$. This and (3.12) mean that the boundary value and initial value are kept in the iterations.

Based on (3.20), (3.21), and straightforward calculation, we can obtain

$$\left| \int_0^t \int_{-L-\sigma}^{0} K^\star \bar{\delta}_y^n \frac{(\bar{w}_y^n + \bar{w}_y^{n-1})}{2} dyd\sigma \right|$$

$$
\begin{aligned}
&\le O(1)\left(\int_0^t\int_{-L-\sigma}^0 |K^\star| e^{\frac{-|y|}{2}} e^{\frac{-\sigma}{4}} e^{\frac{-L}{4}} dyd\sigma\right.\\
&\quad+\int_0^t\int_{-L-\sigma}^0 |K^\star|\left[e^{-|y|}(y+\sigma+L+1)^{-2\alpha+2}\right.\\
&\qquad\qquad\left.+e^{\frac{-|y|}{2}-\frac{|y|}{4}} e^{\frac{-\sigma}{8}}(y+\sigma+L+1)^{-\alpha+1}\right]\\
&\qquad\cdot e^{\frac{-L}{4}} dyd\sigma\\
&\quad+\int_0^t\int_0^\infty |K^\star|\left[(y+\sigma+L+1)^{-\alpha}\right.\\
&\qquad\qquad\left.\left.+e^{\frac{-|y|}{3}}(y+\sigma+L+1)^{-\alpha+1}\right]^2 e^{\frac{-L}{4}} dyd\sigma\right)\cdot |||\bar\delta^n|||\\
&\le O(1)e^{\frac{x}{2}} e^{\frac{-L}{4}}\cdot |||\bar\delta^n|||,
\end{aligned}
$$

$$
\left|\int_0^t\int_{-L-\sigma}^\infty K^\star(1-\phi(y+x_0))\bar\delta_y^n dyd\sigma\right| \le O(1)e^{(1-\frac{1}{2D})x}|||\bar\delta^n|||,
$$

and

$$
\left|\int_0^t\int_{-L-\sigma}^\infty K^\star\bar\delta_y^n R_y dyd\sigma\right| \le O(1)e^{(1-\frac{1}{2D})x} e^{\frac{-L}{6}}\cdot |||\bar\delta^n|||.
$$

Hence, for $-L-t\le x<(-L-t)/2$

$$
|\bar\delta_x^{n+1}(x,t)| \le O(1)e^{\frac{x}{2}} e^{\frac{-L}{6}} |||\bar\delta^n|||, \tag{3.27}
$$

and, in particular,

$$
|\bar\delta_x^{n+1}(-L-t,t)| \le O(1)e^{\frac{-t}{2}} e^{\frac{-L}{2}-\frac{L}{6}}\cdot |||\bar\delta^n|||. \tag{3.28}
$$

Region II: $\{x\ge(-L-t)/2\}$. Substituting (3.28) into the following integrals and then by Lemma 3.2, we can obtain

$$
\left|\int_0^t \bar g(x,t;-L-\sigma,\sigma)\bar\delta_y^{n+1}(-L-\sigma,\sigma)d\sigma\right| \le O(1)e^{\frac{-|x|}{4}} e^{\frac{-3t}{16}} e^{\frac{-L}{4}-\frac{L}{6}} |||\bar\delta^n|||,
$$

and

$$
\left|\int_0^t \bar g_x(x,t;-L-\sigma,\sigma)\bar\delta_y^{n+1}(-L-\sigma,\sigma)d\sigma\right| \le O(1)e^{\frac{-|x|}{4}} e^{\frac{-3t}{16}} e^{\frac{-L}{4}-\frac{L}{6}} |||\bar\delta^n|||.
$$

From Lemmas 3.8–3.10, we have, by straightforward calculations,

$$
\begin{aligned}
&\left|\int_0^t\int_{-L-\sigma}^\infty \bar g\bar\delta_y^n \frac{(\bar w_y^n+\bar w_y^{n-1})}{2} dyd\sigma\right|\\
&\le O(1)\left(\int_0^t\int_{-L-\sigma}^0 \bar g\left[e^{\frac{-|y|}{4}} e^{\frac{-\sigma}{8}}+e^{\frac{-|y|}{2}}(y+\sigma+L+1)^{-\alpha+1}\right]^2 e^{\frac{-L}{4}} dyd\sigma\right.\\
&\qquad\left.+\int_0^t\int_0^\infty \bar g\left[e^{\frac{-|y|}{3}}(y+\sigma+L+1)^{-\alpha+1}+(y+\sigma+L+1)^{-\alpha}\right]^2 e^{\frac{-L}{4}} dyd\sigma\right)\\
&\quad\cdot|||\bar\delta^n|||\\
&\le O(1)|||\bar\delta^n|||e^{\frac{-L}{6}}\begin{cases} e^{\frac{-|x|}{2}}e^{\frac{-t}{8}}+e^{\frac{-|x|}{2}}(x+t+L+1)^{-\alpha+1}, & \dfrac{-L-t}{2}<x\le 0,\\ e^{\frac{-|x|}{3}}(x+t+L+1)^{-\alpha+1}+(x+t+L+1)^{-\alpha}, & x>0,\end{cases}
\end{aligned}
$$

and $|\int_0^t \int_{-L-\sigma}^{\infty} \bar{g}_x \bar{\delta}_y^n \frac{(\bar{w}_y^n + \bar{w}_y^{n-1})}{2} dyd\sigma|$ is also bounded by the same estimates as above, and

$$\left| \int_0^t \int_{-L-\sigma}^{\infty} \bar{g} \bar{\delta}_y^n R_y dyd\sigma \right| \le O(1) e^{\frac{-|x|}{4}} e^{\frac{-3t}{16}} e^{\frac{-L}{2} - \frac{L}{6}} |||\bar{\delta}^n|||,$$
$$\left| \int_0^t \int_{-L-\sigma}^{\infty} \bar{g}_x \bar{\delta}_y^n R_y dyd\sigma \right| \le O(1) e^{\frac{-|x|}{4}} e^{\frac{-3t}{16}} e^{\frac{-L}{2} - \frac{L}{6}} |||\bar{\delta}^n|||.$$

Therefore, for $x > (-L-t)/2$,

$$|\bar{\delta}^{n+1}(x,t)|, |\bar{\delta}_x^{n+1}(x,t)| \le O(1)|||\bar{\delta}^n|||e^{\frac{-L}{6}} \cdot \begin{cases} e^{\frac{-|x|}{4}} e^{\frac{-t}{8}} + e^{\frac{-|x|}{2}} (x+t+L+1)^{-\alpha+1}, & \frac{-L-t}{2} < x \le 0, \\ e^{\frac{-|x|}{3}} (x+t+L+1)^{-\alpha+1} + (x+t+L+1)^{-\alpha}, & x > 0. \end{cases} \tag{3.29}$$

From (3.25), (3.27), and (3.29),

$$|||\bar{\delta}^{n+1}||| \le C e^{\frac{-L}{6}} |||\bar{\delta}^n|||. \tag{3.30}$$

Similar estimates also yield

$$|||\bar{\delta}^1||| \le C e^{\frac{-L}{6}} |||\bar{w}^0|||. \tag{3.31}$$

Consequently, when L is sufficiently large, $\{|||\bar{\delta}^m|||\}$ is a geometric sequence such that

$$|||\bar{w}^{n+1} - \bar{w}^0||| \le \sum_{m=1}^{n+1} |||\bar{\delta}^m||| < \frac{1}{2} |||\bar{w}^0|||, \tag{3.32}$$

which implies that $\bar{w}^{n+1}(x,t)$ and $\bar{w}_x^{n+1}(x,t)$ satisfy the required bounds. By mathematical induction, the theorem is true for all $m \in \mathbf{N}$. As a result, there exists a subsequence converging to a limit $\bar{w}(x,t)$ which is the solution of the initial-boundary problem (3.6)–(3.9). The proof is completed. □

4. Effect of nonlinearity, boundary and initial data. In Theorem 3.6, we show that there are three different convergence rates to the shock: exponential near the boundary, exponential in space and algebraic in time $(1+t)^{-\alpha+1}$ behind and near the shock, and algebraic $(1+x+L+t)^{-\alpha}$ in front of the shock. This is so when initial data decays slowly (e.g., algebraically). On the other hand, similar computations show that when initial data decays faster than $e^{-|x|}$, the solution doesn't converge in space as fast as the initial data. The convergence rate is exponential but depends on the viscosity ε and the nonlinearity, that is, the strength $2|u_-|$ of the shock. Thus, consider the initial-boundary value problem

$$u_t + uu_x = \varepsilon u_{xx}, \tag{4.1}$$
$$u(-L-t,t) = u_-, \qquad u(\infty,t) = -u_-, \tag{4.2}$$
$$u(x,0) = \phi_\varepsilon(x) + \bar{u}_\varepsilon(x), \qquad \bar{u}_\varepsilon(-L) = u_- - \phi_\varepsilon(-L), \tag{4.3}$$
$$\phi_\varepsilon(x) \equiv -u_- \tanh\left(\frac{u_- x}{2\varepsilon}\right).$$

As before, we use two different Green's functions for two divided space-time domains, respectively:

I. Near the boundary, we represent the solutions using the Green's function

$$K^{B,\varepsilon}(x,t;y,\sigma) = \frac{1}{\sqrt{4\pi\varepsilon(t-\sigma)}}\left(e^{-\frac{(x-y-u_-(t-\sigma))^2}{4\varepsilon(t-\sigma)}} - e^{-\frac{(x+y+2L-u_-t+(2+u_-)\sigma)^2}{4\varepsilon(t-\sigma)}} e^{-\left(\frac{1+u_-}{\varepsilon}\right)(y+L+\sigma)}\right)$$

of the initial-boundary value problem with viscosity $\varepsilon > 0$

$$w_t + u_- w_x = \varepsilon w_{xx},$$
$$w(-L-t,t) = 0, \qquad w(\infty,t) = 0.$$

II. In the domain which is far away from the boundary, we represent the solutions using the Green's function G^ε for the initial value problem for $w_t + (\phi_\varepsilon(x)w)_x = \varepsilon w_{xx}$:

$$G^\varepsilon(x,t;y,\sigma) = g^\varepsilon(x,t;y,\sigma) + \frac{\int_{-\infty}^{y} \sinh\left(\frac{u_-(x-\xi)}{2\varepsilon}\right) k_\varepsilon(x-\xi,t-\sigma) e^{-\frac{u_-^2(t-\sigma)}{4\varepsilon}} d\xi}{2\cosh^2\frac{u_- x}{2\varepsilon}},$$

where

$$g^\varepsilon(x,t;y,\sigma) \equiv \frac{\cosh\left(\frac{u_- y}{2\varepsilon}\right)}{\cosh\left(\frac{u_- x}{2\varepsilon}\right)} k_\varepsilon(x-y,t-\sigma) e^{-\frac{u_-^2(t-\sigma)}{4\varepsilon}}$$

is the Green's function for

$$w_t + \phi_\varepsilon(x) w_x - \varepsilon w_{xx} = 0$$

and $k_\varepsilon(x,t) \equiv e^{-\frac{x^2}{4\varepsilon t}}/\sqrt{4\pi\varepsilon t}$ is the heat kernel. Of course, we also use the Green's function $\bar{g}^\varepsilon$ for

$$w_t + \phi_\varepsilon(x + x_0^\varepsilon) w_x = \varepsilon w_{xx}$$

to represent the solution in this region, where x_0^ε is the translation and

$$\bar{g}^\varepsilon(x,t;y,\sigma) = g^\varepsilon(x + x_0^\varepsilon, t; y + x_0^\varepsilon, \sigma).$$

As before, the convergence rate is dictated by the effect of the initial data and boundary for $x \geq -(L+t)/2$:

$$\int_{-L}^{\infty} \bar{g}^\epsilon(x,t;y,0)\bar{w}_y(y,0)dy, \tag{4.4}$$

$$\int_0^t \bar{g}_x^\epsilon(x,t;-L-\sigma,\sigma)\bar{w}_y(-L-\sigma,\sigma)d\sigma, \quad \bar{w} \text{ defined as in (3.5).} \tag{4.5}$$

For the particular initial data with the critical decay

$$\bar{u}_\varepsilon(x) = \begin{cases} (u_- - \phi_\varepsilon(-L))e^{-\frac{u_-(x+L)}{\varepsilon}}, & -L \leq x \leq 0, \\ 0, & x > 0, \end{cases}$$

and therefore

$$\bar{w}_x(x,0) = O(1)\begin{cases} e^{-(\frac{u_-}{\varepsilon}L)\beta}, & -L \leq x \leq 0, \\ e^{-(\frac{u_-}{\varepsilon}L)\beta} e^{-\frac{u_-|x|}{\varepsilon}}, & x > 0, \end{cases} \qquad \text{for some } \beta \text{ between } \tfrac{1}{2} \text{ and } 1;$$

then the convergence rates in space and time are those given by the integral (4.4) and (4.5) and are of the order

$$O(1)e^{-\frac{u_-|x|}{2r\varepsilon}}e^{-\frac{u_-^2 t}{4r\varepsilon}} \quad \text{for any fixed } r > 1.$$

Assuming that the initial perturbation has faster decay, for instance, in the extreme case, is of compact support, then the solutions converge no faster than $e^{-u_-|x|/2r\varepsilon}$ in space and $e^{-u_-^2 t/4r\varepsilon}$ in time. In other words, in this case, the nonlinearity $|u_-|$ and the viscosity ε dictate the convergence rate. Note that the effects of viscosity and nonlinearity are different for space and time decay. In the limiting case of hyperbolic conservation law, $\varepsilon \to 0$, perturbation with compact support will decay to zero in finite time T, which is large when the shock strength is small. This is consistent with the above estimate in that the convergence rate is infinite in the limit $\varepsilon \to 0$. For data with either algebraic or exponential decay, the convergence of solution between the boundary and the shock,

$$O(1)e^{-\frac{u_-|x|}{2\varepsilon}}e^{-\frac{u_-^2 t}{4\varepsilon}},$$

depends solely on the nonlinearity and viscosity. This is so because of the effect of the boundary.

Acknowledgment. This work was done while the first two authors visited the Department of Mathematics, Stanford University. They would like to express their sincere gratitude for the hospitality they received there.

REFERENCES

[1] T.-P. Liu, *Hyperbolic and Viscous Conservation Laws*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 72, SIAM, Philadelphia, 2000.

[2] T.-P. Liu, A. Matsumura, and K. Nishihara, *Behaviors of solutions for the Burgers equation with boundary corresponding to rarefaction waves*, SIAM J. Math. Anal., 29 (1998), pp. 293–308.

[3] T.-P. Liu and K. Nishihara, *Asymptotic behavior for scalar viscous conservation laws with boundary effects*, J. Differential Equations, 133 (1997), pp. 296–320.

[4] T.-P. Liu and S.-H. Yu, *Propagation of a stationary shock layer in the presence of a boundary*, Arch. Rational Mech. Anal., 139 (1997), pp. 57–82.

[5] K. Nishihara, *Boundary effect on a stationary viscous shock wave for scalar viscous conservation laws*, J. Math. Anal. Appl., 255 (2001), pp. 535–550.

[6] S.-H. Yu, *The asymptotic behavior of the Burgers equation on the quarter plane*, unpublished.

SIAM J. MATH. ANAL.
Vol. 36, No. 2, pp. 618–642

ON THE SOBOLEV SPACE THEORY OF PARABOLIC AND ELLIPTIC EQUATIONS IN C^1 DOMAINS*

KYEONG-HUN KIM† AND N. V. KRYLOV†

Abstract. Existence and uniqueness results are given for second-order parabolic and elliptic equations with variable coefficients in C^1 domains in Sobolev spaces with weights allowing the derivatives of solutions to blow up near the boundary. The "number" of derivatives can be negative and fractional. The coefficients of parabolic equations are only assumed to be measurable in time.

Key words. parabolic equations, C^1 domains, Sobolev spaces with weights

AMS subject classifications. 35K20, 35J15

DOI. 10.1137/S0036141003421145

1. Introduction.

In this article we deal with the Sobolev space theory of second-order parabolic and elliptic equations in C^1 domains. Since the boundary is not supposed to be regular enough we have to look for solutions in function spaces with weights allowing derivatives of our solutions to blow up near the boundary. In the framework of Hölder spaces such a setting leads to investigating so-called intermediate (or interior) Schauder estimates, which originated in [2]. For results about these estimates the reader is referred to [2], [4], [5] (elliptic case) and [3], [13] (parabolic case).

The main source of our interest in the Sobolev space theory comes from the theory of stochastic partial differential equations (SPDEs). There the Hölder space approach seems not to allow one to obtain results of reasonable generality. On the contrary, the Sobolev space approach works quite well. However, the Sobolev spaces without weights turn out to be trivially inappropriate. Therefore, even if we investigate SPDEs in smooth domains we need to work with weights. Then, naturally, we first need to understand what happens if we are dealing with usual parabolic equations rather than SPDEs. Interestingly, if one studies the problem under natural assumptions, then it becomes irrelevant whether the domain is C^1 or C^∞ (see Theorem 2.12). This is how we ended up with C^1 domains.

Various Sobolev spaces with weights (say, in domains with irregular boundaries or even in the whole space) and their applications to partial differential equations have long been investigated. We do not want to even try to present all relevant references, some of which can be found in [1]. The reader can find some references related to the subject of this article in the papers [9], [14], [15], and [16], the results of which are extensively used in what follows.

Our main results are stated in section 2 and consist of Theorem 2.10 and Theorem 2.14, on solvability of parabolic equations in domains and half-spaces, respectively; Theorem 2.11, treating elliptic equations; and Theorem 2.12, allowing us to reduce the case of general C^1 domains to the case of C^∞ domains. Notice that in Theorem 2.10 we consider only bounded domains; however, actually, the result is

*Received by the editors January 9, 2003; accepted for publication (in revised form) January 16, 2004; published electronically August 6, 2004. This research was partially supported by NSF grant DMS-0140405.

http://www.siam.org/journals/sima/36-2/42114.html

†127 Vincent Hall, University of Minnesota, Minneapolis, MN 55455 (khkim@math.umn.edu, krylov@math.umn.edu).

also true for the domains Ω which are uniformly C^1 smooth in a natural sense. It is assumed usually in the L_p-theory of parabolic equations that the leading coefficients are continuous in the closure of the domain. In our results the coefficients are only assumed to be measurable in t and may substantially oscillate near the boundary.

In section 3 we prove some auxiliary results, and in section 4 the solvability in half-spaces is investigated and Theorem 2.14 is proved. Then in section 5, Theorems 2.10 and 2.11 are proved. The final section, section 6, is devoted to the proof of Theorem 2.12.

It is certainly worth saying that formally speaking, at least under heavier smoothness assumptions on the domain, the continuity of leading coefficients, and the rate with which lower order coefficients are allowed to grow near the boundary, Theorem 2.10 can be obtained from Theorem 4.1 of [14] after deleting all stochastic terms and then claiming that in this situation the restriction $p \geq 2$ can be relaxed to $p > 1$. However, while reading somewhat sketchy proofs in [14] we came to the conclusion that the argument based on renormalization of spaces may be wrong. This is why we decided to give independent proofs in a more general situation but only for the deterministic case. SPDEs will be considered in a subsequent article.

In this paper, as usual $\mathbb{R}^d$ stands for the Euclidean space of points $x = (x^1, \dots, x^d)$, $B_r(x) = \{y \in \mathbb{R}^d : |x-y| < r\}$, $B_r = B_r(0)$, $\mathbb{R}^d_+ = \{x \in \mathbb{R}^d : x^1 > 0\}$. For $i = 1, \dots, d$, multi-indices $\alpha = (\alpha_1, \dots, \alpha_d)$, $\alpha_i \in \{0, 1, 2, \dots\}$, and functions $u(x)$ we set

$$u_{x^i} = \partial u / \partial x^i = D_i u, \quad D^\alpha u = D_1^{\alpha_1} \times \cdots \times D_d^{\alpha_d} u, \quad |\alpha| = \alpha_1 + \cdots + \alpha_d.$$

2. Main results. Let Ω be an open set in $\mathbb{R}^d$, $\Omega \neq \mathbb{R}^d$. First, we consider the equation

$$u_t(t,x) = a^{ij}(t,x) u_{x^i x^j}(t,x) + b^i(t,x) u_{x^i}(t,x) + c(t,x) u(t,x) + f(t,x) \tag{2.1}$$

given for $x \in \Omega$, $t \geq 0$.

Take an increasing function $\omega_0(\varepsilon)$ defined on $[0, \infty)$ and such that $\omega_0(\varepsilon) \downarrow 0$ as $\varepsilon \downarrow 0$. Also take some numbers $r_0, K_0 \in (0, \infty)$.

Assumption 2.1. The domain Ω is of class C^1_u. In other words, for any $x_0 \in \partial\Omega$, there exists a one-to-one continuously differentiable mapping Ψ of $B_{r_0}(x_0)$ onto a domain $G \subset \mathbb{R}^d$ such that

(i) $G_+ := \Psi(B_{r_0}(x_0) \cap \Omega) \subset \mathbb{R}^d_+$ and $\Psi(x_0) = 0$;
(ii) $\Psi(B_{r_0}(x_0) \cap \partial\Omega) = G \cap \{y \in \mathbb{R}^d : y^1 = 0\}$;
(iii) $\|\Psi\|_{C^1(B_{r_0}(x_0))} \leq K_0$ and $|\Psi^{-1}(y_1) - \Psi^{-1}(y_2)| \leq K_0 |y_1 - y_2|$ for any $y_i \in G$;
(iv) for $x_1, x_2 \in B_{r_0}(x_0)$, we have $|\Psi_x(x_1) - \Psi_x(x_2)| \leq \omega_0(|x_1 - x_2|)$.

To state our assumptions on a, b, c we introduce the following notation. Set

$$\rho(x) = \rho_\Omega(x) = \operatorname{dist}(\mathrm{x}, \partial\Omega), \quad \rho(\mathrm{x}, \mathrm{y}) = \rho_\Omega(\mathrm{x}, \mathrm{y}) = \rho(\mathrm{x}) \wedge \rho(\mathrm{y})$$

and according to [2] and [5] for $\sigma \in \mathbb{R}$, $\alpha \in (0,1)$, and $k = 0, 1, 2, \dots$ introduce

$$[f]^{(\sigma)}_k = [f]^{(\sigma)}_{k,\Omega} = \sup_{|\beta|=k} \sup_{x \in \Omega} \rho^{k+\sigma}(x) |D^\beta f(x)|,$$

$$[f]^{(\sigma)}_{k+\alpha} = [f]^{(\sigma)}_{k+\alpha,\Omega} = \sup_{|\beta|=k} \sup_{x,y \in \Omega} \rho^{k+\alpha+\sigma}(x,y) \frac{|D^\beta f(x) - D^\beta f(y)|}{|x-y|^\alpha}, \tag{2.2}$$

$$|f|^{(\sigma)}_k = |f|^{(\sigma)}_{k,\Omega} = \sum_{j=0}^k [f]^{(\sigma)}_{j,\Omega}, \quad |f|^{(\sigma)}_{k+\alpha} = |f|^{(\sigma)}_{k+\alpha,\Omega} = |f|^{(\sigma)}_{k,\Omega} + [f]^{(\sigma)}_{k+\alpha,\Omega}.$$

Remark 2.2. We did not specify what kind of derivatives are $D^\beta f$. These are either classical derivatives or Sobolev ones. In the latter case, of course, instead of sup we should have used ess sup. Also, it is worth pointing out that the norms $|\cdot|_\gamma^{(\sigma)}$ introduced for all $\gamma \geq 0$ and $\sigma \in \mathbb{R}$ possess quite peculiar properties if γ is not an integer and $\gamma+\sigma < 0$. In that case, for instance, $[f]_\gamma^{(\sigma)} = \infty$ unless $D^\beta f \equiv 0$ whenever $|\beta| = [\gamma]$, so that it may happen that $|f|_\nu^{(\sigma)} < \infty$ for a $\nu > \gamma$ but $|f|_\gamma^{(\sigma)} = \infty$.

We also fix a function $\delta_0(\tau) \geq 0$ defined on $[0,\infty)$ such that $\delta_0(\tau) > 0$ unless $\tau \in \{0,1,2,\ldots\}$. For $\tau \geq 0$ define

$$\tau+ = \tau + \delta_0(\tau).$$

Finally, fix some constants

$$\delta, K \in (0,\infty), \quad \gamma \in \mathbb{R}, \quad p \in (1,\infty).$$

Assumption 2.3. (i) The functions a, b, c are Borel measurable in (t,x), $a^{ij} = a^{ji}$.

(ii) For any $t > 0, x \in \Omega$, and $\lambda \in \mathbb{R}^d$,

$$\delta|\lambda|^2 \leq a^{ij}(t,x)\lambda^i\lambda^j \leq K|\lambda|^2. \tag{2.3}$$

(iii) For each $t > 0$,

$$|a(t,\cdot)|_{|\gamma|+}^{(0)} + |b(t,\cdot)|_{|\gamma|+}^{(1)} + |c(t,\cdot)|_{|\gamma|+}^{(2)} \leq K.$$

Assumption 2.4. (i) The function $a(t,\cdot)$ is continuous at any point $x \in \Omega$ uniformly with respect to t.

(ii) There is a control on the behavior of a, b, and c near $\partial\Omega$, namely,

$$\lim_{\substack{\rho(x)\to 0,\\ x\in\Omega}} \sup_{\substack{y\in\Omega,\\ |x-y|\leq\rho(x,y)}} \sup_t |a(t,x)-a(t,y)| = 0. \tag{2.4}$$

$$\lim_{\substack{\rho(x)\to 0,\\ x\in\Omega}} \sup_t [\rho(x)|b(t,x)| + \rho^2(x)|c(t,x)|] = 0.$$

Remark 2.5. Equation (2.4) has very little to do with the uniform continuity of a in Ω (which is assumed, for instance, in [14]) and, for that matter, even with its pointwise continuity. For instance, if $\delta \in (0,1)$, $d = 1$, and $\Omega = \mathbb{R}_+$, then the function $a(x)$ equal to $2 + \sin(|\ln x|^\delta)$ for $0 < x \leq 1/2$ satisfies (2.4).

Indeed, if $x, y > 0$ and $|x-y| \leq x \wedge y$, then

$$|a(x) - a(y)| = |x-y||a'(\xi)|,$$

where ξ lies between x and y. In addition, $|x-y| \leq x \wedge y \leq \xi \leq 2(x \wedge y)$, and $\xi|a'(\xi)| \leq |\ln[2(x \wedge y)]|^{\delta-1} \to 0$ as $x \wedge y \to 0$. The function $a(x)$ also satisfies Assumption 2.3 for any γ if we change it appropriately for $x > 1/2$.

To proceed further we state a version of well-known results from [4] and [12], which we discuss in section 6.

LEMMA 2.6. *There is a bounded real-valued function ψ defined in $\bar\Omega$ such that*

(i) *$\psi(x) > 0$ in Ω, $\psi = 0$ on $\partial\Omega$, and for any $\varepsilon > 0$ the function ψ is bounded away from zero on the set $\{x \in \Omega : \rho(x) \geq \varepsilon\}$;*

(ii) *ψ_x is uniformly continuous in $\bar\Omega$, and $|\psi_x(x)| \geq 1$ on $\partial\Omega$;*

(iii) *for any multi-index* α *we have*

$$\sup_{\Omega} \rho^{|\alpha|}(x)|D^{\alpha}\psi_x(x)| < \infty.$$

(iv) *for any multi-index* $\alpha \neq 0$ *we have* $\rho^{|\alpha|}(x)|D^{\alpha}\psi_x(x)| \to 0$ *as* $x \in \Omega$ *and* $\rho(x) \to 0$.

Remark 2.7. In the part of a neighborhood of $\partial\Omega$ lying in Ω the functions ψ and ρ are comparable in the following sense.

For $x \in \Omega$, the ratio $\psi(x)/\rho(x)$ equals $\psi_{x^i}(\xi)\tau^i$, where $\tau^i = (x^i - x_0^i)/|x - x_0|$, x_0 is one of the closest points to x on $\partial\Omega$, and ξ is a point between x and x_0. It follows from (iii) that $\psi(x) \leq N\rho(x)$, where N is independent of x. On the other hand, (ii) implies that

$$\psi_{x^i}(\xi)\tau^i = \psi_{x^i}(x_0)\tau^i + o(|x - x_0|) = |\psi_x(x_0)| + o(|x - x_0|) \geq 1 + o(|x - x_0|).$$

It follows that, if $\rho(x)$ is small enough, then $\rho(x) \leq 2\psi(x)$. Thus, there exists an $\varepsilon > 0$ such that in $\{x \in \Omega : \rho(x) \leq \varepsilon\}$ we have $(1/2)\rho(x) \leq \psi(x) \leq N\rho(x)$.

Also notice that the functions ψ and ρ are comparable in Ω if Ω is bounded. Therefore, in many situations one can interchange $\psi(x)$ and $\rho(x)$. An advantage of using ψ on some occasions is that this function is infinitely differentiable. For instance, we prove the following fact in section 3.

LEMMA 2.8. *Let* ψ *be a function as in Lemma* 2.6 *and let* Ω *be bounded. Let* $\mu \in \mathbb{R}$, $\tau \geq 0$, $\kappa \leq \sigma$, *and either* $\sigma + [\tau] \geq 0$ *or* $\tau \in \{0, 1, 2 \ldots\}$. *Then* (i) $|\psi^{-\kappa}|_{\tau}^{(\sigma)} < \infty$; (ii) *for any function* a *we have* $|a|_{\tau}^{(\sigma)} \leq N|\psi^{\sigma}a|_{\tau}^{(0)}$ *and* $|\psi^{\mu}a|_{\tau}^{(\sigma)} \leq N|a|_{\tau}^{(\mu+\sigma)}$, *where* N *is independent of* a.

To describe the assumptions on f we use the Banach spaces introduced in [15]. Let $\xi \in C_0^{\infty}(\mathbb{R}_+)$ be a function satisfying

$$\sum_{n=-\infty}^{\infty} \xi(e^{n+t}) > 0 \quad \forall t \in \mathbb{R}. \tag{2.5}$$

For $x \in \Omega$ and $n \in \mathbb{Z} = \{0, \pm 1, \ldots\}$ define

$$\zeta_n(x) = \xi(e^n\psi(x)).$$

Observe that, due to (2.5), we have $\sum_n \zeta_n \geq \text{const} > 0$ in Ω. Also in Ω by virtue of Lemma 2.6(i)–(iii) we have

$$\zeta_n \in C_0^{\infty}(\Omega), \quad |D^m\zeta_n(x)| \leq Ne^{mn}, \tag{2.6}$$

where N is independent of n and x. For any distribution u on Ω, the first relation in (2.6) allows us to define $u\zeta_n$ as a distribution on $\mathbb{R}^d$ (equal to zero outside of Ω).

Now, for $\theta, \gamma \in \mathbb{R}$, let $H_{p,\theta}^{\gamma}(\Omega)$ be the set of all distributions u on Ω such that

$$\|u\|_{H_{p,\theta}^{\gamma}(\Omega)}^p := \sum_{n \in \mathbb{Z}} e^{n\theta}\|\zeta_{-n}(e^n\cdot)u(e^n\cdot)\|_{H_p^{\gamma}}^p < \infty, \tag{2.7}$$

where $H_p^{\gamma} = (1 - \Delta)^{-\gamma/2}L_p(\mathbb{R}^d, dx)$.

Remark 2.9. It is known (see, for instance, [15]) that up to equivalent norms the space $H_{p,\theta}^{\gamma}(\Omega)$ is independent of the choice of ξ and ψ if Ω is bounded and, if γ is a

nonnegative integer, then $H^\gamma_{p,\theta}(\Omega)$ is the space of all distributions u on Ω such that $\rho^{|\alpha|+(\theta-d)/p}D^\alpha u \in L_p(\Omega, dx)$, $|\alpha| \le \gamma$, provided with a natural norm.

For convenience we may assume that $\xi(t) = 0$ if $t \le \sup_\Omega \psi$. In that case $\zeta_n = 0$ for $n \le 0$ and the sum in (2.7) can be taken only over $n \le -1$.

Denote

$$\mathbb{H}^\gamma_{p,\theta}(\Omega, T) = L_p((0,T), H^\gamma_{p,\theta}(\Omega)), \quad U^\gamma_{p,\theta}(\Omega) = \psi^{1-2/p}H^{\gamma-2/p}_{p,\theta}(\Omega),$$

and by $\mathfrak{H}^\gamma_{p,\theta}(\Omega, T)$ we denote the space of all functions $u \in \psi\mathbb{H}^\gamma_{p,\theta}(\Omega, T)$ such that, for some $u_0 \in U^\gamma_{p,\theta}(\Omega)$ and $f \in \psi^{-1}\mathbb{H}^{\gamma-2}_{p,\theta}(\Omega, T)$, we have

$$(u(t), \phi) = (u_0, \phi) + \int_0^t (f(s), \phi)\, ds \quad \forall t \le T, \phi \in C_0^\infty(\Omega).$$

Naturally, we denote $u_t = f$ and $u(0) = u_0$. The norm in $\mathfrak{H}^\gamma_{p,\theta}(\Omega, T)$ is introduced by

$$\|u\|_{\mathfrak{H}^\gamma_{p,\theta}(\Omega,T)} = \|\psi^{-1}u\|_{\mathbb{H}^\gamma_{p,\theta}(\Omega,T)} + \|\psi u_t\|_{\mathbb{H}^{\gamma-2}_{p,\theta}(\Omega,T)} + \|u(0)\|_{U^\gamma_{p,\theta}(\Omega)}.$$

Finally, let $\mathfrak{H}^\gamma_{p,\theta,0}(\Omega, T) = \mathfrak{H}^\gamma_{p,\theta}(\Omega, T) \cap \{u : u(0) = 0\}$.

From this point on, we assume that

$$d - 1 < \theta < d - 1 + p.$$

Here is our first main result.

THEOREM 2.10. *Let Ω be bounded and $T \in [0, \infty)$. Then under the above assumptions,*

(i) *for any $f \in \psi^{-1}\mathbb{H}^\gamma_{p,\theta}(\Omega, T)$ and $u_0 \in U^{\gamma+2}_{p,\theta}(\Omega)$, (2.1) with initial data u_0 admits a unique solution u in the class $\mathfrak{H}^{\gamma+2}_{p,\theta}(\Omega, T)$;*

(ii) *for this solution*

$$\|\psi^{-1}u\|_{\mathbb{H}^{\gamma+2}_{p,\theta}(\Omega,T)} \le Ne^{NT}\left(\|u_0\|_{U^{\gamma+2}_{p,\theta}(\Omega)} + \|\psi f\|_{\mathbb{H}^\gamma_{p,\theta}(\Omega,T)}\right), \tag{2.8}$$

where the constant N is independent of T, f, and u_0.

The following theorem is obtained in section 5 rather easily from Theorem 2.10. It extends Theorem 5.1 of [15], in which there is the requirement that $c \le -c_0/\rho(x)$ with a sufficiently large constant c_0. On the other hand, it should be noted that in [15] there are no restrictions on θ and no assumptions on the smoothness of Ω.

THEOREM 2.11. *Let Ω be bounded and the above assumptions be satisfied. Let a, b, c be independent of t and let c_0 be a sufficiently large constant (actually, any constant bigger than N from (2.8)). Then for any $f \in \psi^{-1}H^\gamma_{p,\theta}(\Omega)$ there is a unique $u \in \psi H^{\gamma+2}_{p,\theta}(\Omega)$ such that, in Ω,*

$$a^{ij}u_{x^ix^j} + b^iu_{x^i} + (c - c_0)u + f = 0. \tag{2.9}$$

Furthermore,

$$\|\psi^{-1}u\|_{H^{\gamma+2}_{p,\theta}(\Omega)} \le N\|\psi f\|_{H^\gamma_{p,\theta}(\Omega)}, \tag{2.10}$$

where the constant N is independent of f.

One of important ingredients in the proof of Theorem 2.10 is the following result, which allows us to reduce the case of general C^1 domains to the case of C^∞ domains. For $\varepsilon > 0$ set

$$\Omega_\varepsilon = \{x \in \Omega : \psi(x) > \varepsilon\}.$$

THEOREM 2.12. *There is an $\varepsilon > 0$ and a C^∞ diffeomorphism $\mu : \Omega_\varepsilon \to \Omega$ such that, for $\nu = \mu^{-1}$,*

(i) *the functions μ_x and ν_x are uniformly continuous in Ω_ε and Ω, respectively;*

(ii) *for any $n = 0, 1, 2, \dots$, we have $|\mu_x|^{(0)}_{n,\Omega_\varepsilon} + |\nu_x|^{(0)}_{n,\Omega} < \infty$;*

(iii) *for any multi-index $\alpha \neq 0$ we have $(\psi(x) - \varepsilon)^{|\alpha|} D^\alpha \mu_x(x) \to 0$ as $x \in \Omega_\varepsilon$ and $\psi(x) - \varepsilon \downarrow 0$;*

(iv) *for any multi-index $\alpha \neq 0$ we have $\rho^{|\alpha|} D^\alpha \nu_x(x) \to 0$ as $x \in \Omega$ and $\rho(x) \downarrow 0$;*

(v) *in the part of a neighborhood of $\partial\Omega_\varepsilon$ lying in Ω_ε we have $\psi(\mu(x)) = \psi(x) - \varepsilon$;*

(vi) *if Ω is bounded, then $\rho_\Omega \leq N \rho_{\Omega_\varepsilon}(\nu)$ in Ω and $\rho_{\Omega_\varepsilon} \leq N \rho_\Omega(\mu)$ in Ω_ε, where N is a finite constant.*

The proof of Theorem 2.10 is also based on the following result for $\mathbb{R}^d_+$. Below, the spaces $H^\gamma_{p,\theta}$, $\mathbb{H}^\gamma_{p,\theta}(T)$, and $\mathfrak{H}^\gamma_{p,\theta}(T)$ are taken from [9]. They are defined on the basis of (2.7), where we formally take $\Omega = \mathbb{R}^d_+$ and $\psi(x) = x^1$, so that $\zeta_{-n}(e^n x) = \xi(x^1) =: \zeta(x)$ and

$$\|u\|^p_{H^\gamma_{p,\theta}} := \sum_{n \in \mathbb{Z}} e^{n\theta} \|u(e^n \cdot)\zeta\|^p_{H^\gamma_p} < \infty.$$

As in [9] by M^α we denote the operator of multiplying by $(x^1)^\alpha$ and $M = M^1$.

Remark 2.13 (see [9]). If $\gamma = 0, 1, 2, \dots$, then $\|u\|^p_{H^\gamma_{p,\theta}}$ is equivalent to

$$\sum_{|\alpha| \leq \gamma} \int_{\mathbb{R}^d_+} (x^1)^{\theta - d} |(x^1)^{|\alpha|} D^\alpha u(x)|^p \, dx.$$

THEOREM 2.14. *Let $\Omega = \mathbb{R}^d_+$, $\omega \in (0, \infty)$, $T \in (0, \infty]$. Drop Assumption 2.4 and instead suppose that*

$$|a(t,x) - a(t,y)| + x^1 |b(t,x)| + (x^1)^2 |c(t,x)| \leq \omega$$

whenever $t > 0$, $x, y \in \Omega$, and $|x - y| \leq x^1 \wedge y^1$. Suppose that all other assumptions are satisfied.

Then there exists an $\omega_0 \in (0,1)$ depending only on δ, p, θ, γ, $|\gamma|+$, and K, such that, if $\omega \leq \omega_0$, then

(i) *for any $f \in M^{-1}\mathbb{H}^\gamma_{p,\theta}(T)$ and $u_0 \in U^{\gamma+2}_{p,\theta}$, (2.1) with initial data u_0 admits a unique solution u in the class $\mathfrak{H}^{\gamma+2}_{p,\theta}(T)$;*

(ii) *for this solution*

$$\|M^{-1}u\|_{\mathbb{H}^{\gamma+2}_{p,\theta}(T)} \leq N \left(\|u_0\|_{U^{\gamma+2}_{p,\theta}} + \|Mf\|_{\mathbb{H}^\gamma_{p,\theta}(T)} \right), \tag{2.11}$$

where the constant N depends only on $p, \delta, \theta, \gamma$, $|\gamma|+$, and K.

3. Auxiliary results. The goal of this section is to write multidimensional versions of the results of section 3 in [6] and to develop certain techniques for dealing with the norms $|\cdot|^{(\sigma)}_\gamma$. In the following lemma, no restriction on θ is needed. One can prove

that its statement also holds true if one replaces $\mathbb{R}^d_+$ and $H^\gamma_{p,\theta}$ with Ω and $H^\gamma_{p,\theta}(\Omega)$, respectively. However, such a modification is of no use for us because we needed the lemma as it is stated and the norms in $H^\gamma_{p,\theta}(\mathbb{R}^d_+)$ and $H^\gamma_{p,\theta}$ are not equivalent (since ψ is bounded).

LEMMA 3.1. *Let constants $C, \delta \in (0,\infty)$, a function $u \in H^\gamma_{p,\theta}$, and q be the smallest integer such that $|\gamma| + 2 \le q$.*

(i) *Let $\eta_k \in C^\infty(\mathbb{R}^d_+)$, $k = 1, 2, \dots$, satisfy*

$$\sum_k M^{|\alpha|} |D^\alpha \eta_k| \le C \quad in \quad \mathbb{R}^d_+ \tag{3.1}$$

for any multi-index α such that $0 \le |\alpha| \le q$. Then

$$\sum_k \|\eta_k u\|^p_{H^\gamma_{p,\theta}} \le N C^p \|u\|^p_{H^\gamma_{p,\theta}},$$

where the constant N is independent of u, θ, and C.

(ii) *If in addition to the condition in* (i)

$$\sum_k \eta_k^2 \ge \delta \quad on \quad \mathbb{R}^d_+, \tag{3.2}$$

then

$$\|u\|^p_{H^\gamma_{p,\theta}} \le N \sum_k \|\eta_k u\|^p_{H^\gamma_{p,\theta}}, \tag{3.3}$$

where the constant N is independent of u and θ.

Proof. (i) One may assume that $C = 1$ because one can replace η_k with η_k/C. Then since different functions ξ generate equivalent norms, we have

$$\sum_k \|\eta_k u\|^p_{H^\gamma_{p,\theta}} \le N \sum_n e^{n\theta} \sum_k \|u(e^n\cdot)\zeta\eta_{kn}\|^p_{H^\gamma_p},$$

where $\eta_{kn} = \eta_k(e^n\cdot)\zeta$. Furthermore, observe that by the Leibniz rule

$$I_{n\alpha} := \sum_k |D^\alpha \eta_{kn}(x)| \le N \sum_k \sum_{|\beta|+|\gamma|=|\alpha|} e^{n|\beta|} |D^\beta \eta_k|(e^n x)\, |D^\gamma \zeta(x)|$$

and that on the support of ζ we have $e^{n|\beta|} \le N(e^n x^1)^{|\beta|}$. Then, upon recalling (3.1), we see that $I_{n\alpha}$ are bounded by a constant independent of $x \in \mathbb{R}^d$, $n \in \mathbb{Z}$, and α such that $|\alpha| \le q$.

It follows by Theorem 2.1 and Remark 2.1 of [7] that, for each n,

$$\sum_k \|u(e^n\cdot)\zeta\eta_{kn}\|^p_{H^\gamma_p} \le N \|u(e^n\cdot)\zeta\|^p_{H^\gamma_p}.$$

Formally speaking, to use Theorem 2.1 in [7], we need condition (3.1) to be satisfied for all multi-indices α rather than only such that $|\alpha| \le q$. That it suffices to dominate these quantities for $|\alpha| \le q$ follows by inspecting the argument in [7]. Hence,

$$\sum_k \|\eta_k u\|^p_{H^\gamma_{p,\theta}} \le N \sum_n e^{n\theta} \|u(e^n\cdot)\zeta\|^p_{H^\gamma_p} \le N \|u\|^p_{H^\gamma_{p,\theta}}.$$

This proves (i) and allows us to use the same argument as in Remark 2.1 of [7]. Assertion (i) means that the operator mapping $u \in H^\gamma_{p,\theta}$ into $(\eta_k u, k = 1, 2, \ldots) \in \ell_p(H^\gamma_{p,\theta})$ is bounded. Its dual is also bounded, which means that (due to the arbitrariness of p, γ, θ we do not use new parameters for dual spaces) if $\sum_k \|g_k\|^p_{H^\gamma_{p,\theta}} < \infty$, then $\sum_k \eta_k g_k \in H^\gamma_{p,\theta}$ and

$$\left\| \sum_k \eta_k g_k \right\|^p_{H^\gamma_{p,\theta}} \leq N \sum_k \|g_k\|^p_{H^\gamma_{p,\theta}}. \tag{3.4}$$

Under the condition in (ii), it turns out that here in place of η_k one can take $\tilde{\eta}_k := \eta_k/\bar{\eta}$, where $\bar{\eta} = \sum_i \eta_i^2$. Indeed, it is easy to deduce from (3.1) and the inequality $\sum |ab| \leq \sum |a|(\sum |b|)$ that $M^{|\alpha|} D^\alpha \bar{\eta}$ is bounded if $0 \leq |\alpha| \leq q$. Then one gets the same property for $1/\bar{\eta}$ by relying on (3.2). This makes it clear that $\tilde{\eta}_k$ satisfy (3.1) with certain C. Finally, by taking $\tilde{\eta}_k = \eta_k / \sum_i \eta_i^2$ and $\eta_k u$ in place of η_k and g_k, respectively, in (3.4) we get (3.3). The lemma is proved.

Remark 3.2. In Lemma 3.1 we assumed that $u \in H^\gamma_{p,\theta}$. In this connection it is important to observe that the above proof shows also that if the right-hand side of (3.3) is finite, then $u \in H^\gamma_{p,\theta}$.

Notice that the first inequality in (3.5) below is written for η_k^4 and not for η_k^2 as in Lemma 3.1. The purpose of this is to have the possibility to apply Lemma 3.1 to η_k^2 in place of η_k. In this connection it is useful to have in mind that $\sum |ab| \leq \sum |a|(\sum |b|)$ and $\sum a^2 \leq (\sum |a|)^2$.

LEMMA 3.3. *For each $\varepsilon > 0$ and $q = 1, 2, \ldots$ there exist nonnegative functions $\eta_k \in C_0^\infty(\mathbb{R}^d_+)$, $k = 1, 2, \ldots$, such that* (i) *on $\mathbb{R}^d_+$ for each multi-index α with $1 \leq |\alpha| \leq q$ we have*

$$\sum_k \eta_k^4 \geq 1, \quad \sum_k \eta_k \leq N(d), \quad \sum_k M^{|\alpha|} |D^\alpha \eta_k| \leq \varepsilon; \tag{3.5}$$

(ii) *for any k and $x, y \in \operatorname{supp} \eta_k$ we have $|x - y| \leq N(x^1 \wedge y^1)$, where $N = N(d, q, \varepsilon) \in [1, \infty)$.*

Proof. Let

$$\mathbb{R}^{d-1} = \bigcup_{k=1}^\infty Q_k$$

be a decomposition of $\mathbb{R}^{d-1}$ into disjoint unit cubes Q_k. Mollify the indicator function of each Q_k in such a way that thus obtained function χ_k vanish outside of the twice dilated Q_k (naturally, with center of dilation being that of Q_k). Then

$$\delta \leq \sum_k \chi_k^2 \leq \left(\sum_k \chi_k \right)^2 \leq N$$

on $\mathbb{R}^{d-1}$ for some constants $\delta, N \in (0, \infty)$ depending only on d. Furthermore, by Lemma 3.2 of [6] there exists a nonnegative function $\xi \in C_0^\infty(\mathbb{R}_+)$ such that assertion (i) of the present lemma holds for $d = 1$ with the collection $\{\xi(e^n x) : n \in \mathbb{Z}\}$ in place of $\{\eta_k(x) : k = 1, 2, \ldots\}$.

Then write $x=(x^1,x')$, fix a constant $r\in(0,1)$ to be specified later, and introduce

$$\tau_k(x')=\chi_k(rx'),\quad \eta_{nk}(x)=\xi(e^nx^1)\tau_k(e^nx').$$

Then (first sum with respect to k)

$$\delta\leq\sum_{n,k}\eta_{nk}^4\leq\left(\sum_{n,k}\eta_{nk}\right)^4\leq N \tag{3.6}$$

on $\mathbb{R}^d_+$ for some constants $\delta,N\in(0,\infty)$ depending only on d.

Now, for $1\leq|\alpha|\leq q$ and some constants $c_{\beta\gamma}$, we have

$$M^{|\alpha|}D^\alpha\eta_{nk}(x)=(x^1)^{|\alpha|}e^{n|\alpha|}\sum_{\beta+\gamma=\alpha}c_{\beta\gamma}\xi^{(\beta_1)}(e^nx^1)(D^\gamma\tau_k)(e^nx').$$

Hence,

$$\sum_{n,k}|M^{|\alpha|}D^\alpha\eta_{nk}(x)|\leq\sum_{\beta+\gamma=\alpha}c_{\beta\gamma}I_1(\gamma)I_2(\alpha,\beta),$$

where

$$I_1(\gamma)=\sup_{x'}\sum_k|D^\gamma\tau_k(x')|=r^{|\gamma|}\sup_{x'}\sum_k|D^\gamma\chi_k(x')|,$$

$$I_2(\alpha,\beta)=\sup_{t\geq0}\sum_n t^{|\alpha|}e^{n|\alpha|}|\xi^{(\beta_1)}(e^nt)|=\sup_{t\in\mathbb{R}}\sum_n e^{(n+t)|\alpha|}|\xi^{(\beta_1)}(e^{n+t})|.$$

Obviously I_1 is finite. That I_2 is also finite is seen from its representation as the supremum of a continuous 1-periodic function. Moreover, if $\gamma=0$, then $c_{\beta\gamma}\neq0$ only if $\beta_1=|\alpha|$, in which case $c_{\beta\gamma}=1$ and, by the construction of ξ, we have $I_2(\alpha,\beta)\leq\varepsilon$. It follows that

$$\sum_{n,k}|M^{|\alpha|}D^\alpha\eta_{nk}(x)|\leq N(d)\varepsilon+N(\varepsilon,q,d)r. \tag{3.7}$$

We renumber the set $\{\eta_{kn}:n=0,\pm1,\dots,k=1,2,\dots\}$ and write it as $\{\eta_k:k=1,2,\dots\}$. Then from (3.7) we see how to choose r to satisfy the last inequality in (3.5) with $N(d)\varepsilon$ in place of ε.

Equation (3.6) shows that $N(d)\eta_k$, with an appropriate $N(d)$, satisfy the first two inequalities in (3.5). However, ε in (3.5) will be replaced with $N(d)\varepsilon$. This is, of course, irrelevant since from the very beginning we could take a smaller constant instead of ε. This proves (i).

To prove (ii) notice that if $x,y\in\operatorname{supp}\eta_{0k}$, then $x^1,y^1\in\operatorname{supp}\xi$ and x^1,y^1 are separated away from zero and bounded above, whereas $x',y'\in\operatorname{supp}\tau_k$, so that $|x'-y'|$ is bounded above independently of k. In that case $|x-y|\leq N(d,q,\varepsilon)(x^1\wedge y^1)$. This relation is dilation invariant and therefore holds for any n,k, and $x,y\in\operatorname{supp}\eta_{nk}$. The lemma is proved.

LEMMA 3.4. *Let $0\leq\gamma\leq\tau$, $\sigma_1,\sigma_2,\sigma,\kappa\in\mathbb{R}$, and*

$$\textit{either}\quad[\gamma]+\sigma\geq0,\quad[\gamma]+\sigma_1+\sigma_2\geq0\quad\textit{or}\quad\gamma\in\{0,1,2,\dots\}.$$

Then, with $N = N(\gamma, \sigma, \sigma_1, \sigma_2, d)$, we have

$$|a|_\gamma^{(\sigma)} \le N|a|_\tau^{(\sigma)}, \quad |ab|_\gamma^{(\sigma_1+\sigma_2)} \le N|a|_\gamma^{(\sigma_1)}|b|_\gamma^{(\sigma_2)},$$
$$|a|_{\gamma+1}^{(\kappa)} \le N(|a|_0^{(\kappa)} + |a_x|_\gamma^{(\kappa+1)}), \quad |a|_0^{(\kappa)} + |a_x|_\gamma^{(\kappa+1)} \le N|a|_{\gamma+1}^{(\kappa)}.$$

This result is quite standard (for various particular cases of it we refer to [2] and [5]) and is based on simple manipulations. We only mention three main ingredients.

The first is that if we take the sup in (2.2) only over $x, y \in \Omega$ such that $4|x-y| \le \rho(x,y)$, then the norm $|\cdot|_{k+\alpha}^{(\sigma)}$ will be replaced with an equivalent one provided that $k + \sigma \ge 0$. This is because $\rho^{k+\sigma}(x,y) \le \rho^{k+\sigma}(x)$ and $\rho(x,y)/|x-y| \le 4$ when $4|x-y| \ge \rho(x,y)$. This replacement allows one to connect x, y by a straight segment lying in Ω and use that $(\rho(x,y)/|x-y|)^\alpha$ increases with α.

The second ingredient is the observation that if $4|x-y| \le \rho(x,y)$, then $(1/2)\rho(x) \le \rho(x,y) \le \rho(x)$, and one can raise this inequality to any power. The third ingredient is the Leibniz rule.

The following interpolation lemma is a particular case of Proposition 4.2 of [13] (also see www.math.iastate.edu/lieb/book/errata.pdf) stated in more general form and for norms based on parabolic distances. Various versions of the lemma also can be found in many other places (see, for instance, [2] and [5]).

LEMMA 3.5. *If $0 \le \kappa \le \tau < \infty$, then*

$$|a|_\kappa^{(0)} \le N(\kappa, \tau, d) \left(\sup_\Omega |a|\right)^{1-\kappa/\tau} \left(|a|_\tau^{(0)}\right)^{\kappa/\tau}.$$

Notice that we only need Lemma 3.5 for $\Omega = \mathbb{R}^d_+$. The next result for $\Omega = \mathbb{R}^d_+$ bears on multiplicators in $H^\gamma_{p,\theta}$.

LEMMA 3.6. *Let $p \in (1,\infty)$, $\gamma, \theta \in \mathbb{R}$. Then there exists a constant $N = N(\gamma, |\gamma|+, p, d)$ such that if $f \in H^\gamma_{p,\theta}$ and a is a function with finite norm $|a|^{(0)}_{|\gamma|+,\mathbb{R}^d_+}$, then*

$$\|af\|_{H^\gamma_{p,\theta}} \le N|a|^{(0)}_{|\gamma|+,\mathbb{R}^d_+} \|f\|_{H^\gamma_{p,\theta}}. \tag{3.8}$$

In addition, if $\gamma = 0, 1, 2, \dots$, then

$$\|af\|_{H^\gamma_{p,\theta}} \le N \sup_{\mathbb{R}^d_+} |a| \, \|f\|_{H^\gamma_{p,\theta}} + N\|f\|_{H^{\gamma-1}_{p,\theta}} \sup_{\mathbb{R}^d_+} \sup_{1 \le |\alpha| \le \gamma} |M^{|\alpha|} D^\alpha a|. \tag{3.9}$$

Proof. Since the norms in $H^\gamma_{p,\theta}$ constructed from different ζ are equivalent, we have

$$\|af\|^p_{H^\gamma_{p,\theta}} \le N \sum_n e^{n\theta} \|\zeta(\cdot)a(e^n\cdot)\zeta(\cdot)f(e^n\cdot)\|^p_{H^\gamma_p}.$$

Furthermore, for any n (see, for instance, Lemma 5.2 in [8]),

$$\|\zeta(\cdot)a(e^n\cdot)\zeta(\cdot)f(e^n\cdot)\|_{H^\gamma_p} \le N|a(e^n\cdot)\zeta|_{B^{|\gamma|+}} \|f(e^n\cdot)\zeta\|_{H^\gamma_p} =: I,$$

where $|\cdot|_{B^\nu}$ is a natural Hölder's norm in $\mathbb{R}^d$. Now we use, first, that for functions with support belonging to that of ζ the norm $|\cdot|_{B^\nu}$ is equivalent to $|\cdot|_\nu^{(0)}$, then the multiplicative property of $|\cdot|_\nu^{(0)}$ from Lemma 3.4 and, finally, the simple fact that the

norms $|\cdot|^{(0)}_{\nu}$ are dilation invariant. Then we see that

$$I \leq N|a(e^n\cdot)\zeta|^{(0)}_{|\gamma|+,\mathbb{R}^d_+}\|f(e^n\cdot)\zeta\|_{H^\gamma_p} \leq N|a|^{(0)}_{|\gamma|+,\mathbb{R}^d_+}|\zeta|^{(0)}_{|\gamma|+,\mathbb{R}^d_+}\|f(e^n\cdot)\zeta\|_{H^\gamma_p}.$$

Assertion (3.8) easily follows from these inequalities. Inequality (3.9) is straightforward due to Remark 2.13. The lemma is proved.

Now we give a result that will allow us to use change of variables.

LEMMA 3.7. *Let Ω', Ω'' be domains in $\mathbb{R}^{d'}$ and $\mathbb{R}^{d''}$, respectively, constants $N_1, N_2 \in [1,\infty)$, $\sigma \geq 0$, $\gamma = k+\varepsilon$, where $k = 0,1,\dots$ and $\varepsilon \in [0,1)$. Let a be a function on Ω' with $|a|^{(\sigma)}_{\gamma,\Omega'} < \infty$ and let $\mu : \Omega'' \to \Omega'$ be a Lipschitz continuous mapping with Lipschitz constant N_1 such that $\rho_{\Omega''} \leq N_2\rho_{\Omega'}(\mu)$ on Ω'' and $|\mu_x|^{(0)}_{(\gamma-1)_+,\Omega''} < \infty$. Then*

$$|a(\mu)|^{(\sigma)}_{\gamma,\Omega''} \leq N(\gamma,\sigma,d)N_3|a|^{(\sigma)}_{\gamma,\Omega'}, \tag{3.10}$$

where $N_3 = N_3(k,\varepsilon,\sigma) = N_2^{\gamma+\sigma}N_1^{\varepsilon}(1+|\mu_x|^{(0)}_{(\gamma-1)_+,\Omega''})^k$.

Proof. The result is trivial if $k = \varepsilon = 0$ since $\rho_{\Omega''} \leq N_2\rho_{\Omega'}(\mu)$ and $\sigma \geq 0$. If $k = 0$ and $\varepsilon \in (0,1)$, estimate (3.10) follows after observing that for $x, y \in \Omega''$ we have

$$\rho^{\varepsilon+\sigma}_{\Omega''}(x,y)\frac{|a(\mu(x)) - a(\mu(y))|}{|x-y|^\varepsilon} \leq N_2^{\gamma+\sigma}[a]^{(\sigma)}_{\gamma,\Omega'}\frac{|\mu(x)-\mu(y)|^\varepsilon}{|x-y|^\varepsilon} \leq N_2^{\gamma+\sigma}N_1^{\varepsilon}[a]^{(\sigma)}_{\gamma,\Omega'}.$$

We now use the induction on k. Assume that $k = n+1$ and (3.10) holds with $n+\varepsilon$ in place of γ, where $n = 0,1,2,\dots$ and $\varepsilon \in [0,1)$. Observe that

$$(a(\mu(x)))_{x^i} = a_{y^j}(\mu(x))\mu^j_{x^i}(x).$$

Then by the induction hypotheses and Lemma 3.4 we obtain

$$\begin{aligned}|(a(\mu))_x|^{(\sigma+1)}_{\gamma-1,\Omega''} &\leq N|(a_x(\mu))|^{(\sigma+1)}_{\gamma-1,\Omega''}|\mu_x|^{(0)}_{\gamma-1,\Omega''}\\ &\leq NN_3(k-1,\varepsilon,\sigma+1)|a_x|^{(\sigma+1)}_{\gamma-1,\Omega'}|\mu_x|^{(0)}_{\gamma-1,\Omega''} \leq NN_3(k,\varepsilon,\sigma)|a|^{(\sigma)}_{\gamma,\Omega'}.\end{aligned}$$

By using again Lemma 3.4 and the fact that $|a(\mu)|^{(\sigma)}_{0,\Omega''}$ admits the same estimate (recall that $N_1, N_2 \geq 1$), we see that (3.10) holds with $k = n+1$. The lemma is proved.

Proof of Lemma 2.8. (i) Owing to the first assertion of Lemma 3.4 we have $|\psi^{-\kappa}|^{(\sigma)}_{\tau} \leq N|\psi^{-\kappa}|^{(\sigma)}_{n}$, where n is any integer $\geq \tau$. Therefore, it suffices to concentrate on $\tau \in \{0,1,\dots\}$. Furthermore, since Ω is bounded and $\kappa \leq \sigma$, we have $|\psi^{-\kappa}|^{(\sigma)}_{\tau} \leq N|\psi^{-\kappa}|^{(\kappa)}_{\tau}$. Hence we may assume in addition that $\kappa = \sigma$.

Now first let $\sigma \geq 0$. In Lemma 3.7 take $\Omega' = \mathbb{R}_+$, $\Omega'' = \Omega$, and $\mu = \psi$. Then the assumption $\rho_{\Omega''} \leq N_2\rho_{\Omega'}(\mu)$ is satisfied since Ω is bounded, and it remains only to note that, obviously, $a(x) := x^{-\sigma}$, $x > 0$, satisfies $|a|^{(\sigma)}_{\tau,\Omega'} < \infty$ for any $\tau = 0,1,2,\dots$.

If $\sigma < 0$, assertion (i) follows by induction on τ on the basis of the case $\sigma \geq 0$ and the Leibniz rule:

$$\begin{aligned}0 = \psi^{|\gamma|}D^\gamma(\psi^\sigma\psi^{-\sigma}) &= \psi^{|\gamma|+\sigma}D^\gamma(\psi^{-\sigma})\\ &\quad + \sum_{\substack{\alpha+\beta=\gamma,\\ |\alpha|<|\gamma|}} c^\gamma_{\alpha\beta}[\psi^{|\alpha|+\sigma}D^\alpha(\psi^{-\sigma})]\psi^{|\beta|-\sigma}D^\beta(\psi^\sigma),\end{aligned}$$

where $|\gamma| \geq 1$ and $c^\gamma_{\alpha\beta}$ are certain constants.

To prove assertion (ii) notice that by (i) and the second assertion of Lemma 3.4 we have

$$|a|_\tau^{(\sigma)} = |\psi^{-\sigma}\psi^\sigma a|_\tau^{(\sigma)} \le N|\psi^{-\sigma}|_\tau^{(\sigma)}|\psi^\sigma a|_\tau^{(0)} \le N|\psi^\sigma a|_\tau^{(0)}.$$

Also

$$|\psi^\mu a|_\tau^{(\sigma)} \le N|\psi^\mu|_\tau^{(-\mu)}|a|_\tau^{(\mu+\sigma)} \le N|a|_\tau^{(\mu+\sigma)},$$

provided that $[\tau]-\mu \ge 0$ or $\tau \in \{0,1,2,\dots\}$. However, if none of these conditions holds, then $\mu = k\alpha$, where k is an integer ≥ 2 and $0 < \alpha \le [\tau]$. Then, for $\mu(i) = i\alpha$, we have

$$|\psi^{\mu(i+1)}a|_\tau^{(\sigma)} = |\psi^\alpha(\psi^{\mu(i)}a)|_\tau^{(\sigma)} \le N|\psi^{\mu(i)}a|_\tau^{(\alpha+\sigma)}$$

and the rest is obvious. The lemma is proved.

The following lemma about implicit functions will be used on few occasions.

LEMMA 3.8. *Let $n = 1,2,\dots$, $G \subset \mathbb{R}^d$ be a domain and let $d(x)$ be a nonnegative function on G. Let $E(r,x)$ be an $\mathbb{R}^{d_1}$-valued n times continuously differentiable function given in an open set of points $(r,x) \in \mathbb{R}^{d_1+d}$, $r \in \mathbb{R}^{d_1}$, $x \in \mathbb{R}^d$, whose projection on $\mathbb{R}^d$ is G. Assume that for each $x \in G$ there exists a unique solution $r(x)$ of the equation $E(r,x) = 0$. Denote $z(x) = (r(x),x)$ and assume that for $x \in G$ the matrix $E_r(z(x))$ is invertible and the inverse matrix is bounded on G. Finally, assume that*

$$d^{|\alpha|-1}(x)(D^\alpha E)(z(x)) \tag{3.11}$$

is bounded in G for any α such that $n \ge |\alpha| \ge 1$, where, as usual, D^α stands for the derivative of order α in all variables (on that occasion of function E depending on $z = (r,x)$). Then

(i) *it holds that*

$$d^{|\alpha|-1}(x)D^\alpha r(x) \tag{3.12}$$

is bounded in G for any α such that $n \ge |\alpha| \ge 1$;

(ii) *if the sets $\{x \in G : d(x) < \varepsilon\}$ are nonempty for any $\varepsilon > 0$ and if $n \ge 2$ and (3.11) tends to zero as $d(x) \to 0$ for any α such that $n \ge |\alpha| \ge 2$, then (3.12) tends to zero as $d(x) \to 0$ for any α such that $n \ge |\alpha| \ge 2$.*

Proof. (i) It is well known that $r(x)$ is n times continuously differentiable in G and

$$E_{r^j}(z(x))r^j_{x^i}(x) = -E_{x^i}(z(x)). \tag{3.13}$$

It follows that (3.12) is bounded for $|\alpha| = 1$. Assume that it is bounded for $1 \le |\alpha| \le m$, where $m \in \{1,2,\dots,n-1\}$. By differentiating (3.13) we find for $x \in G$ that

$$\begin{aligned} E_{r^j}(z(x))d^{|\alpha|}(x)D^\alpha r^j_{x^i}(x) &= -d^{|\alpha|}(x)D^\alpha\big(E_{x^i}(z(x))\big) \\ &\quad + \sum_{\substack{|\beta|\le|\alpha|-1,\\ |\beta|+|\gamma|=|\alpha|}} c^\alpha_{\beta\gamma}\big[d^{|\beta|}(x)D^\beta r^j_{x^i}(x)\big]d^{|\gamma|}(x)D^\gamma\big(E_{r^j}(z(x))\big), \end{aligned} \tag{3.14}$$

where $c^\alpha_{\beta\gamma}$ are some constants. Owing to the induction hypotheses, we see that to prove the boundedness of (3.12) for $|\alpha| = m+1$ it suffices to prove that $d^{|\alpha|}(x)D^\alpha\big(E_z(z(x))\big)$ is bounded whenever $|\alpha| \le m$.

Observe that $D^\alpha\big(E_z(z(x))\big)$ is represented as the sum of certain constants times terms of the type

$$(D^\beta E_z)(z(x))D^{\gamma_1}z^{i_1}(x)\times\cdots\times D^{\gamma_{|\beta|}}z^{i_{|\beta|}}(x), \tag{3.15}$$

where $|\gamma_i|\geq 1$, $|\gamma_1|+\cdots+|\gamma_{|\beta|}|=|\alpha|$ and z^i, $i=1,\dots,d_1+d$, is the ith coordinate of $z=(r,x)$. Being multiplied by $d^{|\alpha|}(x)$ expression (3.15) is written as

$$d^{|\beta|}(x)(D^\beta E_z)(z(x))d^{|\gamma_1|-1}(x)D^{\gamma_1}z^{i_1}(x)\times\cdots\times d^{\gamma_{|\beta|}-1}(x)D^{\gamma_{|\beta|}}z^{i_{|\beta|}}(x). \tag{3.16}$$

If $|\alpha|\leq m$, then $|\gamma_k|\leq m$ and $d^{|\gamma_k|-1}(x)D^{\gamma_k}z^{i_k}(x)$ is bounded in both cases if $z^{i_k}=r^j$ (by the induction hypotheses) or $z^{i_k}=x^j$ (trivially). Hence the boundedness of (3.11) implies that of (3.16) and finishes the proof of (i).

To prove (ii), first examine (3.14) for $|\alpha|=1$. The first term on the right goes to zero as $d(x)\to 0$ due to the assumption about (3.11) and the fact that r_x is bounded. The treatment of the second term is no different since for β there is only one possibility: $\beta=0$. Next, assume that (3.12) tends to zero as $d(x)\to 0$ for $|\alpha|=2,\dots,m$ ($m\leq n-1$) and, to make one step forward, take α in (3.14) with $|\alpha|=m$.

Notice that in the second term on the right in (3.14) the terms with $|\beta|\geq 1$ tend to zero as $d(x)\to 0$ by the induction hypotheses since $|\beta|+1\leq m$. It follows that to complete the induction it suffices to prove that $d^{|\alpha|}(x)D^\alpha\big(E_z(z(x))\big)\to 0$ if $|\alpha|=m$.

We go back to (3.16) and observe that if $2\leq|\gamma_k|(\leq m)$, then $d^{|\gamma_k|-1}(x)D^{\gamma_k}z^{i_k}(x)\to 0$ as $d(x)\to 0$ in both cases if $z^{i_k}=r^j$ (by the induction hypotheses) or $z^{i_k}=x^j$ (being identically zero). Taking into account assertion (i) we see that it remains only to analyze the terms of type (3.16) with $|\gamma_1|=\cdots=|\gamma_{|\beta|}|=1$. However, for those terms we have $|\beta|=|\alpha|\geq 1$ and they tend to zero by assumption. The lemma is proved.

4. Proof of Theorem 2.14. We closely follow the proof of Theorem 2.16 of [6]. As usual we may assume that $u_0=0$ and $T=\infty$ and, since for $a^{ij}=\delta^{ij}$ and $b=0$ and $c=0$ the result is known from [9], we need only prove the existence of an ω_0, such that the a priori estimate (2.11) holds given that the solution already exists and $\omega\leq\omega_0$. Below, unless explicitly expressed otherwise, we use notation N for various constants which may vary from one occurrence to another and depend only on the data as they should according to the statement of the theorem.

Case 1. $|\gamma|\notin\{0,1,2,\dots\}$. Take the least integer $q\geq|\gamma|+4$. Also take an $\varepsilon\in(0,1)$ to be specified later and take a sequence of functions η_k, $k=1,2,\dots$, from Lemma 3.3 corresponding to ε,q. Then by Lemma 3.1, we have

$$\|M^{-1}u\|^p_{\mathbb{H}^{\gamma+2}_{p,\theta}}\leq N\sum_{k=1}^\infty\|M^{-1}u\eta_k^2\|^p_{\mathbb{H}^{\gamma+2}_{p,\theta}}. \tag{4.1}$$

For any k let x_k be a point in $\operatorname{supp}\eta_k$ and $a_k(t)=a(t,x_k)$. Owing to (2.1), we have

$$(u\eta_k^2)_t=a_k^{ij}(u\eta_k^2)_{x^ix^j}+M^{-1}f_k,$$

where

$$\begin{aligned}f_k=&(a^{ij}-a_k^{ij})\eta_k^2Mu_{x^ix^j}-2a_k^{ij}M(\eta_k^2)_{x^i}u_{x^j}-a_k^{ij}M^{-1}uM^2(\eta_k^2)_{x^ix^j}\\&+\eta_k^2Mb^iu_{x^i}+\eta_k^2M^2cM^{-1}u+Mf\eta_k^2.\end{aligned}$$

It follows from [9] that for each k

$$\|M^{-1}u\eta_k^2\|^p_{\mathbb{H}^{\gamma+2}_{p,\theta}} \le N\|f_k\|^p_{\mathbb{H}^{\gamma}_{p,\theta}}. \tag{4.2}$$

Furthermore, by Lemmas 3.6 and 3.5 after denoting $\gamma' = |\gamma| + \delta_0(|\gamma|)/2$ and $\delta_1 = \delta_0(|\gamma|)/(2|\gamma| + 2\delta_0(|\gamma|))$ (> 0), we get

$$\begin{aligned}\|(a^{ij} - a_k^{ij})\eta_k^2 M u_{x^ix^j}\|_{\mathbb{H}^{\gamma}_{p,\theta}} &\le N\|\eta_k M u_{x^ix^j}\|_{\mathbb{H}^{\gamma}_{p,\theta}} \sup_{t\ge 0} |(a^{ij} - a_k^{ij})(t,\cdot)\eta_k|^{(0)}_{\gamma',\mathbb{R}^d_+}\\ &\le N\|\eta_k M u_{x^ix^j}\|_{\mathbb{H}^{\gamma}_{p,\theta}} \sup_{[0,\infty)\times\mathbb{R}^d_+} |(a^{ij} - a_k^{ij})\eta_k|^{\delta_1}.\end{aligned}$$

Observe that for any k and $x, y \in \operatorname{supp}\eta_k$ we have $|x - y| \le N(\varepsilon)(x^1 \wedge y^1)$, where $N(\varepsilon) = N(d, q, \varepsilon)$, and one can easily find not more than $N(\varepsilon) + 2 \le 3N(\varepsilon)$ points x_i lying on the straight segment connecting x and y and including x and y, such that $|x_i - x_{i+1}| \le x_i^1 \wedge x_{i+1}^1$. It follows from our assumptions that

$$\sup_{[0,\infty)\times\mathbb{R}^d_+} |(a^{ij} - a_k^{ij})\eta_k| \le NN(\varepsilon)\omega.$$

Similarly,

$$\|\eta_k^2 M b^i u_{x^i}\|_{\mathbb{H}^{\gamma}_{p,\theta}} + \|\eta_k^2 M^2 c M^{-1}u\|_{\mathbb{H}^{\gamma}_{p,\theta}} \le NN(\varepsilon)\omega^{\delta_1}\left(\|\eta_k u_x\|_{\mathbb{H}^{\gamma}_{p,\theta}} + \|\eta_k M^{-1}u\|_{\mathbb{H}^{\gamma}_{p,\theta}}\right).$$

Coming back to (4.2) and (4.1) and using Lemma 3.1, we conclude

$$\begin{aligned}\|M^{-1}u\|^p_{\mathbb{H}^{\gamma+2}_{p,\theta}} &\le NN(\varepsilon)\omega^{p\delta_1}\left(\|Mu_{xx}\|^p_{\mathbb{H}^{\gamma}_{p,\theta}} + \|u_x\|^p_{\mathbb{H}^{\gamma}_{p,\theta}} + \|M^{-1}u\|^p_{\mathbb{H}^{\gamma}_{p,\theta}}\right)\\ &\quad + NC^p\left(\|u_x\|^p_{\mathbb{H}^{\gamma}_{p,\theta}} + \|M^{-1}u\|^p_{\mathbb{H}^{\gamma}_{p,\theta}}\right) + N\|Mf\|^p_{\mathbb{H}^{\gamma}_{p,\theta}},\end{aligned} \tag{4.3}$$

where

$$C = \sup_{\mathbb{R}^d_+}\sup_{|\alpha|\le q-2}\sum_{k=1}^{\infty} M^{|\alpha|}(|D^\alpha(M(\eta_k^2)_x)| + |D^\alpha(M^2(\eta_k^2)_{xx})|).$$

By construction, we have $C \le N\varepsilon$. Furthermore (see, for instance, [9]),

$$\|u_x\|_{H^{\gamma}_{p,\theta}} \le N\|M^{-1}u\|_{H^{\gamma+1}_{p,\theta}}, \quad \|Mu_{xx}\|_{H^{\gamma}_{p,\theta}} \le N\|M^{-1}u\|_{H^{\gamma+2}_{p,\theta}}. \tag{4.4}$$

Hence (4.3) yields

$$\|M^{-1}u\|^p_{\mathbb{H}^{\gamma+2}_{p,\theta}} \le N_1(N(\varepsilon)\omega^{p\delta_1} + \varepsilon^p)\|M^{-1}u\|^p_{\mathbb{H}^{\gamma+2}_{p,\theta}} + N\|Mf\|^p_{\mathbb{H}^{\gamma}_{p,\theta}}.$$

We finally choose first ε and then ω_0 so that $N_1(N(\varepsilon)\omega^{p\delta_1} + \varepsilon^p) \le 1/2$ for $\omega \le \omega_0$ and finish the proof of the theorem in the case under consideration.

Case 2. $\gamma \in \{0, 1, 2, \ldots\}$. If $\gamma = 0$, (4.3) obviously holds with $\delta_1 = 1$ and C defined by the same formula in which we drop the supremum with respect to α and take $\alpha = 0$. After this one can follow the previous arguments word for word. If γ is a positive integer, one can proceed as in [6] by induction on γ on the basis of (3.9). We leave the details to the reader.

Case 3. $\gamma \in \{-1, -2, \dots\}$. In this case instead of proving a priori estimates we prove the theorem directly. As above we may assume that $u_0 = 0$.

We proceed by induction on γ and assume that there exists an $\omega_0 > 0$ such that the theorem holds for $\gamma + 1$ in place of γ. We will see that the same ω_0 suits γ. The possibility to start the induction from $\gamma = -1$ is justified by the above result.

Let $\omega \leq \omega_0$. Then the operator $\mathcal{R}$ which maps $f \in M^{-1}\mathbb{H}^{\gamma+1}_{p,\theta}$ into the solution $u \in \mathfrak{H}^{\gamma+3}_{p,\theta,0}$ of (2.1) is well defined and bounded.

Take an $f \in M^{-1}\mathbb{H}^{\gamma}_{p,\theta}$ and recall that $d - 1 < \theta < d - 1 + p$, so that according to Corollary 2.12 of [9] we have the representation

$$f = MD_k f^k,$$

where $f^k \in M^{-1}\mathbb{H}^{\gamma+1}_{p,\theta}$, $k = 1, 2, \dots, d$, and

$$\sum_{k=1}^{d} \|Mf^k\|_{\mathbb{H}^{\gamma+1}_{p,\theta}} \leq N\|Mf\|_{\mathbb{H}^{\gamma}_{p,\theta}}. \tag{4.5}$$

Now let

$$w^k = \mathcal{R}f^k, k = 1, 2, \dots, d, \quad v = MD_k w^k.$$

Owing to the induction hypothesis, (4.4), and (4.5) we have

$$\|M^{-1}v\|_{\mathbb{H}^{\gamma+2}_{p,\theta}} \leq \sum_k \|w^k_x\|_{\mathbb{H}^{\gamma+2}_{p,\theta}} \leq N \sum_k \|M^{-1}w^k\|_{\mathbb{H}^{\gamma+3}_{p,\theta}} \leq N\|Mf\|_{\mathbb{H}^{\gamma}_{p,\theta}}.$$

Furthermore, as is easy to check,

$$v_t = a^{ij}v_{x^i x^j} + b^i v_{x^i} + cv + f + \bar{f}$$

with

$$\begin{aligned} M\bar{f} = {} & Mw^k_{x^i x^j} MD_k a^{ij} + w^k_{x^i} M^2 D_k b^i + M^{-1}w^k M^3 D_k c \\ & - 2a^{i1} M w^k_{x^k x^i} - w^k_{x^k} M b^1. \end{aligned}$$

In addition,

$$|MD_k a|_{|\gamma+1|,\mathbb{R}^d_+} = |MD_k a|_{|\gamma|-1,\mathbb{R}^d_+} \leq N|a|_{|\gamma|,\mathbb{R}^d_+} \leq NK,$$

$$|M^2 D_k b|_{|\gamma+1|,\mathbb{R}^d_+} = |M^2 D_k b|_{|\gamma|-1,\mathbb{R}^d_+} \leq N|b|^{(1)}_{|\gamma|,\mathbb{R}^d_+},$$

and similar estimates hold for $M^3 D_k c$, a, and Mb. Hence from the construction of w^k, (4.4), Lemma 3.6, and (4.5) we infer that

$$\|M\bar{f}\|_{\mathbb{H}^{\gamma+1}_{p,\theta}} \leq N\|Mf\|_{\mathbb{H}^{\gamma}_{p,\theta}}.$$

Finally, we can define $\bar{u} = \mathcal{R}(\bar{f})$ and $u = v - \bar{u}$. Then u belongs to $\mathfrak{H}^{\gamma+2}_{p,\theta,0}$ and satisfies (2.1), and (2.11) follows from the above estimates.

To prove the uniqueness of solutions take an $u \in \mathfrak{H}^{\gamma+2}_{p,\theta,0}$ and assume that it satisfies (2.1) in $(0, \infty) \times \mathbb{R}^d_+$ with $f = 0$. Notice that

$$(\eta_k u)_t = a^{ij}(\eta_k u)_{x^i x^j} + b^i(\eta_k u)_{x^i} + c(\eta_k u) + \tilde{f}, \tag{4.6}$$

where

$$\tilde{f} = -2a^{ij}\eta_{kx^i}u_{x^j} - (a^{ij}\eta_{kx^ix^j} + b^i\eta_{kx^i})u.$$

As is easy to see, $\tilde{f} \in L_p((0,\infty), H_p^{\gamma+1}) =: \mathbb{H}_p^{\gamma+1}$. (Here we use the notation from [8].) Furthermore, (4.6) will not change if we change arbitrarily a, b, c outside of the support of η_k. We do this preserving the uniform ellipticity and smoothness of a, b, c in the whole space, and then by a well-known regularity result we get that (4.6) about $\eta_k u$ is uniquely solvable in $\mathbb{H}_p^{\gamma+2}$ for any $\tilde{f} \in \mathbb{H}_p^{\gamma}$ and also uniquely solvable in $\mathbb{H}_p^{\gamma+3}$ for any $\tilde{f} \in \mathbb{H}_p^{\gamma+1}$. Actually, it is hard to find an exact reference to this "well-known" result, but refer to Remark 5.6 of [8], where one must throw away all stochastic terms and then notice that one can apply this remark for all $p \in (1,\infty)$ rather than $p \in [2,\infty)$. The latter assumption on p in [8] is only related to the presence of stochastic terms.

From the uniqueness in $\mathbb{H}_p^{\gamma+2}$ and the solvability in $\mathbb{H}_p^{\gamma+3}$ we conclude that $\eta_k u \in \mathbb{H}_p^{\gamma+3}$. Since η_k has compact support, we also have $\eta_k u \in \mathbb{H}_{p,\theta}^{\gamma+3}$, $M^{-1}\eta_k u \in \mathbb{H}_{p,\theta}^{\gamma+3}$, and $\eta_k u \in \mathfrak{H}_{p,\theta,0}^{\gamma+3}$. This and the induction hypotheses allow us to get from (4.6) that

$$\|\eta_k M^{-1}u\|_{\mathbb{H}_{p,\theta}^{\gamma+3}}^p \le N\|M\eta_{kx^i}a^{ij}u_{x^j}\|_{\mathbb{H}_{p,\theta}^{\gamma+1}}^p + N\|(a^{ij}M^2\eta_{kx^ix^j} + Mb^iM\eta_{kx^i})M^{-1}u\|_{\mathbb{H}_{p,\theta}^{\gamma+1}}^p.$$

By summing up these estimates with respect to k and using Lemmas 3.1, 3.4, and 3.6 and the fact that $M^{-1}u \in \mathbb{H}_{p,\theta}^{\gamma+2}$, we obtain that $\|M^{-1}u\|_{\mathbb{H}_{p,\theta}^{\gamma+3}} < \infty$, that is, $u \in \mathfrak{H}_{p,\theta,0}^{\gamma+3}$. That $u = 0$ now follows from the induction hypotheses.

The theorem is proved.

5. Proof of Theorems 2.10 and 2.11.

Proof of Theorem 2.10. We split the proof into three steps.

Step 1. First we claim that we may assume that $\partial\Omega$ is infinitely differentiable. To prove the claim, use Theorem 2.12 and notice that, as we know from Theorem 3.2 of [15], due to assertions (i) and (ii) of Theorem 2.12, the mappings μ and ν induce one-to-one linear bounded mappings of the spaces $H_{p,\theta}^{\gamma}(\Omega)$ onto $H_{p,\theta}^{\gamma}(\Omega_\varepsilon)$ and vice versa. Therefore, proving that a function $u \in \mathfrak{H}_{p,\theta}^{\gamma+2}(\Omega,T)$ satisfies (2.1) with initial condition u_0 and admits estimate (2.8) is equivalent to proving that the function $\tilde{u} := u(\mu)$ satisfies the corresponding equation in $(0,T)\times\Omega_\varepsilon$, belongs to $\mathfrak{H}_{p,\theta}^{\gamma+2}(\Omega_\varepsilon,T)$, and admits the natural modification of estimate (2.8). The equation for $\tilde{u}$ is

$$\tilde{u}_t = \tilde{a}^{ij}\tilde{u}_{x^ix^j} + \tilde{b}^i\tilde{u}_{x^i} + \tilde{c}\tilde{u} + \tilde{f}, \tag{5.1}$$

where

$$\tilde{a}^{ij}(t,x) = \bar{a}^{ij}(t,\mu(x)), \quad \tilde{b}^i(t,x) = \bar{b}^i(t,\mu(x)),$$
$$\bar{a}^{ij} = a^{kr}\nu_{x^k}^i\nu_{x^r}^j, \quad \bar{b}^i = a^{kr}\nu_{x^kx^r}^i + b^k\nu_{x^k}^i,$$
$$\tilde{c}(t,x) = c(t,\mu(x)), \quad \tilde{f}(t,x) = f(t,\mu(x)).$$

Since the matrix $\nu_x(\mu)$ and its inverse μ_x are bounded, (5.1) is uniformly parabolic.

Furthermore, by Lemma 3.4,

$$|\bar{a}(t,\cdot)|^{(0)}_{|\gamma|+} \le N|a(t,\cdot)|^{(0)}_{|\gamma|+}(|\nu_x|^{(0)}_{|\gamma|+})^2 \le N_1,$$
$$|\bar{b}(t,\cdot)|^{(1)}_{|\gamma|+} \le N|a(t,\cdot)|^{(0)}_{|\gamma|+}|\nu_{xx}|^{(1)}_{|\gamma|+} + N|b(t,\cdot)|^{(1)}_{|\gamma|+}|\nu_x|^{(0)}_{|\gamma|+} \le N_1,$$

where N_1 is independent of t. Also observe that in a bounded C^1 domain every function having bounded first derivatives satisfies the Lipschitz condition. Then from the above estimates by Lemma 3.7 we conclude that $\tilde{a}, \tilde{b}, \tilde{c}$ satisfy Assumption 2.3 relative to Ω_ε with certain strictly positive and finite constants in place of δ and K.

The function $\tilde{a}(t,x)$ is obviously continuous inside of Ω_ε uniformly in t, so that Assumption 2.4(i) is satisfied. Also, obviously the part of Assumption 2.4(ii) concerning $\tilde{c}$ is satisfied. It is satisfied for $\tilde{b}$ as well since (cf. Theorem 2.12(iv)) $\rho_\Omega(x)(|b(t,x)| + |\nu_{xx}(x)|) \to 0$ as long as $t > 0$ and $\Omega \ni x \to \partial\Omega$, which implies that, under the same conditions, $\rho_\Omega|\bar{b}| \to 0$ and if $x \in \Omega_\varepsilon$ and $t > 0$ and $\rho_{\Omega_\varepsilon}(x) \to 0$, then $\rho_{\Omega_\varepsilon}(x)|\tilde{b}(t,x)| \to 0$.

It remains to check (2.4) for $\tilde{a}$. This turns out to be a more tedious albeit elementary task. Observe that for $x, y \in \Omega_\varepsilon$,

$$\begin{aligned}|\tilde{a}(t,x) - \tilde{a}(t,y)| &= |\bar{a}(t,\mu(x)) - \bar{a}(t,\mu(y))| \\ &\le N|a(t,\mu(x)) - a(t,\mu(y))| + N|\nu_x(\mu(x)) - \nu_x(\mu(y))|.\end{aligned}$$

Here the last term goes to zero as $x - y \to 0$ due to the uniform continuity of μ and ν_x no matter whether x, y approach $\partial\Omega_\varepsilon$ or not. Furthermore, if $x, y \in \Omega_\varepsilon$ and $|x - y| \le \rho_{\Omega_\varepsilon}(x,y)$, then

$$|\mu(x) - \mu(y)| \le N|x - y| \le N\rho_{\Omega_\varepsilon}(x,y) \le N\rho_\Omega(\mu(x),\mu(y)).$$

Therefore, to check that $|\tilde{a}(t,x) - \tilde{a}(t,y)| \to 0$ uniformly in t as $\Omega_\varepsilon \ni x \to \partial\Omega_\varepsilon$ and $|x - y| \le \rho_{\Omega_\varepsilon}(x,y)$ it suffices to prove that, for any constant N_1,

$$\lim_{\substack{\rho(x)\to 0,\\ x\in\Omega}} \sup_{\substack{y\in\Omega,\\ |x-y|\le N_1\rho(x,y)}} \sup_t |a(t,x) - a(t,y)| = 0, \tag{5.2}$$

which is given for $N_1 = 1$ by Assumption 2.4.

By way of getting a contradiction assume that (5.2) is false. Then there exists a point $x_0 \in \partial\Omega$, a $\tau > 0$, and sequences $x_n, y_n \in \Omega$ and $t_n > 0$ such that $x_n, y_n \to x_0$, $|x_n - y_n| \le N_1\rho(x_n,y_n)$, and

$$|a(t_n,x_n) - a(t_n,y_n)| \ge \tau.$$

Take the mapping Ψ from Assumption 2.1 and a number $k = 1, 2, \dots$ to be specified later and define

$$\bar{x}_n(i) = \Psi(x_n)i/k + \Psi(y_n)(k-i)/k, \quad x_n(i) = \Psi^{-1}(\bar{x}_n(i)), \quad i = 0, 1, \dots, k.$$

As is easy to see, due to Assumption 2.1, we have $B_{r_0/K_0} \cap \mathbb{R}^d_+ \subset \Psi(B_{r_0}(x_0) \cap \Omega)$. Therefore, for all large n, which we only concentrate on, we have $x_n(i) \in B_{r_0}(x_0) \cap \Omega$. Furthermore, for $x \in \partial\Omega$ close to x_0, we have

$$\bar{x}^1_n(i) \le |\bar{x}_n(i) - \Psi(x)| \le K_0|x_n(i) - x|.$$

By taking the inf over $x \in B_{r_0}(x_0) \cap \partial\Omega$, we get that $\bar{x}_n^1(i) \le K_0 \rho(x_n(i))$. Similarly, if $z \in \partial\mathbb{R}^d_+$ and z is close to 0, then $\rho(x_n(i)) \le |x_n(i) - \Psi^{-1}(z)| \le K_0 |\bar{x}_n(i) - z|$, which after taking the inf over z yields $\rho(x_n(i)) \le K_0 \bar{x}_n^1(i)$.

Now, notice that the sequence $\bar{x}_n^1(i)$ is monotone in i so that, for $i = 0, 1, \dots, k-1$,

$$\bar{x}_n^1(k) \wedge \bar{x}_n^1(0) \le \bar{x}_n^1(i+1) \wedge \bar{x}_n^1(i).$$

It follows that, for $i = 0, 1, \dots, k-1$,

$$\begin{aligned}
|x_n(i+1) - x_n(i)| &\le K_0 |\bar{x}_n(i+1) - \bar{x}_n(i)| = K_0 k^{-1} |\Psi(x_n) - \Psi(y_n)| \\
&\le K_0^2 k^{-1} |x_n - y_n| \le K_0^2 k^{-1} N_1 \rho(x_n(k), x_n(0)) \\
&\le K_0^3 k^{-1} N_1 [\bar{x}_n^1(k) \wedge \bar{x}_n^1(0)] \le K_0^3 k^{-1} N_1 [\bar{x}_n^1(i+1) \wedge \bar{x}_n^1(i)] \\
&\le K_0^4 k^{-1} N_1 \rho(x_n(i+1), x_n(i)).
\end{aligned}$$

The latter is less than $\rho(x_n(i+1), x_n(i))$ if $k \ge K_0^4 N_1$. With such a k by Assumption 2.4 we conclude

$$|a(t_n, y_n) - a(t_n, x_n)| \le \sum_{i=0}^{k-1} |a(t_n, x_n(i+1)) - a(t_n, x_n(i+1))| \to 0,$$

which is the contradiction in question. Thus indeed in the rest of the proof we may assume that $\partial\Omega$ is infinitely differentiable.

Step 2. We establish a priori estimate (2.8) assuming that $u \in \mathfrak{H}^{\gamma+2}_{p,\theta}(\Omega, T)$ satisfies (2.1). Let $x_0 \in \partial\Omega$ and Ψ be a function from Assumption 2.1. By Step 1 we may assume that Ψ is infinitely differentiable with bounded derivatives.

Define $r = r_0/K_0$ and fix smooth functions $\eta \in C_0^\infty(B_r), \varphi \in C^\infty(\mathbb{R})$ such that $\eta = 1$ in $B_{r/2}$ and $\varphi(t) = 1$ for $t \le -3$ and $\varphi(t) = 0$ for $t \ge -1$ and $0 \ge \varphi' \ge -1$. As we noted above, $\Psi(B_{r_0}(x_0))$ contains B_r. For $k = 1, 2, \dots$, $t > 0$, $x \in \mathbb{R}^d_+$ introduce $\varphi_k(x) = \varphi(k^{-1} \ln x^1)$,

$$\hat{a}_k := \tilde{a}\eta(x)\varphi_k + (1 - \eta\varphi_k)I, \quad \hat{b}_k := \tilde{b}\eta\varphi_k, \quad \hat{c}_k := \tilde{c}\eta\varphi_k,$$

where the functions $\tilde{a}, \tilde{b}$, and $\tilde{c}$ are taken from (5.1) with Ψ and Ψ^{-1} instead of ν and μ, respectively. By using Lemma 3.4 one can check that $\hat{a}_k, \hat{b}_k, \hat{c}_k$ satisfy Assumptions 2.3 with $\Omega = \mathbb{R}^d_+$ and some new constant $\delta', K' \in (0, \infty)$ independent of k and x_0.

Take the ω_0 from Theorem 2.14 corresponding to $\delta', p, \theta, \gamma, |\gamma|+$, and K'. We also fix a $k > 0$ such that

$$|\hat{a}_k(t,x) - \hat{a}_k(t,y)| + x^1 |\hat{b}_k(t,x)| + (x^1)^2 |\hat{c}_k(t,x)| \le \omega_0$$

whenever $t > 0, x, y \in \mathbb{R}^d_+$ and $|x - y| \le x^1 \wedge y^1$. The fact that this condition holds with $\tilde{a}, \tilde{b}, \tilde{c}$ in place of $\hat{a}_k, \hat{b}_k, \hat{c}_k$ if x^1 and y^1 are small enough is proved in Step 1. That multiplying by φ_k preserves the needed property for small x^1 and y^1 and also extends it for all x^1 and y^1 follows from the fact that $\varphi(k^{-1} \ln x^1) = 0$ for $x^1 \ge e^{-k}$ and

$$|\varphi(k^{-1} \ln x^1) - \varphi(k^{-1} \ln y^1)| = k^{-1} \xi^{-1} |\varphi'(k^{-1} \ln \xi)| \cdot |x^1 - y^1| \le k^{-1},$$

where ξ is a point between x^1 and y^1, so that $|x^1 - y^1| \le x^1 \wedge y^1 \le \xi$. Now we fix a $\rho_0 < r_0$ such that

$$\Psi(B_{\rho_0}(x_0)) \subset B_{r/2} \cap \{x : x^1 \le e^{-3k}\}.$$

Let ζ be a smooth function with support in $B_{\rho_0}(x_0)$ and denote $v := (u\zeta)(\Psi^{-1})$ and continue v as zero in $\mathbb{R}^d_+ \setminus \Psi(B_{\rho_0}(x_0))$. Since $\eta\varphi_k = 1$ on $\Psi(B_{\rho_0}(x_0))$, the function v satisfies

$$v_t = \hat{a}^{ij}_k v_{x^i x^j} + \hat{b}^i_k v_{x^i} + \hat{c}_k v + \hat{f},$$

where $\hat{f} = \tilde{f}(\Psi^{-1})$, $\tilde{f} = -2a^{ij}u_{x^i}\zeta_{x^j} - ua^{ij}\zeta_{x^i x^j} - ub^i\zeta_{x^i} + \zeta f$. Similar to what was said in the beginning of Step 1, we have $v \in \mathfrak{H}^{\gamma+2}_{p,\theta}(T)$. It follows from Theorem 2.14 that for any $t \le T$,

$$\|M^{-1}v\|_{\mathbb{H}^{\gamma+2}_{p,\theta}(t)} \le N\|M\hat{f}\|_{\mathbb{H}^{\gamma}_{p,\theta}(t)} + N\|u_0(\Psi^{-1})\zeta(\Psi^{-1})\|_{U^{\gamma+2}_{p,\theta}}.$$

To transform this estimate we observe that by Theorem 3.2 in [15], for any $\nu, \alpha \in \mathbb{R}$ and $g \in \psi^{-\alpha}H^{\nu}_{p,\theta}(\Omega)$ with support in $B_{\rho_0}(x_0)$,

$$\|\psi^{\alpha}g\|_{H^{\nu}_{p,\theta}(\Omega)} \sim \|M^{\alpha}g(\Psi^{-1})\|_{H^{\nu}_{p,\theta}}.$$

Then we find

$$\begin{aligned}\|\psi^{-1}u\zeta\|_{\mathbb{H}^{\gamma+2}_{p,\theta}(\Omega,t)} &\le N\|M^{-1}v\|_{\mathbb{H}^{\gamma+2}_{p,\theta}(t)} \le N\|a\zeta_x\psi u_x\|_{\mathbb{H}^{\gamma}_{p,\theta}(\Omega,t)}\\ &+ N\|a\zeta_{xx}\psi u\|_{\mathbb{H}^{\gamma}_{p,\theta}(\Omega,t)} + \|\zeta_x\psi bu\|_{\mathbb{H}^{\gamma}_{p,\theta}(\Omega,t)}\\ &+ N\|\zeta\psi f\|_{\mathbb{H}^{\gamma}_{p,\theta}(\Omega,t)} + \|\zeta u_0\|_{U^{\gamma+2}_{p,\theta}(\Omega)}.\end{aligned}$$

Next we use a natural counterpart of Lemma 3.6 for general domains, which is stated as Theorem 3.1 in [15]. We also use that, by Lemma 2.8, Assumption 2.3(iii) implies that $|\psi b(t,\cdot)|^{(0)}_{|\gamma|+}$ is bounded on $[0,T]$. Then we conclude

$$\begin{aligned}\|\psi^{-1}u\zeta\|_{\mathbb{H}^{\gamma+2}_{p,\theta}(\Omega,t)} &\le N\|\psi u_x\|_{\mathbb{H}^{\gamma}_{p,\theta}(\Omega,t)} + N\|u\|_{\mathbb{H}^{\gamma}_{p,\theta}(\Omega,t)}\\ &+ N\|\psi f\|_{\mathbb{H}^{\gamma}_{p,\theta}(\Omega,t)} + N\|u_0\|_{U^{\gamma+2}_{p,\theta}(\Omega)}.\end{aligned}$$

Note that the above constants ρ_0, k, δ', K', and N are independent of x_0. Therefore, to estimate the norm $\|\psi^{-1}u\|_{\mathbb{H}^{\gamma+2}_{p,\theta}(\Omega,t)}$, one introduces a partition of unity $\zeta_{(i)}, i = 0,1,2,\dots,N$, such that $\zeta_{(0)} \in C_0^{\infty}(\Omega)$ and $\zeta_{(i)} \in C_0^{\infty}(B_{\rho_0}(x_i))$, $x_i \in \partial\Omega$, for $i \ge 1$. Then one estimates $\|\psi^{-1}u\zeta_{(0)}\|_{\mathbb{H}^{\gamma+2}_{p,\theta}(\Omega,t)}$ using Theorem 5.1 in [8] and the other norms as above. By summing up those estimates one gets

$$\begin{aligned}\|\psi^{-1}u\|_{\mathbb{H}^{\gamma+2}_{p,\theta}(\Omega,t)} &\le N\|\psi u_x\|_{\mathbb{H}^{\gamma}_{p,\theta}(\Omega,t)} + N\|u\|_{\mathbb{H}^{\gamma}_{p,\theta}(\Omega,t)}\\ &+ N\|\psi f\|_{\mathbb{H}^{\gamma}_{p,\theta}(\Omega,t)} + N\|u_0\|_{U^{\gamma+2}_{p,\theta}(\Omega)}.\end{aligned} \tag{5.3}$$

Furthermore, we know from Theorem 4.1 of [15] (cf. also (4.4)) that

$$\|\psi u_x\|_{H^{\gamma}_{p,\theta}(\Omega)} \le N\|u\|_{H^{\gamma+1}_{p,\theta}(\Omega)}.$$

Therefore (5.3) yields

$$\|u\|^p_{\mathfrak{H}^{\gamma+2}_{p,\theta}(\Omega,t)} \le N\|u\|^p_{\mathbb{H}^{\gamma+1}_{p,\theta}(\Omega,t)} + N\|\psi f\|^p_{\mathbb{H}^{\gamma}_{p,\theta}(\Omega,t)} + N\|u_0\|^p_{U^{\gamma+2}_{p,\theta}(\Omega)}.$$

Now (2.8) follows from inequality (2.21) of [16] and Gronwall's inequality. Actually, there is a restriction that $p \ge 2$ in inequality (2.21) of [16], but by inspecting the proofs

of Theorem 4.2 and Theorem 7.1 in [8] one can easily check that in our (deterministic) case the result holds for all $p > 1$.

Step 3. The a priori estimate from Step 2 combined with the method of continuity shows that it remains only to prove solvability in the case of the heat equation. At this moment the fact that the domain is infinitely smooth turns out to be extremely handy.

Since $C_0^\infty(\Omega)$ is dense in $U^{\gamma+2}_{p,\theta}(\Omega)$, it suffices to concentrate on $u_0 \in C_0^\infty(\Omega)$. Then passing from u to $u - u_0$ we see that we may assume $u_0 = 0$. Again using the fact that $C_0^\infty(\Omega)$ is dense in the spaces $H^\kappa_{q,\tau}(\Omega)$ we easily convince ourselves that it suffices to consider only f's that are bounded on $\Omega \times [0,T]$ along with each derivative in (t,x) and vanish if x is in a neighborhood of the boundary of Ω.

In that case it is well known (see, for instance, Theorem 4.5.2 in [10]) that there exists a classical solution u of the heat equation with zero boundary and initial data. Moreover, u is infinitely differentiable and each of its derivatives in (t,x) is bounded.

Next, it turns out that u/ψ is infinitely differentiable and has bounded derivatives. Indeed, this is a local property which is preserved under C^∞ transformations of coordinates. Moreover, inside of Ω the property is obvious and near the boundary points it follows after flattening the boundary from the formula

$$v(x)/x^1 = \int_0^1 v_{x^1}(rx^1, x^2, \dots, x^d)\, dr,$$

valid for any smooth function v on $\bar{\mathbb{R}}^d_+$ vanishing for $x^1 = 0$. In particular, we infer that $\psi^{|\alpha|} D^\alpha(\psi^{-1}u)$ is bounded for any multi-index α. Hence, by Proposition 2.2 in [15] we conclude that $u \in \mathfrak{H}^{\gamma+2}_{p,\theta}(\Omega, T)$. The theorem is proved.

Proof of Theorem 2.11. Let $u \in \psi H^{\gamma+2}_{p,\theta}(\Omega)$ be a solution of (2.9). Observe that, due to Theorem 3.1 of [15] and Lemma 2.8(i) implying that $|\psi^{2/p}|^{(0)}_\tau < \infty$ for all $\tau \geq 0$, we have $u \in U^{\gamma+2}_{p,\theta}(\Omega)$. Furthermore, $v := ue^{c_0 t}$ satisfies (2.1) with $fe^{c_0 t}$ in place of f. For v estimate (2.8) becomes

$$g(T)\|u/\psi\|_{H^{\gamma+2}_{p,\theta}(\Omega)} \leq N_1 e^{N_1 T}\left(\|u/\psi\|_{H^{\gamma+2}_{p,\theta}(\Omega)} + g(T)\|\psi f\|_{H^{\gamma}_{p,\theta}(\Omega)}\right),$$

where

$$g(T) = \left(\int_0^T e^{ptc_0}\, dt\right)^{1/p}.$$

If $c_0 > N_1$, then the ratio $N_1 e^{N_1 T}/g(T)$ tends to zero as $T \to \infty$. Then after finding a T such that this ratio is less than $1/2$ one gets (2.10).

Having thus proved the a priori estimate (2.10), we can proceed as in Steps 1 and 3 of the above proof of Theorem 2.10. The theorem is proved.

6. Proof of Theorem 2.12. First we discuss Lemma 2.6. Its assertions (i)–(iii) are stated as Lemma 2.8 in [4] and one finds all the assertions for $|\alpha| = 1$ in Theorems 1.3 and 2.1 of [12]. Since the lemma plays a crucial role in the present article, we give a short proof.

Lemma 2.8 of [4] is obtained on the basis of Lemma 2.3 of [4], and assertion (iv) of our Lemma 2.6 is obtained by analyzing the proof of Lemma 2.3 of [4], which treats a generalization of the following. Given a C^1 function f on $\mathbb{R}^{d-1}$ with compact support

and a $\varphi \in C_0^\infty(\mathbb{R}^{d-1})$, for $x = (x^1, x') \in \mathbb{R}^d_+$, introduce

$$F(x) = (x^1)^{1-d} \int_{\mathbb{R}^{d-1}} f(y')\varphi((x'-y')/x^1)\,dy' = \int_{\mathbb{R}^{d-1}} f(x'-x^1y')\varphi(y')\,dy'.$$

We have

$$F_{x^j}(x) = \int_{\mathbb{R}^{d-1}} f_{y^j}(x'-x^1y')\varphi(y')\,dy', \quad j = 2,\dots,d, \tag{6.1}$$

$$F_{x^1}(x) = -\int_{\mathbb{R}^{d-1}} f_{y^i}(x'-x^1y')y^i\varphi(y')\,dy', \tag{6.2}$$

$$|F_x(x)| \le N \sup_{y'} |f_{y'}(y')|. \tag{6.3}$$

Furthermore, by induction one easily gets that for any multi-index $\alpha \neq 0$

$$\begin{aligned}(MD)^\alpha F(x) &:= (MD_1)^{\alpha_1} \times \cdots \times (MD_d)^{\alpha_d} F(x)\\ &= (x^1)^{1-d} \int_{\mathbb{R}^{d-1}} f(y')\varphi_\alpha((x'-y')/x^1)\,dy' = \int_{\mathbb{R}^{d-1}} f(x'-x^1y')\varphi_\alpha(y')\,dy',\end{aligned}$$

where $\varphi_\alpha \in C_0^\infty(\mathbb{R}^{d-1})$ and $\int \varphi_\alpha\,dy' = 0$. Applying this to (6.1) and (6.2) we obtain

$$(MD)^\alpha F_{x^i}(x) = \int_{\mathbb{R}^{d-1}} f_{y^j}(x'-x^1y')\varphi_{ij\alpha}(y')\,dy',$$

where $\varphi_{ij\alpha} \in C_0^\infty(\mathbb{R}^{d-1})$ and $\int \varphi_{ij\alpha}\,dy' = 0$ if $\alpha \neq 0$. It follows that if $\alpha \neq 0$, then

$$\begin{aligned}(MD)^\alpha F_{x^i}(x) &= \int_{\mathbb{R}^{d-1}} [f_{y^j}(x'-x^1y') - f_{y^j}(x')]\varphi_{ij\alpha}(y')\,dy',\\ |(MD)^\alpha F_{x^i}(x)| &\le N \sup_{y':|y'|\le x^1R} |f_x(x'+y') - f_x(x')|,\\ |(x^1)^{|\alpha|} D^\alpha F_x(x)| &\le N \sup_{y':|y'|\le x^1R} |f_x(x'+y') - f_x(x')|, \qquad (6.4)\end{aligned}$$

where R is such that $\operatorname{supp}\varphi \in B_R$.

Now, if a portion of $\partial\Omega$ is given by the equation $x^1 = f(x')$ and Ω near this portion lies in $x^1 > f(x')$, then there the function ψ is constructed after Lemma 2.5 of [4] by means of solving the equation

$$x^1 = \psi(x) + F(\varepsilon\psi(x), x') \tag{6.5}$$

under the additional harmless assumptions that $\int \varphi\,dy' = 1$. The constant $\varepsilon > 0$ is chosen in such a way that $\varepsilon|F_{x^1}| \le 1/2$, so that (6.5) admits a smooth solution by the implicit function theorem. In that case also, for the function $E(r,x) := r + F(\varepsilon r, x') - x^1$, we have $E_r \le -1/2$. By Lemma 3.8 and estimates (6.3) and (6.4) we conclude that assertion (iv) of Lemma 2.6 holds for ψ defined from (6.5). For general C^1 domains one constructs ψ by "piecing together such local definitions of ψ by partitions of unity" (see [4]); one can find more detail in [11] and [12].

Proof of Theorem 2.12. This proof consists of five steps.

Step 1. First we construct $\mu(x)$ near the boundary of Ω_ε as a mapping that moves $x \in \Omega_\varepsilon = \{\psi > \varepsilon\}$ along the straight line $x(r) = x - r\psi_x(x)$ toward $\partial\Omega$ to a point y at which $\psi(y) = \psi(x) - \varepsilon$. The value of $\varepsilon > 0$ will be taken in a special way.

To make a preliminary choice, notice that we can take 2ψ in place of ψ, so that without losing generality we may assume that for an $\bar{\varepsilon} > 0$, we have $|\psi_x| > 1$ in $\Omega \setminus \Omega_{\bar{\varepsilon}}$. Then recalling that r_0 and K_0 are the constants from Assumption 2.1, define

$$M = \sup_{\Omega} |\psi_x|, \quad R = 10K_0^2 M^2,$$

and only concentrate on ε satisfying

$$0 < \varepsilon < (\bar{\varepsilon}/R) \wedge (r_0/(8K_0^2 M)). \tag{6.6}$$

Keeping in mind that $\psi(x)$ is equivalent to $\rho(x)$ near $\partial\Omega$ and that ψ_x is uniformly continuous, choose ε (satisfying (6.6) and) such that

$$x, y \in \bar{\Omega} \setminus \Omega_{R\varepsilon}, |x - y| \leq 2M\varepsilon \Longrightarrow \rho(x) < \frac{r_0}{2K_0^2}, \quad |\psi_x(x) - \psi_x(y)| \leq \frac{1}{8K_0^2 M}. \tag{6.7}$$

Next, for $x \in \bar{\Omega} \setminus \Omega_{R\varepsilon}$ and $r \in \mathbb{R}$ such that $x - r\psi_x(x) \in \bar{\Omega} \setminus \Omega_{R\varepsilon}$, introduce the functions

$$u = u(r, x) = x - r\psi_x(x), \quad E(r, x) = \psi(u) - \psi(x) + \varepsilon.$$

Observe that if $x, u(r, x) \in \bar{\Omega} \setminus \Omega_{R\varepsilon}$ and $|r| \leq 2\varepsilon$, then $|x - u| \leq 2M\varepsilon$ and $|\psi_x(u)| \geq 1$ (owing to $R\varepsilon \leq \bar{\varepsilon}$; see (6.6)) and

$$\begin{aligned} E_r(r, x) = &- \psi_{x^i}(x)\psi_{x^i}(u) = -|\psi_x(u)|^2 \\ &+ \big(\psi_{x^i}(u) - \psi_{x^i}(x)\big)\psi_{x^i}(u) \leq -1 + M/(8K_0^2 M) < -1/2. \end{aligned}$$

Hence, in this range of r, the functions $\psi(u(r, x))$ and $E(r, x)$ are strictly locally decreasing in r. In particular, if $x \in \partial\Omega$, then on the interval $-2\varepsilon \leq r < 0$ we have $\psi(u(r, x)) > |r|/2$. Hence $u(-2\varepsilon, x) \in \Omega_\varepsilon$, so that any point on $\partial\Omega$ can be connected by a straight line with a point in Ω_ε.

Furthermore, it is seen that for any $x \in \Omega_\varepsilon \setminus \Omega_{R\varepsilon}$ we have $E(0, x) = \varepsilon$ and there is a unique $r = r(x) \in (0, 2\varepsilon)$ such that

$$\varepsilon \geq E(r, x) > 0 \quad \text{for} \quad r \in [0, r(x)), \quad E(r(x), x) = 0. \tag{6.8}$$

Now we apply Lemma 3.8 to $G = \Omega_\varepsilon \setminus \bar{\Omega}_{R\varepsilon}$ and $d = \rho_{\Omega_\varepsilon}$. Notice that in G every derivative of $\psi(x)$ is bounded and that d on G is comparable with $\psi(x) - \varepsilon$ (cf. Remark 2.7). Also we introduce

$$u(x) = u(r(x), x) = x - r(x)\psi_x(x)$$

and notice that by definition $\psi(x) - \varepsilon = \psi(u(x))$ and the latter is comparable with $\rho(u(x))$. Then we see that instead of dealing with (3.11) it suffices to treat

$$(\psi(x) - \varepsilon)^{|\alpha|-1}(D^\alpha \psi)(u(x)),$$

which equals

$$\psi^{|\alpha|-1}(u(x))(D^\alpha \psi)(u(x)),$$

and the latter is bounded on G for any $\alpha \neq 0$ and, for $|\alpha| \geq 2$, tends to zero as $\psi(x) - \varepsilon = \psi(u(x)) \downarrow 0$ by Lemma 2.6.

Thus, by Lemma 3.8 we have that $(\psi(x)-\varepsilon)^{|\alpha|-1}D^\alpha r(x)$ is bounded in $\Omega_\varepsilon \setminus \Omega_{R\varepsilon}$ for any $\alpha \neq 0$ and, for $|\alpha| \geq 2$, tends to zero as $\psi(x)-\varepsilon \downarrow 0$.

Step 2. Now we can define μ in Ω_ε. Observe that $R > 10$, so that $1/(R-2) < 5/(4R)$ and there is an infinitely differentiable function $\kappa(t)$, $t \in \mathbb{R}$, such that $\kappa(t) = 1$ for $t \leq 2$, $\kappa(t) = 0$ for $t \geq R$, and $0 \leq -\kappa' \leq 5/(4R)$. In Ω_ε we define

$$\mu(x) = x - r(x)\kappa(\psi(x)/\varepsilon)\psi_x(x). \tag{6.9}$$

Strictly speaking, for $x \in \Omega_{R\varepsilon}$ the above formula has no sense since we have not defined $r(x)$ for $x \in \Omega_{R\varepsilon}$, but for those x we have $\psi(x) > R\varepsilon$, so that $\kappa(\psi(x)/\varepsilon) = 0$ and in that case we set by definition $\mu(x) = x$. Certainly, assertions (ii), (iii), (v), and (vi) hold in what concerns μ. Assertion (i) for μ follows from the formula

$$r_{x^i}(x)\psi_{x^j}(u)\psi_{x^j}(x) = \psi_{x^i}(u) - \psi_{x^i}(x) - r(x)\psi_{x^j}(u)\psi_{x^ix^j}(x), \tag{6.10}$$

which holds for $x \in \Omega_\varepsilon \setminus \Omega_{R\varepsilon}$ with $u = u(x)$ and which yields, first, the modulus of continuity of $r(x)$ and then that of $r_x(x)$ in $\Omega_\varepsilon \setminus \Omega_{R\varepsilon}$.

Step 3. Next, we come to defining ν. To be sure that μ is locally one-to-one, we fine tune the choice of ε in the following way. Observe that in $\Omega_\varepsilon \setminus \Omega_{R\varepsilon}$ as we know, $r(x) \in (0, 2\varepsilon)$ so that $r \leq 2\psi$ in $\Omega_\varepsilon \setminus \Omega_{R\varepsilon}$ and the last term in (6.10) is less than $2M\psi\psi_{x^ix^j}$. Furthermore, $\psi\psi_{xx}$ can be made arbitrary small in $\Omega \setminus \Omega_{R\varepsilon}$ on the account of appropriate choice of ε. Therefore, (6.10) and the estimates in the beginning of the proof (remember $|u-x| \leq 2\varepsilon M$) show that for sufficiently small ε we have

$$|r_x| \leq 2|\psi_x(u) - \psi_x(x)| + 1/(4MK_0^2) \leq 1/(2MK_0^2)$$

in $\Omega_\varepsilon \setminus \Omega_{R\varepsilon}$.

Now, by differentiating (6.9) in the direction of a unit vector $e \in \mathbb{R}^d$ we obtain in $\Omega_\varepsilon \setminus \Omega_{R\varepsilon}$ that

$$\begin{aligned} |\mu_{(e)}(x) - e| &\leq |r_{(e)}(x)|\kappa(\psi/\varepsilon)M + 2\varepsilon|\kappa'(\psi/\varepsilon)|M^2/\varepsilon + 2\psi|\psi_{x(e)}| \\ &\leq 1/(2K_0^2) + 10M^2/(4R) + 2\psi|\psi_{x(e)}| = 3/(4K_0^2) + 2\psi|\psi_{x(e)}|. \end{aligned}$$

By reducing further ε we arrive at a situation in which

$$|\mu_{(e)}(x) - e| \leq 4/(5K_0^2) < 1 \tag{6.11}$$

in $\Omega_\varepsilon \setminus \Omega_{R\varepsilon}$ and, actually, in Ω_ε since the first expression is zero in $\Omega_{R\varepsilon}$.

Estimate (6.11) allows us to solve the equation

$$\mu(\nu(x)) - x = 0 \tag{6.12}$$

near any point $x_0 = \mu(y_0)$ with $y_0 \in \Omega_\varepsilon$ (and the solution satisfying $\nu(x_0) = y_0$ is unique). Furthermore,

$$\rho^{|\alpha|-1}(x)(D^\alpha\mu)(\nu(x)) = \rho^{|\alpha|-1}(\mu(y))D^\alpha\mu(y)|_{y=\nu(x)}$$

and, due to the part of assertion (ii) proved for μ, the latter is bounded. Also observe that, for any solution of (6.12) and $x \in \Omega \setminus \Omega_{R\varepsilon}$, we have $\nu(x) \in \Omega_\varepsilon \setminus \Omega_{R\varepsilon}$ (otherwise $\mu(\nu(x)) = \nu(x)$), so that, by virtue of (6.8) and the fact that $0 \leq \kappa \leq 1$, we have

$$\psi(x) = \psi(\mu(\nu(x))) \geq \psi(u(\nu(x))) = \psi(\nu(x)) - \varepsilon.$$

It follows that if $\Omega \ni x \to \partial\Omega$, then $y = \nu(x) \to \partial\Omega_\varepsilon$ and

$$|\rho^{|\alpha|-1}(\mu(y))D^\alpha \mu(y)| \le N\psi^{|\alpha|-1}(\mu(y))D^\alpha\mu(y)| \to 0,$$

as have been pointed out above. This and Lemma 3.8 applied to (6.12) would have finished the proof of the theorem if we already knew that $\nu(x)$ was uniquely defined in Ω.

Step 4. To show that $\nu(x)$ is indeed uniquely defined in Ω, we first prove that

$$\Omega' := \mu(\Omega_\varepsilon) = \Omega. \tag{6.13}$$

Referring to (6.11), we conclude that Ω' is an open subset of Ω. On the other hand, if $y_n \in \Omega'$ and $y \in \Omega$ and $y_n \to y$, then for $x_n \in \Omega_\varepsilon$ such that $\mu(x_n) = y_n$ we have

$$x_n - y_n = r(x_n)\kappa(\psi(x_n)/\varepsilon)\psi_x(x_n).$$

Since the right-hand side is bounded, there is a subsequence $x_{n'}$ converging to a point $x \in \bar\Omega_\varepsilon$.

It turns out that $x \in \Omega_\varepsilon$. Indeed, if $x \in \partial\Omega_\varepsilon$, then $\kappa(\psi(x_{n'})/\varepsilon) = 1$ for all large n' and $\psi(x_{n'}) = \psi(y_{n'}) + \varepsilon$, so that $\psi(x) = \psi(y) + \varepsilon$. Here $y \in \Omega$ and $\psi(y) > 0$, implying $\psi(x) > \varepsilon$ and contradicting $x \in \partial\Omega_\varepsilon$.

By passing to the limit in $\mu(x_{n'}) = y_{n'}$ we now see that $\mu(x) = y$ and $y \in \Omega'$. This means that Ω' is not only open but also is closed in the relative topology of Ω. It follows that Ω' is the union of some connected components of Ω. In addition, since as was noted above, any point on $\partial\Omega$ can be connected by a straight line with a point in Ω_ε, each connected component of Ω contains points of Ω_ε and thus points of Ω'. We have proved (6.13).

Step 5. It remains to prove that the mapping μ is one-to-one. To this end we make the final adjustment of ε which gives us the possibility of connecting close points in Ω_ε by paths of length comparable with the distance between the points. So far the relation of ε to r_0 and K_0 did not play much of a role. Now it becomes crucial.

As is easy to see, due to Assumption 2.1 for any $x_0 \in \partial\Omega$, we have $B_{r_0/K_0} \cap \mathbb{R}^d_+ \subset \Psi(B_{r_0}(x_0) \cap \Omega)$. Furthermore, the function $\psi(\Psi^{-1})$ is continuously differentiable in the closure of $B_{r_0/K_0} \cap \mathbb{R}^d_+$ and vanishes on the set $\{y^1 = 0\} \cap B_{r_0/K_0}$, so that its gradient on this set is parallel to the y^1 axis. It follows that for sufficiently small $\varepsilon > 0$ the angle, which the gradient of $\psi(\Psi^{-1})$ makes with the y^1 axis on the surface $\{\psi(\Psi^{-1}) = \varepsilon\} \cap B_{r_0/K_0}$, is as small as we like. We make it so small that any two points $y_1, y_2 \in \{\psi(\Psi^{-1}) > \varepsilon\} \cap B_{r_0/K_0}$ could be connected by a path lying in $\{\psi(\Psi^{-1}) > \varepsilon\} \cap B_{r_0/K_0}$ and consisting of two straight segments of total length $\le (10/9)|y_1 - y_2|$.

Next, if $x_1, x_2 \in \Omega$ and

$$\rho(x_1) < r_0/(2K_0^2) \quad \text{and} \quad |x_1 - x_2| < r_0/(2K_0^2),$$

then there is an $x_0 \in \partial\Omega$ such that

$$x_1, x_2 \in B_{r_0/K_0^2}(x_0) \cap \Omega$$

and the images y_1 and y_2 of x_1, x_2 under Ψ lie in $B_{r_0/K_0} \cap \mathbb{R}^d_+$ and $|y_1 - y_2| \le K_0|x_1 - x_2|$. If in addition $x_1, x_2 \in \Omega_\varepsilon$, then by the above paragraph there is a path in Ω_ε connecting x_1 and x_2 and having length $\le K_0(10/9)|y_1 - y_2| \le (10/9)K_0^2|x_1 - x_2|$.

Now, assume that $\mu(x_1) = \mu(x_2)$ for some $x_i \in \Omega_\varepsilon$. If $x_1, x_2 \in \Omega_{R\varepsilon}$, then $\mu(x_1) = x_1 = \mu(x_2) = x_2$. If, say $x_1 \in \Omega \setminus \Omega_{R\varepsilon}$, then (see (6.7)) we have $\rho(x_1) < r_0/(2K_0^2)$ and

$$|x_1 - x_2| = |r(x_1)\kappa(\psi(x_1)/\varepsilon)\psi_x(x_1) - r(x_2)\kappa(\psi(x_2)/\varepsilon)\psi_x(x_2)|$$
$$\leq |r(x_1)\kappa(\psi(x_1)/\varepsilon)\psi_x(x_1)| + |r(x_2)\kappa(\psi(x_2)/\varepsilon)\psi_x(x_2)| \leq 4M\varepsilon.$$

In addition (see (6.6)) $4M\varepsilon < r_0/(2K_0^2)$, so that the points x_1 and x_2 can be connected by a path $s(t)$, $t \in [0,1]$, lying in Ω_ε and having length $\leq (10/9)K_0^2|x_1 - x_2|$. In light of (6.11), we obtain

$$|x_1 - x_2| = |\mu(x_1) - \mu(x_2) - (x_1 - x_2)| \leq \int_0^1 |(\mu(s(t)))' - s'(t)|\,dt$$
$$\leq 4/(5K_0^2)\int_0^1 |s'(t)|\,dt \leq (8/9)|x_1 - x_2|,$$

implying that $|x_1 - x_2| = 0$. We see that the conditions $\mu(x_1) = \mu(x_2)$ and $x_1, x_2 \in \Omega_\varepsilon$ imply that $x_1 = x_2$. Therefore, $\nu = \mu^{-1}$ is well defined indeed and the theorem is proved.

Acknowledgments. The authors are sincerely grateful to the referees for useful comments and suggestions.

REFERENCES

[1] J. Chabrowski, *The Dirichlet Problem with L^2 Boundary Data for Elliptic Equations*, Lecture Notes in Math. 1482, Springer-Verlag, New York, 1991.
[2] A. Douglis and L. Nirenberg, *Interior estimates for elliptic systems of partial differential equations*, Comm. Pure Appl. Math., 8 (1955), pp. 503–538.
[3] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.
[4] D. Gilbarg and L. Hörmander, *Intermediate Schauder estimates*, Arch. Rational Mech. Anal., 74 (1980), pp. 297–318.
[5] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin, 1983.
[6] K.-H. Kim and N. V. Krylov, *On SPDEs with variable coefficients in one space dimension*, Potential Anal., 21 (2004), pp. 209–239.
[7] N. V. Krylov, *A generalization of the Littlewood-Paley inequality and some other results related to stochastic partial differential equations*, Ulam Quart., 2 (1994), pp. 16–26 (electronic). Also available online at http://www.ulam.usm.edu/VIEW2.4/krylov.ps.
[8] N. V. Krylov, *An analytic approach to SPDEs*, in Stochastic Partial Differential Equations: Six Perspectives, Math. Surveys Monogr. 64, AMS, Providence, RI, 1999, pp. 185–242.
[9] N. V. Krylov, *Weighted Sobolev spaces and Laplace equations and the heat equations in a half space*, Comm. Partial Differential Equations, 24 (1999), pp. 1611–1653.
[10] O. A. Ladyzhenskaya, V. A. Solonnikov, and N. N. Ural'tceva, *Linear and Quasi-Linear Parabolic Equations*, Nauka, Moscow, 1967 (in Russian); AMS, Providence, RI, 1968 (in English).
[11] S. K. Lapic, *On the First-Initial Boundary Value Problem for Stochastic Partial Differential Equations*, Ph.D. thesis, University of Minnesota, Minneapolis, MN, 1994.
[12] G. Lieberman, *Regularized distance and its applications*, Pacific J. Math., 117 (1985), pp. 329–352.
[13] G. Lieberman, *Second Order Parabolic Differential Equations*, World Scientific, River Edge, NJ, 1996.
[14] S. V. Lototksy, *Dirichlet problem for stochastic parabolic equations in smooth domains*, Stochastics Stochastics Rep., 68 (1999), pp. 145–175.
[15] S. V. Lototksy, *Sobolev spaces with weights in domains and boundary value problems for degenerate elliptic equations*, Methods Appl. Anal., 1 (2000), pp. 195–204.
[16] S. V. Lototksy, *Linear stochastic parabolic equations, degenerating on the boundary of a domain*, Electron. J. Probab., 6 (2001), pp. 1–14.

SIAM J. MATH. ANAL.
Vol. 36, No. 2, pp. 643–658

INFINITE JACOBI MATRICES WITH UNBOUNDED ENTRIES: ASYMPTOTICS OF EIGENVALUES AND THE TRANSFORMATION OPERATOR APPROACH*

JAN JANAS† AND SERGUEI NABOKO‡

Abstract. In this paper the exact asymptotics of eigenvalues $\lambda_n(J)$, $n \to \infty$, of a class of unbounded self-adjoint Jacobi matrices J with discrete spectrum are given. Their calculation is based on a successive diagonalization approach—a new version of the classical transformation operator method. The approximations of the transformation operator are constructed step by step using a successive diagonalization procedure, which results in higher order approximations of the $\lambda_n(J)$.

Key words. unbounded Jacobi operator, asymptotics of eigenvalues, transformation operator, successive diagonalization

AMS subject classifications. 47B36, 47A75

DOI. 10.1137/S0036141002406072

1. Introduction. In this work we consider a class of infinite Jacobi matrices acting in $l^2 = l^2(\mathbb{N})$. Although the class we study looks rather special, it is large enough to contain examples of physical interest. More precisely we investigate the asymptotics of the energy spectrum of a molecule in the homogeneous electric field. The asymptotics of the spectrum are obtained via a general successive diagonalization method which seems to be of independent interest and can be used for other Jacobi matrices with discrete spectrum. We emphasize that the special class of Jacobi matrices under consideration has been chosen for simplicity only. Our goal is to present the general ideas of the method, avoiding tedious calculations and complicated formulations.

Recall that an essentially self-adjoint Jacobi matrix

$$J = \begin{pmatrix} q_1 & \lambda_1 & 0 & 0 & \cdots \\ \lambda_1 & q_2 & \lambda_2 & 0 & \cdots \\ 0 & \lambda_2 & q_3 & \lambda_3 & \cdots \\ 0 & 0 & \lambda_3 & q_4 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

which is a relatively compact perturbation of the diagonal operator Q (given by $Qf_n = q_n f_n$ in the canonical basis f_n of l^2) under suitable conditions on its entries, has a compact resolvent and so its spectrum is discrete. A simple sufficient condition for the discreteness of the spectrum of J was given, for example, in [19] and says

$$\text{if } \lim_n \inf \frac{q_n^2}{\lambda_n^2 + \lambda_{n-1}^2} > 2, \text{ then } J \text{ has discrete spectrum}$$

(we assume that $\lambda_n > 0$). Another sufficient condition for discreteness of the spectrum of J was obtained in [36] (using continuous fraction representation of the Weyl function

*Received by the editors April 22, 2002; accepted for publication (in revised form) January 16, 2004; published electronically August 27, 2004. This work was supported by KBN grant 5 PO3A/026/21.

http://www.siam.org/journals/sima/36-2/40607.html

†Institute of Mathematics, Polish Academy of Sciences, Cracow Branch, Św. Tomasza 30, 31-027 Kraków, Poland (najanas@cyf-kr.edu.pl).

‡Department of Mathematical Physics, Institute of Physics, St. Petersburg University, Ulianovskaia 1, 198904, St. Petergoff, St. Petersburg, Russia (naboko@snoopy.phys.spbu.ru).

of J). The class of J we consider below has bounded weights λ_n and $q_n = n^2 + c_1 n + O(1)$. The idea of successive diagonalization uses two general ingredients.

First, for a given compact perturbation R of a diagonal self-adjoint operator D ($De_n = \mu_n e_n$, where e_n is an orthonormal basis in a Hilbert space) we assume that (a) dist $(\mu_k, \mu_l) \geq \varepsilon_0 > 0$ if $k \neq l$ and at least k or l is large enough, and (b) R is a band matrix in the basis e_n. We stress that R can be non-self-adjoint (as it will be in our situation of self-adjoint Jacobi matrices). Under the above assumption $\sigma(D+R)$ is obviously discrete and the eigenvalues $\lambda_k(D+R) = \mu_k + O(\|R^* e_n\|)$. In this way we are able to control the distance $|\lambda_k(D+R) - \mu_k|$. Note that even for a rank-one R it may happen that the above distance can be an arbitrary l^1 sequence (see Example 2.3).

Second, we look for a diagonal matrix that is similar to the original J modulo some compact band matrix (in the same basis f_n).

This idea in some important features goes back to the old method of transformation operator presented by Delsarte [10], Delsarte and Lions [11], Marchenko [32], Levitan [30], and Naimark [33]. It was applied to calculate the exact asymptotics of the eigenvalues of Sturm–Liouville operators with smooth coefficients. However, we do not try to find exactly the transformation operator; instead, a successive approximation procedure is suggested. A somewhat similar idea of approximate similarity was used by Rosenblum in his remarkable paper [34] on asymptotics of the eigenvalues of some pseudodifferential operators. But our class of unbounded Jacobi matrices is quite different from the one of pseudodifferential operators considered by Rosenblum. Note that in this way we have to leave the class of self-adjoint operators, but we stay in the class of band matrices that are a weak perturbation of the diagonal matrix. We repeat this procedure several times (successive diagonalization), obtaining finally as weak perturbation of the diagonal operator, as we require. This allows us to find the value of the eigenvalues $\lambda_n(J)$ with arbitrary precision as $n \to \infty$. In this paper we restrict ourselves to three steps to avoid the tedious calculation necessary for higher order approximation. Although we concentrate on finding asymptotics of $\lambda_n(J)$, the approach also works simultaneously for computing approximate formulas of the eigenvectors of J. Surely this is a natural consequence of realization of the transformation operator idea. Actually, the spectral analysis of infinite self-adjoint Jacobi matrices is a classical topic related to the theory of orthogonal polynomials [1], [2], [3], [4], [8]. Numerous methods were elaborated on for studies of spectral problems for various classes of Jacobi matrices in $l^2(\mathbb{N})$ and $L^2(\mathbb{Z})$; see [12], [13], [14], [15], [16], [19], [20], [21], [22], [23], [24], [25], [26], [27], [35].

Note that calculation of exact asymptotics of eigenvalues of ordinary differential operators is a classical problem studied by Birkhoff and Tamarkin and for partial differential operators by Weyl. The same question about the asymptotics of eigenvalues can be investigated for difference operators (Jacobi matrices).

In turn, in physical models it is important to analyze the influence of physical parameters on the spectrum of Hamiltonians. Since the exact calculation of their eigenvalues is usually impossible, it is of considerable interest to study the appearance of the parameters in different terms of asymptotics of the eigenvalues. This allows better understanding of the role of parameters in the spectral picture of Hamiltonians and therefore in the corresponding physical model as well.

Let us also recall that Jacobi matrices appear in applications to many fields, including quantum mechanics, quantum optics, solid state physics, and numerical analysis. Already classical operators of creation a and annihilation $a^+ = a^*$ of quan-

tum physics can be represented by the nonsymmetric Jacobi matrix (a_{ij}) with only one nonzero diagonal $a_{ij} = \sqrt{i-1}\delta_{ij+1}$, where δ_{ij} is the Kronecker delta function. In quantum optics, Hamiltonians in simple models can be represented by polynomials in a and a^+. Therefore in a suitable orthonormal basis they are given by Hermitian band matrices. In some cases such models can be reduced to symmetric Jacobi matrices, for example, in the case of the Jaynes–Cumming model without rotating wave approximation (RWA) studied in [31]. In turn, in solid state physics models appear as Jacobi matrices with almost periodic entries [9].

The paper is organized as follows. In section 2 the first abstract fact is proved. The above-mentioned three steps of successive diagonalization are presented in section 3. Finally, the application to a physical model is considered briefly in section 4.

2. Preliminaries. Before proceeding further let us recall some notation and notions. Let $\{f_n\}_{n=1}^{\infty}$ be the canonical orthonormal basis in $l^2 = l^2(\mathbb{N})$. For given sequences $\{\lambda_n\}_{n=1}^{\infty}$, $\{q_n\}_{n=1}^{\infty}$ of real numbers one defines the Jacobi matrix J by $J = SD + DS^* + Q$, where D and Q are the diagonal operators defined by $\{\lambda_n\}_{n=1}^{\infty}$ and $\{q_n\}_{n=1}^{\infty}$, respectively, and the unilateral shift is defined by $Sf_n = f_{n+1}$. As mentioned, the class of Jacobi operators we study in this paper is given by $q_n = n^2 + c_1 n + c_2 n^{-1} + c_3 n^{-2} + O(\frac{1}{n^3})$, $\lambda_n = g + b_1 n^{-1} + b_2 n^{-2} + O(\frac{1}{n^3})$, where all c_j, b_j are real and such that $\lambda_n \neq 0$ (as usual). We omit the zero order term in q_n since it produces the shift of the spectrum only. The physical motivation for so special a choice of q_n and λ_n will be given in section 4. Denote by $J(= J(g, b_1, b_2, c_1, c_2, c_3))$ the Jacobi operator induced by the above sequences. The spectrum of J is obviously discrete and consists of the eigenvalues $\{\lambda_k(J)\}_{k=1}^{\infty}$ [18].

In what follows $\Sigma_{1/p}$ $(p > 0)$ stands for the weak ideal of Schatten–von Neumann compact operators T such that their singular numbers $s_k(T)$ satisfy estimates $s_k(T) = O(\frac{1}{k^p})$ [5]. Below $\| \|$ denotes the operator norm and $[A, B] := AB - BA$ is the commutator of operators A and B.

We start by proving a folklore-type, simple lemma, which will be used as a tool of successive approximation of the eigenvalues $\lambda_k(J)$.

LEMMA 2.1. *Let D be a self-adjoint diagonal operator in a Hilbert space H given by $De_n = \mu_n e_n$, where $\{e_n\}$ is an orthonormal basis of eigenvectors in H and simple eigenvalues $\mu_n \to \infty$ are ordered by $|\mu_i| \leq |\mu_{i+1}|$. Assume that*

$$\text{(i)} \quad \textit{dist}\,(\mu_i, \mu_k) \geq \varepsilon_0 \textit{ for some } \varepsilon_0 > 0 \textit{ and all } i \neq k.$$

If R is a compact (not necessary self-adjoint) operator in H, then the operator $T = D + R$ has discrete spectrum when the complex eigenvalues $\lambda_n(T)$ of T are numerated properly. Moreover, for large values of n the eigenvalues of T are simple and $\lambda_n(T)$ satisfy the estimates $\lambda_n(T) = \mu_n + O(\|R^* e_n\|)$.[1]

Proof. Let

$$K_n := \{\lambda \in \mathbb{C},\ |\lambda - \mu_n| \leq r_n\}$$

with $r_n = O(\|R^* e_n\|)$. We claim two facts. First, there exists a collection of the eigenvalues of T such that for arbitrary n the distance between μ_n and an eigenvalue of T from the collection fulfills the estimate $O(\|R^* e_n\|)$, as $n \to \infty$. Second, all the eigenvalues of T belong to the collection and therefore satisfy the estimates of

[1] The above formula induces a proper numeration of $\lambda_n(T)$ which does not coincide in general with monotonic ordering of their moduli, due to possible jumps of the signs of μ_n.

Lemma 2.1 after a proper numeration. Note that the above inessential complicated formulation appears due to not fixed signs of μ_n.

To prove the first fact we find numbers $r_n < \varepsilon_0/2$ such that for large n in the disc K_n there is exactly one simple eigenvalue of T. Below we essentially follow the idea of the proof of the stability of the multiplicity from [5]. For $\lambda \in \mathbb{C} \setminus \sigma(D)$ we have

$$(T-\lambda)^{-1} = [I + (D-\lambda)^{-1}R]^{-1}(D-\lambda)^{-1}. \tag{2.1}$$

Assume that for $n \gg 1$ we have chosen r_n such that

$$(*)\sup\|(D-\lambda)^{-1}R\| \le \frac{1}{3}, \quad \lambda \in \Gamma_n, \text{where } \Gamma_n = \partial K_n.$$

Since $r_n < \varepsilon_0/2$, using (2.1) and (i) we see that $\sigma(T) \cap \Gamma_n = \phi$. Fix large n for which $(*)$ holds. Now we follow more or less standard reasoning. Let P_T and P_D be the Riesz projections given by

$$P_T := -\frac{1}{2\pi i}\int_{\Gamma_n}(T-\lambda)^{-1}d\lambda, \quad P_D := -\frac{1}{2\pi i}\int_{\Gamma_n}(D-\lambda)^{-1}d\lambda = (\cdot, e_n)e_n.$$

Denote $F_\lambda := (D-\lambda)^{-1}R$. Then using (2.1) we have

$$P_T = P_D - \frac{1}{2\pi i}\int_{\Gamma_n}\left[\sum_{k=1}^{\infty}(-1)^k F_\lambda^k\right](D-\lambda)^{-1}d\lambda.$$

Due to $(*)$ one can estimate

$$\|P_T - P_D\| \le \frac{1}{2\pi}|\Gamma_n| \sup_{\Gamma_n}\sum_{k=1}^{\infty}\|F_\lambda\|^k r_n^{-1} = \sup_{\Gamma_n}[\|F_\lambda\|(1-\|F_\lambda\|)^{-1}] \le 1/2.$$

Hence rank P_T = rank $P_D = 1$. Therefore there is in K_n exactly one eigenvalue of T provided one can find r_n for which $(*)$ is satisfied.

Consider the Schmidt decomposition $R = \sum_{k=1}^{\infty} s_k(R)(\cdot, \varphi_k)\psi_k$, where $\{\varphi_n\}$ and $\{\psi_n\}$ are some orthonormal bases in H and $s_k(R)$ are the singular numbers of the compact operator R [5], [28]. Let $P_n := (\cdot, e_n)e_n$ and $P_n^\perp := I - P_n$.

Then $\|R^*(D-\overline{\lambda})^{-1}\| \le \|R^*(D-\overline{\lambda})^{-1}P_n\| + \|R^*(D-\overline{\lambda})^{-1}P_n^\perp\|$. However,

$$\|R^*(D-\overline{\lambda})^{-1}P_n\| = \|R^* e_n\| r_n^{-1}, \quad \lambda \in \Gamma_n. \tag{2.2}$$

On the other hand, for $f \in H$

$$\begin{aligned}\|R^*(D-\overline{\lambda})^{-1}P_n^\perp f\|^2 &= \sum_{k=1}^{\infty} s_k(R)^2 |(P_n^\perp(D-\overline{\lambda})^{-1}P_n^\perp f, \psi_k)|^2 \\ &\equiv \sum_{k=1}^{N} s_k(R)^2|\cdots|^2 + \sum_{k=N+1}^{\infty} s_k(R)^2|\cdots|^2.\end{aligned}$$

Since $\|P_n^\perp(D-\overline{\lambda})^{-1}P_n^\perp\| \le 2\varepsilon_0^{-1}$ (remember that $r_n \le \varepsilon_0/2$) and $\lambda \in \Gamma_n$,

$$\sum_{k=N+1}^{\infty} s_k(R)^2 |(P_n^{\perp}(D-\overline{\lambda})^{-1}P_n^{\perp}f,\psi_k)|^2$$

$$\leq s_{N+1}(R)^2 \cdot (2\varepsilon_0^{-1}\|f\|)^2 \leq \frac{1}{32}\|f\|^2 \tag{2.3}$$

for N sufficiently large.

Fix such a large value of N. Then $|(P_n^{\perp}(D-\overline{\lambda})^{-1}P_n^{\perp}f,\psi_k)|^2 \to 0$ as $n\to\infty$ (uniformly in $\lambda\in\Gamma_n$ and f in the unit ball) and $k=1,\dots,N$.

Indeed, if $Q_L=\sum_{l=1}^{L}P_l$, then

$$\begin{aligned}|(P_n^{\perp}(D-\overline{\lambda})^{-1}P_n^{\perp}f,\psi_k)| &\leq |(P_n^{\perp}(D-\overline{\lambda})^{-1}P_n^{\perp}f,Q_L\psi_k)| \\ &\quad + |(P_n^{\perp}(D-\overline{\lambda})^{-1}P_n^{\perp}f,(I-Q_L)\psi_k)| \\ &\leq \sum_{l=1}^{L}|(P_n^{\perp}(D-\overline{\lambda})^{-1}P_n^{\perp}f,P_l\psi_k)| + 2\varepsilon_0^{-1}\|(I-Q_L)\psi_k\|\cdot\|f\|. \end{aligned} \tag{2.4}$$

The second term from the above line can be made arbitrarily small by choosing L sufficiently large. For such fixed L and k each term $|(P_n^{\perp}(D-\overline{\lambda})^{-1}P_n^{\perp}f,P_l\psi_k)|$, $1\leq l\leq L$, obviously tends to zero uniformly in f from the unit ball as $n\to\infty$ (since $\sup_{\lambda\in\Gamma_n}\|P_l(D-\overline{\lambda})^{-1}\|\to 0$ as $n\to\infty$). Fix L so large that

$$2(2/\varepsilon_0)^2\sum_{k=1}^{N}s_k(R)^2\|(I-Q_L)\psi_k\|^2 \leq \frac{1}{32}.$$

For such large fixed L and $n\gg 1$ and $\lambda\in\Gamma_n$,

$$\sum_{k=1}^{N}s_k(R)^2|(P_n^{\perp}(D-\overline{\lambda})^{-1}P_n^{\perp}f,\psi_k)|^2 \leq \frac{1}{32}\|f\|^2. \tag{2.5}$$

Put $r_n=12\|R^*e_n\|$. Combining (2.2), (2.3), (2.4), and (2.5), we obtain the desired estimate $(*)$ (since all the above estimates were uniform with respect to $\lambda\in\Gamma_n$ and $\frac{1}{12}+\sqrt{\frac{1}{32}+\frac{1}{32}}=\frac{1}{3}$).

To prove the second fact, choose the radii $R_n\to+\infty$, $n\to\infty$, such that the circles $\Gamma_{R_n}:=\{z,|z|=R_n\}$ satisfy the estimates $\operatorname{dist}(\Gamma_{R_n},\sigma(D))\geq\varepsilon_0/4$ (using assumption (i)). As above, we want to prove the new estimate (compare to $(*)$) for large n,

$$\sup\|(\mathrm{D}-\lambda)^{-1}R\|\leq\frac{1}{3},\qquad \lambda\in\Gamma_{R_n}. \tag{2.6}$$

Since $s_k(R)\to 0$, as $k\to\infty$,

$$\sum_{k=N+1}^{\infty}s_k(R)^2|(P_n^{\perp}(\mathrm{D}-\overline{\lambda})^{-1}P_n^{\perp}f,\psi_k)|^2 \leq s_{N+1}(R)^2(4\varepsilon_0^{-1})^2\|f\|^2 \leq \frac{1}{4}\|f\|^2 \tag{2.7}$$

for N sufficiently large. Fix such N. To complete the proof it is enough to estimate

$$\sum_{k=1}^{N}s_k(R)^2|(P_n^{\perp}(\mathrm{D}-\overline{\lambda})^{-1}P_n^{\perp}f,\psi_k)|^2.$$

If $\lambda \in \Gamma_{R_n}$ we have

$$\|(\mathrm{D}-\overline{\lambda})^{-1}P_n^{\perp}f\|^2 = \sum_{l=1}^{\infty}|(P_n^{\perp}f,e_l)|^2|\mu_l-\lambda|^{-2} \le \sum_{l=1}^{\infty}|(P_n^{\perp}f,e_l)|^2(4\varepsilon_0^{-1})^2.$$

Since $|\mu_l+R_n|^{-2}+|\mu_l-R_n|^{-2}\to 0$, as $n\to\infty$, for $l=1,2,\dots$, we get

$$\sum_{k=1}^{N} s_k(R)^2|(P_n^{\perp}(\mathrm{D}-\overline{\lambda})^{-1}P_n^{\perp}f,\psi_k)|^2 = o(1)N\|f\|^2 \tag{2.8}$$

as $n\to\infty$. Here $o(1)$ tends to zero uniformly in f when $n\to\infty$. Finally (2.7) and (2.8) imply the desired estimate (2.6) for large n.

Our second fact will be proved once we show that for the discs $K(0,R_n)$ of radius R_n and the center at zero that

$$\sharp(K(0,R_n)\cap\sigma(T)) = \sharp(K(0,R_n)\cap\sigma(\mathrm{D})) \tag{2.9}$$

for n sufficiently large.

Note that due to the separation assumption (i), the circles Γ_n separate pairs of neighboring eigenvalues of D and therefore they do the same for neighboring eigenvalues of T for sufficiently large n. Finally, combining (2.9) and the formula $r_n = 12\|Re_n^*\|$ we get the desired asymptotic estimates $\lambda_n(T)=\mu_n+O(\|R^*e_n\|)$. It remains to prove (2.9). As above, it is enough to show that $\|P_T-P_{\mathrm{D}}\|<1$, where P_T (resp., P_{D}) are the Riesz projections corresponding to $\sigma(T)\cap K(0,R_n)$ ($\sigma(\mathrm{D})\cap K(0,R_n)$, resp.).

Applying (2.6) we have

$$\begin{aligned}\|P_T-P_{\mathrm{D}}\| &\le \frac{1}{2\pi}\int_{\Gamma_{R_n}}\|(D-\lambda)^{-1}R(T-\lambda)^{-1}\||d\lambda| \\ &\le \frac{1}{2\pi}\cdot\frac{3}{2}\int_{\Gamma_{R_n}}\|(\mathrm{D}-\lambda)^{-1}R(\mathrm{D}-\lambda)^{-1}\||d\lambda|.\end{aligned}$$

We claim that

$$\lim_n \int_{\Gamma_{R_n}}\|(\mathrm{D}-\lambda)^{-1}R(\mathrm{D}-\lambda)^{-1}\||d\lambda| = 0. \tag{2.10}$$

First observe that by the separation condition

$$\int_{\Gamma_{R_n}}\|(\mathrm{D}-\lambda)^{-1}R(\mathrm{D}-\lambda)^{-1}\||d\lambda| \le \|R\|\int_{\Gamma_{R_n}}\|(\mathrm{D}-\lambda)^{-1}\|^2|d\lambda| \le C\|R\| \tag{2.11}$$

for some positive C, $n=1,2,\dots$.

Fix $\varepsilon>0$ for which $C\cdot\varepsilon \ll 1$ and choose M so large that

$$R = R_\varepsilon + \sum_{k=1}^{M} s_k(R)(\dots,\varphi_k)\psi_k, \text{ where } \|R_\varepsilon\|\le\varepsilon. \tag{2.12}$$

Using (2.11) and (2.12), the above claim is reduced to the convergence

$$\lim_n \int_{\Gamma_{R_n}}\|(\mathrm{D}-\lambda)^{-1}(\langle\cdot,\varphi_k\rangle)\psi_k(\mathrm{D}-\lambda)^{-1}\||d\lambda| = 0,\ k=1,\dots,M.$$

Below we omit the index k. Applying the Cauchy–Schwarz inequality, it is enough to prove that for any $\varphi \in H$

$$\lim_n \int_{\Gamma_{R_n}} \|(\mathrm{D}-\lambda)^{-1}\varphi\|^2 |d\lambda| = 0. \tag{2.13}$$

The above integral can be estimated easily:

$$\begin{aligned}\int_{\Gamma_{R_n}} \|(\mathrm{D}-\lambda)^{-1}\varphi\|^2 |d\lambda| &= \int_{\Gamma_{R_n}} \sum_{k=1}^{\infty} |(\varphi, e_k)|^2 |(\mu_k-\lambda)^{-2}||d\lambda| \\ &= \sum_{k=1}^{\infty} |(\varphi, e_k)|^2 \int_{\Gamma_{R_n}} |(\mu_k-\lambda)^{-2}||d\lambda|.\end{aligned}$$

As it has been noticed above,

$$\int_{\Gamma_{R_n}} |(\mu_k-\lambda)^{-2}||d\lambda| \le \int_{\Gamma_{R_n}} \|(\mathrm{D}-\lambda)^{-1}\|^2 |d\lambda| \le C$$

for all n and k. On the other hand, for large n and some positive constant C_1,

$$\int_{\Gamma_{R_n}} |(\mu_k-\lambda)^{-2}||d\lambda| \le C_1 \int_{\Gamma_{R_n}} |\lambda|^{-2} |d\lambda| = O(R_n^{-1})$$

(uniformly in k). Combining the last two estimates we obtain the relation (2.13). This completes the proof of Lemma 2.1. □

Remark 2.2. A similar result holds in the case in which the eigenvalues $\{\mu_n\}$ of D are not simple. If $|\mu_l| \le |\mu_{l+1}|$ and $P_{\{\mu_l\}}$ is the orthogonal projection on the space of eigenvectors corresponding to μ_l, and n_l stands for the multiplicity of μ_l, then the eigenvalues $\lambda_{n(l)}(T)$ (counted properly with algebraic multiplicity) satisfy for $l \gg 1$

$$\lambda_{n(l)}(T) = \mu_l + O(\|R^* P_{\{\mu_l\}}\|),$$

where $n_1 + \cdots + n_{l-1} < n(l) \le n_1 + \cdots + n_l$. In the applications given in the next section we shall need estimates: $\|R^* e_n\| = O(\frac{1}{n^p})$, $p > 0$. Note that the above estimate on R cannot be replaced by $R \in \Sigma_{1/p}$.

Indeed, one can add to a given operator D a diagonal operator $R = \mathrm{diag}(\frac{1}{n_k^p})$, where n_k is an arbitrary permutation of $\mathbb{N}$. It is clear that by a proper choice of the permutation n_k we have $R \in \Sigma_{1/p}$, but we are not able to prove any estimates of $\mathrm{dist}(\lambda_n(T), \mu_n)$ $(T = D + R)$, only its decreasing to 0, as n tends to infinity.

Moreover, even for rank-one perturbation of R, the statement $\lambda_n(T) - \mu_n = O(\frac{1}{n^p})$ does not hold true.

Example 2.3. Indeed, choose the operator T in l^2 given by $T = \mathrm{D} + (\cdot, \varphi)\varphi$, where $\mathrm{D}e_n = ne_n$, $\varphi \in l^2, \|\varphi\| = 1$; i.e., $R = (\cdot, \varphi)\varphi$. Straightforward calculation shows that in the interval $(n, n+1)$ there exists exactly one eigenvalue λ_n of T satisfying the following condition:

$$\sum_{k=1}^{\infty} |\varphi_k|^2 (k - \lambda_n)^{-1} = -1, \tag{2.14}$$

where $\varphi = \sum_{k=1}^{\infty} \varphi_k e_k$, and the Fourier coefficients $\varphi_n \neq 0$ for all $_n$.

We have for fixed n

$$|\varphi_n|^2(n-\lambda_n)^{-1}+|\varphi_{n+1}|^2(n+1-\lambda_n)^{-1}+\sum_{k\neq n,n+1}|\varphi_k|^2(k-\lambda_n)^{-1}=-1.$$

However,

$$\sum_{k\neq n,n+1}\varphi_k^2(k-\lambda_n)^{-1}|\leq 2\sum_{k\neq n}|\varphi_k|^2|n-k|^{-1}=o(1)\quad\text{as}\quad n\to\infty.$$

Thus

$$-1+o(1)=|\varphi_n|^2(n-\lambda_n)^{-1}+|\varphi_{n+1}|^2(n+1-\lambda_n)^{-1}\geq|\varphi_n|^2(n-\lambda_n)^{-1}$$

and so $\lambda_n-n=|\lambda_n-n|\leq|\varphi_n|^2(1+o(1))$.

It follows that $\lim_n(\lambda_n-n)=0$, which in turn implies equality $-1+o(1)=|\varphi_n|^2(n-\lambda_n)^{-1}+o(1)$. Therefore $|\varphi_n|^2(\lambda_n-n)^{-1}=1+o(1)$. However, we know only that $|\varphi_n|^2\in l^1$ and so λ_n-n can tend to zero as an arbitrary l^1 sequence (exactly like $|\varphi_n|^2$).

Before we turn to some applications of Lemma 2.1 let us define by $\Sigma^b_{1/p}$ the set of all operators in H that are in $\Sigma_{1/p}$ and possess a band-type matrix in the basis $\{e_n\}_{n=1}^\infty$ of the eigenvectors of D. Consider a band-type operator R of the following form: $R=\sum_{k=-N}^N S^k\Lambda_k$, where Λ_k are diagonal operators and $S^{-k}:=S^{*^k}$, $k>0$. It will be proved below that $R\in\Sigma^b_{1/p}$ iff the diagonal operators Λ_k are defined by sequences from l^p (for $k=-N,\dots,N$).

PROPOSITION 2.3. *Let R be a compact operator that has a band matrix representation (r_{ij}) of width N in the basis $\{e_n\}$.*

Then $R\in\Sigma_{1/p}$ iff $r_{ii+k}=O(i^{-p})$, $k=-N,\dots,N$.

Proof. Let $R=\sum_{k=-N}^N\Lambda_kS^k$. If $r_{ii+k}=O(i^{-p})$ for all k, then $\Lambda_k\in\Sigma_{1/p}$ and so $R\in\Sigma_{1/p}$. On the other hand, if $R\in\Sigma_{1/p}$ and is given by the above formula, then $S^NR=\sum_{k=0}^{2N}\tilde\Lambda_kS^k$, where $\tilde\Lambda_k:=S^N\Lambda_{k-N}S^{*^N}$. Since $S^NR\in\Sigma_{1/p}$, it follows that its singular numbers satisfy estimates $s_k(S^NR)=O(k^{-p})$, and applying [18, Cor. 3.2] we know that the same estimate holds for its eigenvalues $\lambda_k(S^NR)=O(k^{-p})$. Hence $\tilde\Lambda_o\in\Sigma_{1/p}$ (note that the matrix S^NR is lower triangular). It follows that $\Lambda_{-N}\in\Sigma_{1/p}$ and also (by symmetry) that $\Lambda_N\in\Sigma_{1/p}$. Therefore $\sum_{k=-N+1}^{N-1}\Lambda_kS^k\in\Sigma_{1/p}$ and so on. The proof is complete. □

Below we also shall need the following elementary result.

PROPOSITION 2.4. *Let $W=I+K$, where K is a compact operator. If $0\in\sigma(W)$, then for any small $\varepsilon>0$ there exists a natural number $N=N(\varepsilon)$ such that $W+\varepsilon Q_N$ is invertible; here Q_N is the orthogonal projection in H onto $[e_1,\dots,e_N]$, and e_n is the orthonormal basis in H.*

Proof. Choose $\varepsilon>0$ so small that $W+\varepsilon I$ is invertible. Then $W+\varepsilon Q_N=[I+\varepsilon(Q_N-I)(W+\varepsilon I)^{-1}](W+\varepsilon I)$. Since $(W+\varepsilon I)^{-1}=(1+\varepsilon)^{-1}I+K_\varepsilon$, for a certain compact K_ε, it follows that $\|(Q_N-I)K_\varepsilon\|\to0$, as $N\to\infty$. Therefore $W+\varepsilon Q_N$ is invertible for N sufficiently large (as the product of two invertible operators). □

3. Successive diagonalization and the eigenvalues of Jacobi matrices. In this section a general method of successive diagonalization will be applied to J_g. The choice of the special class of J,s (in what follows we omit the letter g in J_g) is motivated by applications that will be given in section 4. The method seems strong enough (as will become clear from its proof) for possible applications to other classes of

Jacobi matrices with dominating diagonal. However, the method admits its simplest form for $\alpha > 2\beta + 1$, where α is the power order of the diagonal growth and β is the power order of the weight growth. We hope to consider this and other situations in a future paper. The procedure of the successive diagonalization method is the following.

Step 1. We look for self-adjoint diagonal operators $\tilde{J}_1,\ \Lambda_1$, where $\Lambda_1 \in \Sigma_1^b$ such that for anti-Hermitian $T_1 := \Lambda_1 S - S^*\Lambda_1$ we have

$$\tilde{J}_1(I + T_1) \doteq (I + T_1)J, \tag{3.1}$$

and the equality $\doteq$ means one modulo Σ_1^b. By this notation $J \doteq \Lambda^2 + c_1\Lambda + g(S + S^*)$. Assumptions on $\tilde{J}_1$ and T_1 imply (using (3.1)) that

$$\tilde{J}_1 - \Lambda^2 - c_1\Lambda \in \Sigma_1^b. \tag{3.2}$$

This can be easily checked by looking at the main diagonal entries of the difference of the left- and right-hand sides of (3.1). In turn, the first lower diagonal entries of the difference between the left- and right-hand sides of (3.1) are obtained by grouping all the terms of the difference that contain only S. Denoting this difference by A_1 we get that

$$A_1 \doteq \tilde{J}_1\Lambda_1 S - gS - \Lambda_1 S\Lambda^2 - c_1\Lambda_1 S\Lambda.$$

We search for Λ_1 satisfying the condition

$$\tilde{J}_1\Lambda_1 S - gS - \Lambda_1 S\Lambda^2 - c_1\Lambda_1 S\Lambda \in \Sigma_1^b. \tag{3.3}$$

It means

$$\Lambda_1(\tilde{J}_1 S - S\Lambda^2 - c_1 S\Lambda) - gS \in \Sigma_1^b. \tag{3.4}$$

If we define $\tilde{J}_1 := \Lambda^2 + c_1\Lambda$ (see (3.2)), then using the elementary identities

$$[\Lambda, S] = S,\ [\Lambda^2, S] = (2\Lambda - I)S, \tag{3.5}$$

the condition (3.4) can be written as

$$\Lambda_1[2\Lambda - I + c_1]S - gS \in \Sigma_1^b. \tag{3.6}$$

The inclusion (3.6) implies the choice

$$\Lambda_1 := \frac{g}{2}\Lambda^{-1}.$$

Observe that the first upper diagonal entries of the same difference are given by collecting all the terms of the difference that contain only S^*. Moreover, if B_1 denotes the above collection, then one can easily check (again using (3.1), and (3.6)) that

$$B_1 \doteq S^*[2\Lambda - I + c_1]\Lambda_1 - gS^*. \tag{3.7}$$

Hence $B_1 - A_1^* \in \Sigma_1^b$ and our choice of Λ_1 gives simultaneously that $B_1 \in \Sigma_1^b$. Concerning terms containing S^2 or S^{*^2} (in (3.1)), all of them belong to Σ_1^b already. The above choice of $\tilde{J}_1$ and Λ_1 implies the condition (2.1). Now we obtain the first three terms (in power asymptotic expansion) of $\lambda_n(J)$. Indeed, we rewrite (3.1) as

$$K_1 + \tilde{J}_1(I + T_1) = (I + T_1)J$$

for a certain $K_1 \in \Sigma_1^b$. Since $I + T_1$ is invertible (because $T_1^* = -T_1$), $J = (I + T_1)^{-1}[\tilde{J}_1 + K_1(I + T_1)^{-1}](I + T_1)$ and so

$$\{\lambda_n(J)\}_{n=1}^{\infty} = \sigma(J) = \sigma(\tilde{J}_1 + K_1(I + T_1)^{-1}).$$

Applying Lemma 2.1 with $R = K_1(I + T_1)^{-1}$ and $D = \tilde{J}_1 = \Lambda^2 + c_1\Lambda$ we have

$$\lambda_n(J) = n^2 + c_1 n + O(n^{-1}).$$

Due to the above choices and the form of R it is clear that $\|R^* e_n\| = O(\frac{1}{n})$.

Step 2. We look for new self-adjoint diagonal operators $\tilde{J}_2, \Lambda_1 \in \Sigma_1^b$, and $\Lambda_2 \in \Sigma_{1/2}^b$ such that for

$$T_1 := \Lambda_1 S - S^* \Lambda_1, \qquad T_2 := \Lambda_2 S^2 + S^{*^2} \Lambda_2$$

we have

$$\tilde{J}_2(I + T_1 + T_2) \doteq (I + T_1 + T_2)J, \tag{3.8}$$

where $\doteq$ means the equality modulo $\Sigma_{1/2}^b$.

Below diagonal operators Λ_j, band operators T_j and equalities $\doteq$ will have slightly different meanings at each of Steps 2 and 3. We keep the same notation for similar notions for reader convenience to avoid complicated symbols, hoping it will not lead to misunderstanding. Again in the new notation,

$$J \doteq \Lambda^2 + c_1\Lambda + c_2\Lambda^{-1} + g(S + S^*) + b_1(S\Lambda^{-1} + \Lambda^{-1}S^*).$$

Note that Λ_1 can be different from Λ_1 given in the first step, but for reader convenience we keep the same notation as above. Moreover, anticipating consequences of the above assumptions and (3.8) we also impose the following smoothness condition on Λ_1:

$$\Lambda_1 - S^*\Lambda_1 S \in \Sigma_{1/2}^b. \tag{3.9}$$

The operators Λ_j and $\tilde{J}_2$ will be chosen more precisely than in Step 1. Comparing the main diagonal entries of the difference between the left- and the right-hand side of (3.8) we put

$$\tilde{J}_2 = \Lambda^2 + c_1\Lambda + c_2\Lambda^{-1}. \tag{3.10}$$

One can check that the diagonal part of the above difference has the form $\tilde{J}_2 - \Lambda^2 - c_1\Lambda - c_2\Lambda^{-1} - g(\Lambda_1 - S^*\Lambda_1 S)$ modulo $\Sigma_{1/2}^b$. This clarifies the choice of $\tilde{J}_2$ in (3.10) and the condition (3.9)

As in the first step we collect all the terms of the difference of both sides of (3.8) containing only S and we obtain using (3.10)

$$\begin{aligned} A_2 &:= \tilde{J}_2\Lambda_1 S - (SD + \Lambda_1 SQ + \Lambda_2 S^2 DS^*) \doteq \tilde{J}_2\Lambda_1 S - (DS + \Lambda_1 SQ) \\ &= -DS + [2\Lambda - I + c_1]\Lambda_1 S, \end{aligned} \tag{3.11}$$

Again, as in the first step one can check that for the term B_2 which appears at the difference of both sides of (3.8) with terms containing only S^*,

$$A_2 \doteq B_2^*. \tag{3.12}$$

In turn collecting all terms, which appear on both sides of (3.8), containing only S^2, we have

$$C_2 := \tilde{J}_2\Lambda_2 S^2 - \Lambda_1 S^2 D - \Lambda_2 S^2 Q.$$

By the same token, for terms containing only S^{*^2} we have

$$D_2 := \tilde{J}_2 S^{*^2}\Lambda_2 + S^*\Lambda_1 D S^* - S^{*^2}\Lambda_2 Q.$$

Direct computation shows that again

$$D_2 \doteq C_2^*. \tag{3.13}$$

Combining (3.12) and (3.13) it is obvious that the condition $B_2 \in \Sigma_{1/2}^b$ (resp., $D_2 \in \Sigma_{1/2}^b$) holds provided one can check that $A_2 \in \Sigma_{1/2}^b$ (resp., $C_2 \in \Sigma_{1/2}^b$). First we are going to satisfy the condition $A_2 \in \Sigma_{1/2}^b$. Using (3.10) the claim $A_2 \in \Sigma_{1/2}^b$ is equivalent to $(2\Lambda - I + c_1)\Lambda_1 - D \in \Sigma_{1/2}^b$. The above equation suggests that the simplest choice of Λ_1 is

$$\Lambda_1 := D(2\Lambda - I + c_1 I)^{-1}. \tag{3.14}$$

Note that Λ_1 defined by (3.14) satisfies both conditions $\Lambda_1 \in \Sigma_1^b$ and $\Lambda_1 - S^*\Lambda_1 S \in \Sigma_{1/2}^b$. Put

$$\Lambda_2 := g\Lambda_1[4\Lambda + 2c_1 - 4]^{-1} \tag{3.15}$$

to ensure that $C_2 \in \Sigma_{1/2}^b$.

If for a certain k it happens that $2k - 2 + c_1 = 0$, then one can put $\Lambda_1 e_k = 0$. This makes no problem in checking all the above conditions. (Finite dimensional perturbations in the canonical basis are allowed.) Since $\Lambda_1 \in \Sigma_1^b$, we see that $\Lambda_2 \in \Sigma_{1/2}$. Using the formulas (3.14) and (3.15) one can verify that $C_2 \in \Sigma_{1/2}^b$. Since we can reverse the implication in the above reasoning, the desired equality (3.8) holds. Therefore there is $K_2 \in \Sigma_{1/2}^b$ such that

$$K_2 + \tilde{J}_2(I + T_1 + T_2) = (I + T_1 + T_2)J.$$

By Proposition 2.4 we can find a finite dimensional projection Q such that $I + T_1 + T_2 + \varepsilon Q$ is invertible for some $\varepsilon > 0$. Since $\tilde{J}_2 Q - QJ$ is a finite rank band matrix, the above equality can be written as

$$\tilde{K}_2 + \tilde{J}_2(I + T_1 + T_2 + \varepsilon Q) = (I + T_1 + T_2 + \varepsilon Q)J$$

for a certain $\tilde{K}_2 \in \Sigma_{1/2}^b$. We can repeat the reasoning given at the end of Step 1 and applying Lemma 2.1 we have

$$\lambda_n(J) = n^2 + c_1 n + c_2 n^{-1} + O(n^{-2}). \tag{3.16}$$

Step 3. One might think that the method of diagonalization presented in the above steps carries over easily to the third step. This is not the case, however, as we shall see below. At this step we look again for new self-adjoint diagonal operators

$\Lambda_j \in \Sigma_{1/j}$ $(j = 1, 2, 3)$ and $\tilde{\Lambda}_2 \in \Sigma_{1/2}$ such that for $T_j = \Lambda_j S^j - S^{*^j}\Lambda_j$ $(j = 1, 3)$ and $T_2 = \Lambda_2 S^2 + S^{*^2}\tilde{\Lambda}_2$,

$$\tilde{J}_3(I + T_1 + T_2 + T_3) - (I + T_1 + T_2 + T_3)J \in \Sigma_{1/3}, \tag{3.17}$$

where $\tilde{J}_3$ is a diagonal operator given by

$$\tilde{J}_3 = \Lambda^2 + c_1\Lambda + c_2\Lambda^{-1} + \Delta \quad \text{for a certain } \Delta \in \Sigma_{1/2}. \tag{3.18}$$

Note that at this step the diagonalizing matrix $I+T_1+T_2+T_3$ not only has one upper and one lower diagonal more (than in Step 2) but also loses its symmetry with respect to the main diagonal. Using (3.17) we will determine (similarly as in the above steps) $\Lambda_j, \tilde{\Lambda}_2$, and Δ. In what follows $\doteq$ stands for the equality modulo $\Sigma_{1/3}$. Comparing the diagonal entries of both sides of (3.17) we can write

$$\tilde{J}_3 \doteq \Lambda^2 + c_1\Lambda + c_2\Lambda^{-1} + c_3\Lambda^{-2} + \Lambda_1 SDS^* - S^*\Lambda_1 SD.$$

Following the ideas presented in the above steps we also need some extra smoothing assumptions on Λ_1 and Λ_2. Namely, in what follows we assume that

(3.19)
(a) $[\Lambda_1, S] \in \Sigma^b_{1/2}$, (b) $[\Lambda_2, S] \in \Sigma^b_{1/3}$, (c) $[\tilde{\Lambda}_2, S] \in \Sigma^b_{1/3}$, (d) $\tilde{\Lambda}_2 - \Lambda_2 \in \Sigma^b_{1/3}$.

Substituting $D \doteq gI + b_1\Lambda^{-1} + b_2\Lambda^{-2}$ into the above formula for $\tilde{J}_3$ we obtain (using (3.19) (a)) that $\Delta = c_3\Lambda^{-2} + g(\Lambda_1 - S^*\Lambda_1 S)$. Then we can express $\tilde{J}_3$ as

$$\tilde{J}_3 := \Lambda^2 + c_1\Lambda + c_2\Lambda^{-1} + c_3\Lambda^{-2} + g(\Lambda_1 - S^*\Lambda_1 S). \tag{3.20}$$

Similarly as in the second step one can verify (by using (3.19) (c), (d) and assumptions $\Lambda_j \in \Sigma_{1/j}$ $(j = 1, 2, 3)$, $\tilde{\Lambda}_2 \in \Sigma_{1/2}$) that the coefficients A_3 and E_3 (resp., $\tilde{A}_3$ and $\tilde{B}_3$) of the difference of the left- and right-hand sides of (3.16) at S and S^3 (resp., S^* and S^{*^3}) satisfy $\tilde{A}_3 \doteq A^*_3$, $\tilde{B}_3 \doteq E^*_3$. As calculations to be given below show, the situation for the coefficients at S^2 and S^{*^2} terms is not so symmetric. This is exactly the reason for introducing two slightly different diagonal matrices Λ_2 and $\tilde{\Lambda}_2$. Below we shall see that $\tilde{\Lambda}_2 \doteq \Lambda_2$. Note that (3.19) (c) follows from (3.19) (b) and (d). In what follows we shall find formulas for Λ_j, $j = 1, 2, 3$, and $\tilde{\Lambda}_2$. These formulas will be derived from the basic equation (3.17). First we look for Λ_1, Λ_2. Since $\Lambda_j \in \Sigma_{1/j}$ and due to the definition of $\tilde{J}_3$, and J_g, we seek $\Lambda_j (j = 1, 2)$ in the form

$$\Lambda_1 := a\Lambda^{-1} + b\Lambda^{-2} + c\Lambda^{-3}, \tag{3.21}$$

$$\Lambda_2 := d\Lambda^{-2} + e\Lambda^{-3} \tag{3.22}$$

for some constants a, b, c, d, e to be determined below. The ansatz for Λ_1, Λ_2 follows straightforwardly from simple analysis of (3.24) and (3.26) (see below). The proper formula for the matrix $\tilde{\Lambda}_2$ will be given by formula (3.30). It turns out that (3.17) defines the above constants uniquely. Indeed, by looking at terms of (3.17) containing only S we see that

$$(\Lambda^2 + c_1\Lambda + c_2\Lambda^{-1})\Lambda_1 S - SD - \Lambda_1 S(\Lambda^2 + c_1\Lambda + c_2\Lambda^{-1}) - \Lambda_2 S^2 DS^* \doteq 0. \tag{3.23}$$

Using again the commutation formulas (3.5) and $[\Lambda^{-1}, S] = -S\Lambda^{-1}(\Lambda + I)^{-1}$, one can check that (3.23) is equivalent to

$$\Lambda_1[2\Lambda + (c_1 - 1)I] - gI - b_1\Lambda^{-1} - (b_1 + b_2)\Lambda^{-2} \doteq 0. \tag{3.24}$$

In turn by collecting all terms of (3.17) containing only S^2 we have

$$(\Lambda^2 + c_1\Lambda)\Lambda_2 S^2 - \Lambda_1 S^2 D - \Lambda_2 S^2(\Lambda^2 + c_1\Lambda) \doteq 0. \tag{3.25}$$

Since $S^2\Lambda^{-1} \doteq (\Lambda^{-1} + 2\Lambda^{-2})S^2$ one can verify that (3.25) is equivalent to

$$\Lambda_2[4(\Lambda - I) + 2c_1 I] \doteq \Lambda_1(gI + b_1\Lambda^{-1}). \tag{3.26}$$

By substituting expressions (3.21) into (3.24) one gets in particular that $a = g/2$. The explicit solution of (3.24), (3.26) is given by the following formulas:

$$\Lambda_1 := [gI + b_1\Lambda^{-1} + (b_1 + b_2)\Lambda^{-2}][2\Lambda + c_1 - 1]^{-1} \quad (\text{modulo } \Sigma_{1/4}),$$

$$\Lambda_2 := \Lambda_1(gI + b_1\Lambda^{-1})[4\Lambda - (4 - 2c_1)I]^{-1} \quad (\text{modulo } \Sigma_{1/4}). \tag{3.27}$$

Again in formula (3.27) possible two zero of the denominators are inessential and can be removed similarly as in Step 2 (both diagonal operators are defined to be equal to zero for critical indexes). One can compute Λ_1, Λ_2 using (3.27) in the form of (3.21), (3.22), respectively; however, it is necessary only if we want to obtain an approximate formula for the eigenvectors of J.

What about $\tilde{\Lambda}_2$ and Λ_3? By grouping all terms with S^{*^2} of (3.17) we have

$$(\Lambda^2 + c_1\Lambda)S^{*^2}\tilde{\Lambda}_2 + S^*\Lambda_1 D S^* - S^{*^2}\tilde{\Lambda}_2(\Lambda^2 + c_1\Lambda) \doteq 0. \tag{3.28}$$

Using the standard commutator formulas for $[S^{*^2}, \Lambda^2]$ and $[S^{*^2}, \Lambda]$ we rewrite (3.28) into the equivalent form

$$\tilde{\Lambda}_2[4(\Lambda - I) + 2c_1 I] \doteq \Lambda_1(gI + b_1\Lambda^{-1}) + g(S\Lambda_1 S^* - \Lambda_1). \tag{3.29}$$

Thus

$$\tilde{\Lambda}_2 \doteq \Lambda_1[4(\Lambda - I) + 2c_1 I]^{-1}(gI + b_1\Lambda^{-1}) = \Lambda_2 \tag{3.30}$$

because $g(S\Lambda_1 S^* - \Lambda_1) \in \Sigma^b_{1/2}$. However, (3.29) is not sufficiently precise: substitution of Λ_2 instead of $\tilde{\Lambda}_2$ into (3.28) contradicts it. Indeed, due to the multiplier Λ in the left-hand side of (3.29) we should calculate $\tilde{\Lambda}_2$ modulo $\Sigma_{1/4}$ but $\tilde{\Lambda}_2 = \Lambda_2$ modulo $\Sigma_{1/3}$ only because $(\Lambda_1 - S\Lambda_1 S^*)$ belongs to $\Sigma_{1/2}$ and not to $\Sigma_{1/3}$ in general! Moreover, one can choose $\tilde{\Lambda}_2 = \Lambda_2$ iff $\Lambda_1 - S\Lambda_1 S^* \in \Sigma_{1/3}$.

Now by collecting all terms with S^3 we have

$$(\Lambda^2 + c_1\Lambda)\Lambda_3 S^3 - \Lambda_2 S^3 b - \Lambda_3 S^3(\Lambda^2 + c_1\Lambda) \doteq 0. \tag{3.31}$$

Applying the relations $[S^3, \Lambda^2] = -3(2\Lambda - 3)S^3$, $[S^3, \Lambda] = -3S^3$, we see that (3.31) is equivalent to $6\Lambda_3\Lambda \doteq g\Lambda_2$. The last equation has the solution $\Lambda_3 := g\Lambda_2(6\Lambda)^{-1}$. In turn by bringing together all terms of (3.17) that contain S^{*^3}, one easily finds that the sum of these terms must belong to $\Sigma^b_{1/3}$ because $\tilde{\Lambda}_2 \doteq \Lambda_2$ and $\tilde{\Lambda}_2 - S\tilde{\Lambda}_2 S^* \in \Sigma_{1/3}$. Note

that the above explicit formulas for Λ_j and $\tilde{\Lambda}_2$ also imply that all the requirements concerning $\Lambda_j (j = 1, 2, 3)$ and $\tilde{\Lambda}_2$ (in particular assumptions (3.19) (a), (b), (c), (d)) are satisfied. Finally, since all the above equations are equivalent to suitable ones, we check that the above choice of $T_j (j = 1, 2, 3)$ leads to the desired relation (3.17). By repeating the reasoning given at the end of Step 1 we obtain the following more precise asymptotics of the eigenvalues of J.

THEOREM 3.1. *The asymptotic formula for $\lambda_n(J_g)$ is given by*

$$\lambda_n(J_g) = n^2 + c_1 n + c_2 n^{-1} + c_3 n^{-2} + \frac{g^2}{2} n^{-2} + O(n^{-3}). \tag{3.32}$$

Proof. It is enough to use the formula (3.20) and observe that $\Lambda_1 - S^* \Lambda_1 S \doteq \frac{g}{2} \Lambda^{-2}$ (use here (3.27) for the matrix Λ_1). □

Remark 3.2. Surely one can continue this process of successive diagonalization of J. We have stopped after making three steps because the weights λ_n and the diagonal q_n are known only up to $O(n^{-3})$.

Moreover, the idea of successive diagonalization can be applied for other classes of Jacobi operators with uniformly separated discrete spectrum (at least provided the main diagonal has more than one order higher-power-like growth compared to off-diagonal terms). Observe that one can use the same method to approximate the eigenvectors f_k of J for large indices k. In the next section we give an application of Theorem 3.1 to spectral analysis of energy spectrum of a molecule in the homogeneous electric field, which motivated our special choice of J entries.

Remark 3.3. It seems somewhat surprising that the influence of the zero order term $g(S + S^*)$ in J causes only the minus second order variation of the eigenvalues (see [36]).

4. Application to asymptotics of the energy spectrum of a molecule in a homogeneous electric field. This section contains an application of Theorem 3.1 to a polar molecule of symmetric top type in a homogeneous electric field [6], [7], [17]. The Hamiltonian of a symmetric pendulum has the form

$$\hat{H}_o = \frac{\hat{L}^2}{2I} + \left(\frac{1}{2I_3} - \frac{1}{2I} \right) \hat{L}^2_{z'},$$

where $\hat{L}^2$ (resp., $\hat{L}^2_z$) are the operators of the full moment (resp., its projection on the z'-axis) acting in $L^2(\mathbb{R}^3)$, and I, I_3 are the corresponding inertia moments. It turns out that $\hat{H}_o$ has eigenfunctions ϕ^n_{mk} coinciding (up to the normalization constant) with the generalized spherical function $D^n_{mk}(a, b, c)$, where $(a, b, c) \in (0, 2\pi)x(0, \pi)x(0, 2\pi)$. More precisely,

$$\phi^n_{mk}(a, b, c) = \sqrt{(2n + 1)/8\pi^2} D^n_{mk}(a, b, c),$$

$k, m = -n, \ldots, 0, \ldots, n$. The eigenvalues of $\hat{H}_o$ corresponding to ϕ^n_{mk} are given by

$$E_{n,|k|} = \frac{h^2}{2I} n(n + 1) + \frac{h^2}{2} \left(\frac{1}{I_3} - \frac{1}{I} \right) k^2;$$

here h is the Planck constant. Let $\hat{V} = -dE(t)\cos b$ be the potential energy of the perturbation by the electric field $E(t)$ directed along the z-axis (in the fixed coordinate system (x, y, z) connected with the pendulum), where d is the dipole moment of the

molecule and b is the angle between the axes z and z'. For more physical details of the model, see [17].

It turns out that the full Hamiltonian $\hat{H} := \hat{H}_o + \hat{V}$ can be reduced in the basis ϕ^n_{mk} to a Jacobi matrix $J_{\tilde{g}}$, where $\tilde{g} = dE(t)$ is the constant of interaction for fixed time t. More precisely, for fixed indices $k, m \in \mathbb{Z}$, the diagonal and the weights of $J_{\tilde{g}}$, after straightforward calculations in the basis ϕ^n_{km}, are given by

$$q_n = \frac{h^2}{2I} n(n+1) + \frac{h^2}{2}\left(\frac{1}{I_3} - \frac{1}{I}\right) k^2 - \tilde{g}\frac{mk}{n(n+1)},$$

$$\lambda_n = \frac{-\tilde{g}}{(n+1)(2n+1)} |[((n+1)^2 - m^2)((n+1)^2 - k^2)]|^{1/2}.$$

These calculations use the explicit form of the matrix elements of $\hat{V}$ in the basis ϕ^n_{mk} as the Clebsch–Gordan coefficients c^{rt}_{klmn}; see [31]. Namely, we have

$$(\hat{V}\phi^n_{mk}, \phi^p_{rs}) = -dE(t) c^{nm}_{1r0m} c^{nk}_{1p0k} \delta_{mr} \delta_{ks}.$$

Although the study of this type of molecule in an electric field is of special interest, our goal in this paper is rigorous mathematical analysis of related classes of unbounded Jacobi matrices only. It is clear that $J_{\tilde{g}}$ is of the form considered in Theorem 3.1. Applying Theorem 3.1 we immediately obtain the following result.

THEOREM 4.1. *The asymptotics as $n \to \infty$ of the energy spectrum $E_n(k, m)$ (for fixed integer numbers k, m) are given by*

$$E_n(k,m) = \frac{h^2}{2I} n^2 + \frac{h^2}{2I} n + \frac{h^2}{2}\left(\frac{1}{I_2} - \frac{1}{I}\right) k^2 + \left(\frac{\tilde{g}mk + I\frac{\tilde{g}^2}{(4h^2)}}{n^2}\right) + O\left(\frac{1}{n^3}\right), \ n \to \infty.$$

Acknowledgments. We thank Bozena Skoczylas for her valuable help in preparation of the manuscript. We appreciate A. Laptev's useful comments and information about the paper by Rosenblum. We also are grateful to E. Tur for helpful discussion related to the physical aspect of the problem.

REFERENCES

[1] N. AKHIEZER, *The Classical Moment Problem*, Oliver and Boyd, London, 1965.
[2] W. VAN ASSCHE, *Weighted zero distribution for polynomials orthogonal on an infinite interval*, SIAM J. Math. Anal., 16 (1985), pp. 1317–1334.
[3] W. VAN ASSCHE, *Asymptotics for Orthogonal Polynomials*, Lecture Notes in Math. 1265, Springer-Verlag, Berlin, 1987.
[4] YU. M BEREZANSKII, *Expansions in Eigenfunctions of Selfadjoint Operators*, Naukova Dumka, Kiev, 1965 (in Russian).
[5] M. BIRMAN AND M. SOLOMYAK, *Spectral Theory of Selfadjoint Operators in Hilbert Space*, D. Reidel, Dordrecht, The Netherlands, 1987.
[6] P. A. BRAUN, E. V. EGOROV, AND G. P. MIROSHNICHENKO, *Effects of molecule orientation in external fields*, in Interaction of Atoms and Molecules with Electromagnetic Field, LGU, Leningrad, 1987, pp. 200–217 (in Russian).
[7] P. A. BRAUN AND A. A. KISELEV, *Introduction to the Theory of Molecular Spectra*, LGU, Leningrad, 1983 (in Russian).
[8] T. S. CHIHARA, *An Introduction to Orthogonal Polynomials*, Math. Appl. 13, Gordon and Breach, New York, London, Paris, 1978.
[9] H. L. CYCON, R. G. FROESE, W. KIRSCH, AND B. SIMON, *Schroedinger Operators*, Springer, Berlin, Heidelberg, New York, 1987.

[10] J. Delsarte, *Sur une extension de la formulate de Taylor*, J. Math. Pures Appl., 17 (1938), pp. 213–230.
[11] J. Delsarte and J. Lions, *Transmutations d'operateurs differentieles dans le domaine complexe*, Comment. Math. Helv., 32 (1957), pp. 113–128.
[12] J. Dombrowski, *Tridiagonal matrix representations of cyclic selfadjoint operators*, Pacific J. Math., 114 (1984), pp. 325–334.
[13] J. Dombrowski, *Tridiagonal matrix representations of cyclic selfadjoint operators* II, Pacific J. Math., 120 (1985), pp. 47–53.
[14] J. Dombrowski, *Spectral measures corresponding to orthogonal polynomials with unbounded recurrence coefficients*, Constr. Approx., 5 (1989), pp. 371–381.
[15] J. Dombrowski and S. Pedersen, *Spectral measures, and Jacobi matrices related to Laguerre-type systems of orthogonal polynomials*, Constr. Approx., 13 (1997), pp. 421–433.
[16] J. Edward, *Spectra of Jacobi matrices, differential equations on the circle, and the* $su(1,1)$ *Lie algebra*, SIAM J. Math. Anal., 24 (1993), pp. 824–831.
[17] A. V. Gaponov, Yu. N. Demkov, V. A. Protopopova, and V. M. Fain, *Stark-effect for rotating levels of molecules in strong fields*, Optics Spectroscopy, 19 (1965), pp. 501–506.
[18] I. Gohberg and M. G. Krein, *Introduction to the Theory of Linear Nonselfadjoint Operators in Hilbert Space*, AMS, Providence, RI, 1969.
[19] J. Janas and S. Naboko, *Multithreshold spectral phase transition examples in a class of unbounded Jacobi matrices*, in Recent Advances in Operator Theory, Oper. Theory Adv. Appl., 124, Birkhäuser-Verlag, Basel, 2001, pp. 267–285.
[20] J. Janas and S. Naboko, *On the point spectrum of some Jacobi matrices*, J. Operator Theory, 40 (1998), pp. 113–132.
[21] J. Janas and S. Naboko, *Jacobi matrices with absolutely continuous spectrum*, Proc. Amer. Math. Soc., 127 (1999), pp. 791–800.
[22] J. Janas and S. Naboko, *Jacobi matrices with power like weights*, J. Funct. Anal., 166 (1999), pp. 218–243.
[23] J. Janas and S. Naboko, *Asymptotics of generalized eigenvectors for unbounded Jacobi matrices with power-like weights, Pauli matrices commutation relations and Cesaro averaging*, in Differential Operators and Related Topics, Vol. 1, Oper. Theory Adv. Appl. 117, Birkhaüser-Verlag, Basel, 2000, pp. 165–186.
[24] J. Janas and S. Naboko, *Spectral analysis of selfadjoint Jacobi matrices with periodically perturbated entries*, J. Funct. Anal., 191 (2002), pp. 318–342.
[25] J. Janas and S. Naboko, *Spectral properties of selfadjoint Jacobi matrices coming from birth and death processes*, in Recent Advances in Operator Theory and Related Topics, Oper. Theory Adv. Appl. 127, Birkhäuser-Verlag, Basel, 2001, pp. 387–397.
[26] J. Janas and S. Naboko, *Criteria for semiboundedness in a class of unbounded Jacobi operators*, St. Petersburg Math. J., 14 (2003), pp. 156–165.
[27] J. Janas and M. Moszynski, *Alternative approaches to the absolute continuity of Jacobi matrices*, Int. Equat. Oper. Theory, 43 (2002), pp. 397–413.
[28] T. Kato, *Perturbation Theory for Linear Operators*, 2nd ed., Springer, Berlin, Heidelberg, New York, 1980.
[29] L. D. Landau and E. M. Lifshitz, *Quantum Mechanics*, Mir, Moscow, 1972.
[30] V. M. Levitan, *Generalized Shift Operators and Their Applications*, Nauka, Moscow, 1972.
[31] K. M. Ng, C. F. Lo, and K. L. Liu, *Exact eigenstates of the density-dependent Jaynes-Cummings model with the counter-rotating term*, Phys. A, 275 (2000), pp. 463–474.
[32] V. A. Marchenko, *Sturm-Liouville Operators and Their Applications*, Naukova Dumka, Kiev, 1977.
[33] M. A. Naimark, *Linear Differential Operators. Part* II: *Linear Differential Operators in Hilbert Space,* Ungar, New York, 1968.
[34] G. V. Rosenblum, *Almost similarity of operators and spectral asymptotics of pseudodifferential operators on the unit circle*, Transl. Moscow Math. Soc., 2 (1979), pp. 57–82.
[35] G. Stolz, *Spectral theory for slowly oscillating potentials* I. *Jacobi matrices*, Manuscripta Math., 84 (1994), pp. 245–260.
[36] E. A. Tur, *Weyl's function continued fraction representation for Jacobi matrix and its applications*, in Proceedings of the Banach Center Research Seminar IV, Warsaw, Poland, 2000, pp. 21–26.

SIAM J. MATH. ANAL.
Vol. 36, No. 2, pp. 659–677

A SHARP DECAY ESTIMATE FOR POSITIVE NONLINEAR WAVES*

ALBERTO BRESSAN† AND TONG YANG‡

Abstract. We consider a strictly hyperbolic, genuinely nonlinear system of conservation laws in one space dimension. A sharp decay estimate is proved for the positive waves in an entropy weak solution. The result is stated in terms of a partial ordering among positive measures, using symmetric rearrangements and a comparison with a solution of Burgers's equation with impulsive sources.

Key words. hyperbolic conservation laws, positive nonlinear waves, Burgers's equation

AMS subject classifications. 35L65

DOI. 10.1137/S0036141003427774

1. Introduction.

Consider a strictly hyperbolic system of n conservation laws

$$u_t + f(u)_x = 0 \tag{1.1}$$

and assume that all characteristic fields are genuinely nonlinear. Call $\lambda_1(u) < \cdots < \lambda_n(u)$ the eigenvalues of the Jacobian matrix $A(u) \doteq Df(u)$. We shall use bases of left and right eigenvectors $l_i(u)$, $r_i(u)$ normalized so that

$$\nabla\lambda_i(u)\, r_i(u) \equiv 1\,, \qquad\qquad l_i(u)\, r_j(u) = \begin{cases} 1 & \text{if} \quad i = j, \\ 0 & \text{if} \quad i \neq j. \end{cases} \tag{1.2}$$

Given a function $u : \mathbb{R} \mapsto \mathbb{R}^n$ with small total variation following [BC], [B], one can define the measures μ^i of i-waves in u as follows. Since $u \in BV$, its distributional derivative $D_x u$ is a Radon measure. We define μ^i as the measure such that

$$\mu^i \doteq l_i(u) \cdot D_x u \tag{1.3}$$

restricted to the set where u is continuous, while, at each point x where u has a jump, we define

$$\mu^i\big(\{x\}\big) \doteq \sigma_i\,, \tag{1.4}$$

where σ_i is the strength of the i-wave in the solution of the Riemann problem with data $u^- = u(x-)$, $u^+ = u(x+)$. In accordance with (1.2), if the solution of the Riemann problem contains the intermediate states $u^- = \omega_0, \omega_1, \ldots, \omega_n = u^+$, the strength of the i-wave is defined as

$$\sigma_i \doteq \lambda_i(\omega_i) - \lambda_i(\omega_{i-1}). \tag{1.5}$$

Observing that

$$\sigma_i = l_i(u^+) \cdot (u^+ - u^-) + O(1) \cdot |u^+ - u^-|^2,$$

*Received by the editors May 13, 2003; accepted for publication (in revised form) October 24, 2003; published electronically August 27, 2004.
http://www.siam.org/journals/sima/36-2/42777.html

†S.I.S.S.A., Via Beirut 4, Trieste 34014, Italy (bressan@sissa.it). The research of this author was supported by Italian M.I.U.R. research project 2002017219, "Equazioni iperboliche e paraboliche non lineari."

‡Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong (Matyang@math.cityu.edu.hk). The research of this author was supported by CityU Direct Allocation Grant 7100198.

we can find a vector $l_i(x)$ such that

$$\big|l_i(x) - l_i\big(u(x+)\big)\big| = \mathcal{O}(1)\cdot\big|u(x+) - u(x-)\big|, \tag{1.6}$$

$$\sigma_i = l_i(x)\cdot\big(u(x+) - u(x-)\big). \tag{1.7}$$

We can thus define the measure μ^i equivalently as

$$\mu^i \doteq l_i\cdot D_x u\,, \tag{1.8}$$

where $l_i(x) = l_i\big(u(x)\big)$ at points where u is continuous, while $l_i(x)$ is some vector which satisfies (1.6)–(1.7) at points of jump. For all $x\in\mathbb{R}$ there holds

$$\Big|l_i(x) - l_i\big(u(x)\big)\Big| = \mathcal{O}(1)\cdot\big|u(x+) - u(x-)\big|\,. \tag{1.9}$$

We call μ^{i+}, μ^{i-}, respectively, the positive and negative parts of μ^i, so that

$$\mu^i = \mu^{i+} - \mu^{i-}, \qquad\qquad |\mu^i| = \mu^{i+} + \mu^{i-}. \tag{1.10}$$

It is our purpose to prove a sharp estimate on the decay of the density of the measures μ^{i+}. This will be achieved by introducing a partial ordering within the family of positive Radon measures. In the following, $meas(A)$ denotes the Lebesgue measure of a set A.

Definition 1. *Let μ,μ' be two positive Radon measures. We say that $\mu\preceq\mu'$ if and only if*

$$\sup_{meas(A)\leq s}\mu(A) \;\leq\; \sup_{meas(B)\leq s}\mu'(B) \qquad\qquad \text{for every } s>0\,. \tag{1.11}$$

In some sense, the above relation means that μ' is more singular than μ. Namely, it has a greater total mass, concentrated on regions with higher density. Notice that the usual order relation

$$\mu\leq\mu' \qquad \text{if and only if} \qquad \mu(A)\leq\mu'(A) \quad \text{for every} \quad A\subset\mathbb{R} \tag{1.12}$$

is much stronger. Of course $\mu\leq\mu'$ implies $\mu\preceq\mu'$, but the converse does not hold.

Following [BC], [B], together with the measures μ^i, we define the Glimm functionals

$$V(u) \doteq \sum_i |\mu^i|(\mathbb{R})\,, \tag{1.13}$$

$$Q(u) \doteq \sum_{i<j}\big(|\mu^j|\otimes|\mu^i|\big)\big\{(x,y)\,;\;x<y\big\} + \sum_i\big(\mu^{i-}\otimes|\mu^i|\big)\big\{(x,y)\,;\;x\neq y\big\}\,. \tag{1.14}$$

Now let $u=u(t,x)$ be an entropy weak solution of (1.1). If the total variation of u is small and the constant C_0 is large enough, it is well known that the quantities

$$Q(t) \doteq Q\big(u(t)\big)\,, \qquad\qquad \Upsilon(t) \doteq V\big(u(t)\big) + C_0\,Q\big(u(t)\big) \tag{1.15}$$

are nonincreasing in time. The decrease in Q controls the amount of interaction, while the decrease in Υ controls both the interaction and the cancellation in the solution.

An accurate estimate on the measure μ^{i+}_t of positive i-waves in $u(t,\cdot)$ will be obtained by a comparison with a solution of Burgers's equation with source terms.

THEOREM 1. *For some constant κ and for every small BV solution $u = u(t,x)$ of the system* (1.1) *the following holds. Let $w = w(t,x)$ be the solution of the scalar Cauchy problem with impulsive source term*

$$w_t + (w^2/2)_x = -\kappa\,\mathrm{sgn}(x)\cdot\frac{d}{dt}Q\big(u(t)\big)\,, \tag{1.16}$$

$$w(0,x) = \mathrm{sgn}(x)\cdot\sup_{meas(A)<2|x|}\frac{\mu_0^{i+}(A)}{2}\,. \tag{1.17}$$

Then, for every $t \geq 0$,

$$\mu_t^{i+} \preceq D_x w(t)\,. \tag{1.18}$$

As shown in the next section, the initial data in (1.17) represents the *odd rearrangement* of the function $v_i(x) \doteq \mu_0^{i+}\big(]-\infty,x]\big)$. The above theorem improves the earlier estimate derived in [BC]. For a scalar conservation law with strictly convex flux, a classical decay estimate was proved by Oleinik [O]. In the case of genuinely nonlinear systems, results related to the decay of nonlinear waves were also obtained in [GL], [L1], [L2], [L3], [BG]. An application of the present analysis can be found in [BY], where Theorem 1 plays a key role in the estimate of the rate of convergence of vanishing viscosity approximations.

2. Lower semicontinuity. Let μ be a positive Radon measure on $\mathbb{R}$, so that $\mu \doteq D_x v$ is the distributional derivative of some bounded, nondecreasing function $v : \mathbb{R} \mapsto \mathbb{R}$. We can decompose

$$\mu = \mu^{\mathrm{sing}} + \mu^{ac}$$

as the sum of a singular and an absolutely continuous part, w.r.t. Lebesgue measure. The absolutely continuous part corresponds to the usual derivative $z \doteq v_x$, which is a nonnegative $\mathbf{L}^1$ function defined at a.e. point. We shall denote by $\hat{z}$ the *symmetric rearrangement* of z, i.e., the unique even function such that

$$\hat{z}(x) = \hat{z}(-x)\,, \qquad \hat{z}(x) \geq \hat{z}(x') \quad \text{if} \quad 0 < x < x'\,, \tag{2.1}$$

$$\mathrm{meas}\Big(\big\{x\,;\ \hat{z}(x) > c\big\}\Big) = \mathrm{meas}\Big(\big\{x\,;\ z(x) > c\big\}\Big) \qquad \text{for every } c > 0\,. \tag{2.2}$$

Moreover, we define the *odd rearrangement* of v as the unique function $\hat{v}$ such that (Figure 1)

$$\hat{v}(-x) = -\hat{v}(x)\,, \qquad \hat{v}(0+) = \frac{1}{2}\mu^{\mathrm{sing}}(\mathbb{R})\,, \tag{2.3}$$

$$\hat{v}(x) = \hat{v}(0+) + \int_0^x z(y)\,dy \qquad \text{for } x > 0\,. \tag{2.4}$$

By construction, the function $\hat{v}$ is convex for $x < 0$ and concave for $x > 0$.

The relation between the odd rearrangement $\hat{v}$ and the partial ordering (1.10) is clarified by the following result, which is an easy consequence of the definitions.

PROPOSITION 1. *Let $\mu = D_x v$ and $\mu' = D_x v'$ be positive Radon measures. Call $\hat{v}, \hat{v}'$ the odd rearrangements of v, v', respectively. Then $\mu \preceq D_x\hat{v} \preceq \mu$ and moreover*

$$\hat{v}(x) = \mathrm{sgn}(x)\cdot\sup_{meas(A)\leq 2|x|}\frac{\mu(A)}{2}\,, \tag{2.5}$$

$$\mu \preceq \mu' \qquad \textit{if and only if} \qquad \hat{v}(x) \leq \hat{v}'(x) \quad \textit{for all } x > 0\,. \tag{2.6}$$

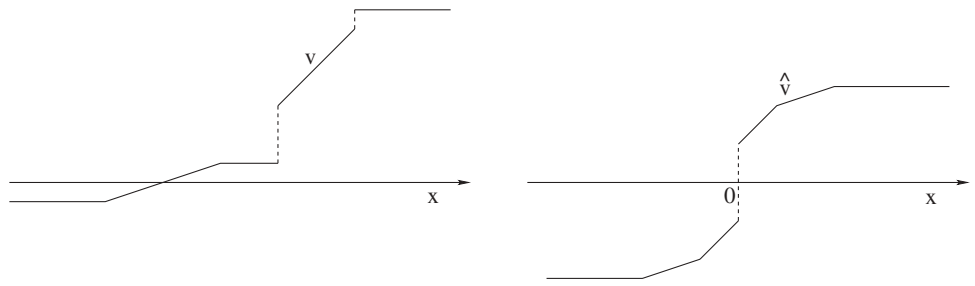

FIG. 1.

Two more results will be used in what follows. By the restriction of a measure μ to a set J, we mean the measure

$$(\mu\lfloor J)(A) \doteq \mu(A \cap J)\,.$$

PROPOSITION 2. *Let μ, μ' be positive measures. Consider any finite partition $\mathbb{R} = J_1 \cup \cdots \cup J_N$. If the restrictions of μ, μ' to each set J_ℓ satisfy $\mu\lfloor J_\ell \preceq \mu'\lfloor J_\ell$, then $\mu \preceq \mu'$.*

PROPOSITION 3. *Assume that $\mu \preceq D_s w$ for some nondecreasing odd function w. If $|\mu^\sharp - \mu|(\mathbb{R}) \leq \varepsilon$, then*

$$\mu^\sharp \preceq D_s \left[w + \mathrm{sgn}(s) \cdot \frac{\varepsilon}{2} \right] .$$

The next result is concerned with the lower semicontinuity of the partial ordering $\preceq$ w.r.t. weak convergence of measures.

PROPOSITION 4. *Consider a sequence of measures μ_ν converging weakly to a measure μ. Assume that the positive parts satisfy $\mu_\nu^+ \preceq D w_\nu$ for some odd, nondecreasing functions $s \mapsto w_\nu(s)$, concave for $s > 0$. Let w be the odd function such that*

$$w(s) \doteq \liminf_{\nu \to \infty} w_\nu(s) \qquad\qquad \text{for } \ s > 0\,.$$

Then the positive part of μ satisfies

$$\mu^+ \preceq D_s w\,. \tag{2.7}$$

Proof. By possibly taking a subsequence, we can assume that $w_\nu(s) \to w(s)$ for all $s \neq 0$. Moreover, we can assume the weak convergence

$$\mu_\nu^+ \rightharpoonup \tilde{\mu}^+\,, \qquad\qquad \mu_\nu^- \rightharpoonup \tilde{\mu}^-$$

for some positive measures $\tilde{\mu}^+$, $\tilde{\mu}^-$. We thus have

$$\mu = \tilde{\mu}^+ - \tilde{\mu}^-\,, \qquad \mu^+ \leq \tilde{\mu}^+\,, \qquad \mu^- \leq \tilde{\mu}^-\,. \tag{2.8}$$

By (2.8) it suffices to prove that $\tilde{\mu}^+ \preceq D_s w$, i.e.,

$$\mathrm{meas}\,(A) \leq 2s \qquad \Longrightarrow \qquad \tilde{\mu}^+(A) \leq 2w(s) \tag{2.9}$$

for every $s > 0$ and every Borel measurable set $A \subset \mathbb{R}$. If (2.9) fails, there exists $s > 0$ and a set A such that

$$\mathrm{meas}\,(A) = 2s\,, \qquad \tilde{\mu}^+(A) > 2w(s) = 2 \lim_{\nu \to \infty} w_\nu(s).$$

Since w is continuous for $s > 0$, we can choose an open set $A' \supseteq A$ such that, setting $s' \doteq \text{meas}\,(A')/2$, one has $2w(s') < \tilde{\mu}^+(A)$. By the weak convergence $\mu^+_\nu \rightharpoonup \tilde{\mu}^+$ one obtains

$$\tilde{\mu}^+(A') \leq \liminf_{\nu\to\infty} \mu^+_\nu(A') \leq 2w(s') < \tilde{\mu}^+(A)\,,$$

reaching a contradiction. Hence (2.9) must hold. □

Toward the proof of Theorem 1 we shall need a lower semicontinuity property for wave measures, similar to what was proved in [BaB]. In the following, C_0 is the same constant as in (1.15).

LEMMA 1. *Consider a sequence of functions u_ν with uniformly small total variation and call μ^{i+}_ν the corresponding measures of positive i-waves. Let $s \mapsto w_\nu(s)$, $\nu \geq 1$, be a sequence of odd, nondecreasing functions, concave for $s > 0$, such that*

$$\mu^{i+}_\nu \preceq D_s\Big[w_\nu + C_0\,\text{sgn}(s)\big(Q_0 - Q(u_\nu)\big)\Big] \tag{2.10}$$

for some Q_0. Assume that $u_\nu \to u$ and $w_\nu \to w$ in $\mathbf{L}^1_{\text{loc}}$. Then the measure of positive i-waves in u satisfies

$$\mu^{i+} \preceq D_s\Big[w + C_0\,\text{sgn}(s)\big(Q_0 - Q(u)\big)\Big]\,. \tag{2.11}$$

Proof. The main steps follow the proof of Theorem 10.1 in [B].

1. By possibly taking a subsequence we can assume that $u_\nu(x) \to u(x)$ for every x and that the measures of total variation converge weakly, say,

$$|\mu_\nu| \doteq \big|D_x u_\nu\big| \rightharpoonup \mu^\sharp \tag{2.12}$$

for some positive Radon measure $\mu^\sharp$. In this case one has $\mu^\sharp \geq |\mu|$, in the sense of (1.12).

2. Let any $\varepsilon > 0$ be given. Since the total mass of $\mu^\sharp$ is finite, one can select finitely many points $y_1, \ldots, y_N$ such that

$$\mu^\sharp\big(\{x\}\big) < \varepsilon \qquad \text{for all} \quad x \notin \{y_1, \ldots, y_N\}. \tag{2.13}$$

We now choose disjoint open intervals $I_k \doteq]y_k - \rho,\ y_k + \rho[$ such that

$$\mu^\sharp\big(I_k \setminus \{y_k\}\big) < \frac{\varepsilon}{N}, \qquad k = 1, \ldots, N. \tag{2.14}$$

Moreover, we choose $R > 0$ such that

$$\bigcup_{k=1}^N I_k \subset [-R, R], \qquad \mu^\sharp\big(]-\infty, -R] \cup [R, \infty[\big) < \varepsilon. \tag{2.15}$$

Because of (2.13), we can now choose points $p_0 < -R < p_1 < \cdots < R < p_r$ which are continuity points for u and for every u_ν, such that

$$\mu^\sharp\big(\{p_h\}\big) = 0, \qquad u_\nu(p_h) \to u(p_h) \qquad \text{for all} \ \ h = 0, \ldots, r \tag{2.16}$$

and such that either

$$p_h - p_{h-1} < \frac{\varepsilon}{N}\,, \qquad p_{h-1} < y_k < p_h, \qquad [p_{h-1}, p_h] \subset I_k \tag{2.17}$$

for some $k \in \{1, \dots, N\}$, or else

$$|\mu|\big([p_{h-1}, p_h]\big) \leq \mu^\sharp\big([p_{h-1}, p_h]\big) < \varepsilon. \tag{2.18}$$

Call $J_h \doteq [p_{h-1}, p_h]$. If (2.18) holds, by weak convergence for some ν_0 sufficiently large one has

$$|\mu_\nu|(J_h) < \varepsilon \qquad \text{for all} \quad \nu \geq \nu_0. \tag{2.19}$$

On the other hand, if (2.17) holds, from (2.14) it follows that

$$|\mu|\big(J_h \setminus \{y_k\}\big) \leq \mu^\sharp\big(J_h \setminus \{y_k\}\big) < \frac{\varepsilon}{N}\,. \tag{2.20}$$

In the remainder of the proof, the main strategy is as follows.

- On the intervals $J_{h(k)}$ containing a point y_k of large oscillation, we first replace each u_ν by a piecewise constant function $\bar u_\nu$ having a single jump at y_k. The relations between the corresponding measures μ^i_ν and $\bar\mu^i_\nu$ are given by Lemma 10.2 in [B]. Then we take the limit as $\nu \to \infty$.
- On the remaining intervals J_h with small oscillation, we replace the left eigenvectors $l_i(u_\nu)$ by a constant vector $l_i(u^*_h)$. Then we use Proposition 4 to estimate the limit as $\nu \to \infty$.

3. We first take care of the intervals J_h containing a point y_k of large oscillation, so that (2.17) holds. For each $k = 1, \dots, N$, let $h = h(k) \in \{1, \dots, r\}$ be the index such that $y_k \in J_h \doteq [p_{h-1},\ p_h]$. For every $\nu \geq 1$ consider the function

$$\bar u_\nu(x) \doteq \begin{cases} u_\nu(x) & \text{if} \quad x \notin \cup_k J_{h(k)}, \\ u_\nu(p_{h(k)-1}) & \text{if} \quad x \in]p_{h(k)-1},\ y_k[, \\ u_\nu(p_h) & \text{if} \quad x \in [y_k,\ p_{h(k)}]. \end{cases}$$

Observe that all functions $u, \bar u_\nu$ are continuous at every point $p_0, \dots, p_r$ and have jumps at $y_1, \dots, y_N$. Call $\bar\mu^i_\nu$, $i = 1, \dots, n$, the corresponding measures, defined as in (1.8) with u replaced by $\bar u_\nu$. Clearly $\bar\mu^i_\nu = \mu^i_\nu$ outside the intervals $J_{h(k)}$ of large oscillation. By Lemma 10.2 at page 203 in [B], there holds

$$Q(\bar u_\nu) \leq Q(u_\nu), \qquad\qquad V(\bar u_\nu) + C_0\, Q(\bar u_\nu) \leq V(u_\nu) + C_0 \cdot Q(u_\nu),$$

$$\bar\mu^{i+}_\nu(\mathbb{R}) - \mu^{i+}_\nu(\mathbb{R}) \ \leq\ C_0\,\big[Q(u_\nu) - Q(\bar u_\nu)\big].$$

As a consequence, from (2.10) we deduce

$$\bar\mu^{i+}_\nu \preceq D_s\Big[T^\varepsilon w_\nu + C_0\,\mathrm{sgn}(s)\big(Q_0 - Q(\bar u_\nu)\big)\Big]\,, \tag{2.21}$$

where

$$T^\varepsilon w(s) \doteq \begin{cases} w(s + \varepsilon/2) & \text{if} \quad s > 0, \\ w(s - \varepsilon/2) & \text{if} \quad s < 0. \end{cases}$$

Indeed, all the mass which in μ^{i+}_ν lies on the set

$$\Omega \doteq \bigcup_{k=1}^N J_{h(k)}\,, \qquad\qquad J_h \doteq [p_{h-1},\ p_h]$$

is replaced in $\bar\mu^{i+}_\nu$ by point masses at $y_1,\ldots,y_N$. We obtain (2.21) by observing that, by (2.17), $meas(\Omega)<\varepsilon$. Moreover, the increase in the total mass is $\leq C_0\big[Q(u_\nu)-Q(\bar u_\nu)\big]$.

Since $u_\nu(p_h)\to u(p_h)$ for every h, there holds

$$\begin{aligned}\Big|\mu^i\big(\{y_k\}\big)-\bar\mu^i_\nu\big(\{y_k\}\big)\Big| &= \mathcal{O}(1)\cdot\Big\{\big|u(y_k-)-u(p_{h(k)-1})\big|+\big|u(y_k+)-u(p_{h(k)})\big|\\ &\qquad +\big|u(p_{h(k)-1})-u_\nu(p_{h(k)-1})\big|+\big|u(p_{h(k)})-u_\nu(p_{h(k)})\big|\Big\}\\ &= \mathcal{O}(1)\cdot\frac{\varepsilon}{N}\end{aligned}\tag{2.22}$$

for each $k=1,\ldots,N$ and all ν sufficiently large. By construction we also have

$$|\bar\mu^i_\nu|\big(J_{h(k)}\setminus\{y_k\}\big)=0,\qquad |\mu^i|\big(J_{h(k)}\setminus\{y_k\}\big)=\mathcal{O}(1)\cdot\frac{\varepsilon}{N}\,.\tag{2.23}$$

4. Next, call $\mathcal{S}\doteq\big\{h\,;\ \mu^\sharp(J_h)<\varepsilon\big\}$ the family of intervals where the oscillation of every u_ν is small, so that (2.18) holds. If $h\in\mathcal{S}$, for every $x,y\in J_h$ and ν sufficiently large, one has

$$\big|u_\nu(x)-u_\nu(y)\big|\leq|\mu_\nu|(J_h)<\varepsilon,$$

$$\big|u(x)-u(y)\big|\leq|\mu|(J_h)\leq\mu^\sharp(J_h)<\varepsilon.$$

Set $u^*_h\doteq u(p_h)$. By the pointwise convergence $u_\nu(p_h)\to u(p_h)$ and the two above estimates it follows that

$$\big|u_\nu(x)-u^*_h\big|<\varepsilon,\qquad \big|u(x)-u^*_h\big|<\varepsilon\qquad\text{for all }\ x\in J_h\,.\tag{2.24}$$

5. We now introduce the measures $\hat\mu^i_\nu$ such that

$$\hat\mu^i_\nu\doteq l_i(u^*_h)\cdot D_xu_\nu$$

restricted to each interval J_h, $h\in\mathcal{S}$, where the oscillation is small, while

$$\hat\mu^i_\nu=\bar\mu^i_\nu$$

on each interval $J_h=J_{h(k)}$ where the oscillation is large. Observe that the restriction of $\hat\mu^i_\nu$ to $J_{h(k)}$ consists of a single mass at the point y_k. Namely, $\hat\mu^i_\nu\big(\{y_k\}\big)$ is precisely the size of the ith wave in the solution of the Riemann problem with data $u^-=u_\nu(p_{h(k)-1})$, $u^+=u_\nu(p_{h(k)})$.

We define $\hat w_\nu$ as the nondecreasing odd function such that

$$\hat w_\nu(s)\ \doteq\ \sup_{meas(A)\leq 2s}\frac{\hat\mu^{i+}_\nu(A)}{2}\,,\qquad s>0.\tag{2.25}$$

By possibly taking a further subsequence we can assume the convergence

$$Q(\bar u_\nu)\to\overline{Q}\,,\qquad \hat\mu^i_\nu\rightharpoonup\hat\mu^i\,,\qquad \hat w_\nu(s)\to\hat w(s)\,.$$

Using (2.16), we can apply Proposition 4 on each interval J_h and obtain

$$\hat\mu^{i+}\preceq D_s\hat w\,.\tag{2.26}$$

6. Observe that, by (2.24) and (2.19),

$$|\hat\mu^i_\nu - \mu^i_\nu|(J_h) = \mathcal{O}(1)\cdot\varepsilon\,\mu^\sharp(J_h), \qquad\qquad h\in\mathcal{S}. \tag{2.27}$$

From (2.21) and the definition of $\hat w_\nu$ at (2.25) it thus follows that

$$\hat w_\nu(s) \leq T^\varepsilon w_\nu(s) + C_0\big[Q_0 - Q(\bar u_\nu)\big] + \mathcal{O}(1)\cdot\varepsilon, \qquad\qquad s>0\,. \tag{2.28}$$

Letting $\nu\to\infty$ we obtain

$$\hat w(s) \leq T^\varepsilon w(s) + C_0[Q_0 - \overline{Q}] + \mathcal{O}(1)\cdot\varepsilon, \qquad\qquad s>0\,, \tag{2.29}$$

$$\overline{Q} = \lim_{\nu\to\infty} Q(\bar u_\nu) \geq \lim_{\nu\to\infty} Q(u_\nu) - \mathcal{O}(1)\cdot\varepsilon \geq Q(u) - \mathcal{O}(1)\cdot\varepsilon\,, \tag{2.30}$$

because of the lower semicontinuity of the functional $u\mapsto Q(u)$. From (2.26), (2.29), and (2.30) we deduce

$$\hat\mu^{i+} \preceq D_s\Big[T^\varepsilon w + \mathrm{sgn}(s)\,\big(C_0[Q_0 - Q(u)] + \mathcal{O}(1)\cdot\varepsilon\big)\Big]\,.$$

By (2.22)–(2.24), our construction of the measure $\hat\mu^i$ achieves the property

$$\big|\mu^{i+} - \hat\mu^{i+}\big|(\mathbb{R}) = \mathcal{O}(1)\cdot\varepsilon\,.$$

Hence, by Proposition 3,

$$\mu^{i+} \preceq D_s\Big[T^\varepsilon w + \mathrm{sgn}(s)\,\big(C_0[Q_0 - Q(u)] + \mathcal{O}(1)\cdot\varepsilon\big)\Big]\,.$$

Since $\varepsilon>0$ was arbitrary, this proves (2.11). □

3. A decay estimate. The second basic ingredient in the proof is the following lemma, which refines the estimate in [BC].

LEMMA 2. *For some constant $\kappa>0$ the following holds. Let $u=u(t,x)$ be any entropy weak solution of* (1.1)*, with initial data $u(0,x)=\bar u(x)$ having small total variation. Then the measure μ^{i+}_t of positive i-waves in $u(t,\cdot)$ can be estimated as follows.*

Let $w:[0,\tau[\,\times\mathbb{R}\mapsto\mathbb{R}$ be the solution of Burgers's equation

$$w_t + (w^2/2)_x = 0 \tag{3.1}$$

with initial data

$$w(0,x) = \mathrm{sgn}(x)\cdot\sup_{meas(A)\leq 2|x|}\frac{\mu^{i+}_0(A)}{2}\,. \tag{3.2}$$

Set

$$w(\tau,x) = w(\tau-,\ x) + \kappa\,\mathrm{sgn}(x)\cdot\big[Q(\bar u) - Q\big(u(\tau)\big)\big]\,. \tag{3.3}$$

Then

$$\mu^{i+}_\tau \preceq D_x w(\tau)\,. \tag{3.4}$$

Proof. The main steps follow the proof of Theorem 10.3 in [B]. We first prove the estimate (3.3) under the following additional hypothesis.

(H) There exist points $y_1 < \cdots < y_m$ such that the initial data $\bar u$ is smooth outside such points, constant for $x < y_1$ and $x > y_m$, and the derivative component $l_i(u)\, u_x$ is constant on each interval $]y_\ell,\, y_{\ell+1}[$. Moreover, the Glimm functional $t \mapsto Q\big(u(t)\big)$ is continuous at $t = \tau$.

1. The solution $u = u(t,x)$ can be obtained as the limit of front tracking approximations. In particular, we can consider a particular converging sequence $(u_\nu)_{\nu\geq 1}$ of ε_ν-approximate solutions with the following additional properties:

(i) Each i-rarefaction front x_α travels with the characteristic speed of the state on the right:

$$\dot x_\alpha = \lambda_i\big(u(x_\alpha+)\big).$$

(ii) Each i-shock front x_α travels with a speed strictly contained between the right and the left characteristic speeds:

$$\lambda_i\big(u(x_\alpha+)\big) < \dot x_\alpha < \lambda_i\big(u(x_\alpha-)\big). \tag{3.5}$$

(iii) As $\nu \to \infty$, the interaction potentials satisfy

$$Q\big(u_\nu(0,\cdot)\big) \to Q(\bar u). \tag{3.6}$$

2. Let u_ν be an approximate solution constructed by the front tracking algorithm. By a *(generalized) i-characteristic* we mean an absolutely continuous curve $x = x(t)$ such that

$$\dot x(t) \in \big[\lambda_i(u_\nu(t,x-)),\ \lambda_i(u_\nu(t,x+))\big]$$

for a.e. t. If u_ν satisfies the above properties (i)–(ii), then the i-characteristics are precisely the polygonal lines $x : [0,\tau] \mapsto \mathbb{R}$ for which the following holds. For a suitable partition $0 = t_0 < t_1 < \cdots < t_m = \tau$, on each subinterval $[t_{j-1},\, t_j]$ either $\dot x(t) = \lambda_i\big(u_\nu(t,x)\big)$, or else x coincides with a wave front of the ith family. For a given terminal point $\bar x$ we shall consider the *minimal backward i-characteristic* through $\bar x$, defined as

$$y(t) = \min\big\{x(t)\,;\quad x \ \text{ is an } i\text{-characteristic},\quad x(\tau) = \bar x\big\}.$$

Observe that $y(\cdot)$ is itself an i-characteristic. By (3.5), it cannot coincide with an i-shock front of u on any nontrivial time interval.

In connection with the exact solution u, we define an i-characteristic as a curve

$$t \mapsto x(t) = \lim_{\nu\to\infty} x_\nu(t)$$

which is the limit of i-characteristics in a sequence of front tracking solutions $u_\nu \to u$.

3. Let $\varepsilon > 0$ be given. If the assumption (H) holds, the measure μ^{i+}_τ of i-waves in $u(\tau)$ is supported on a bounded interval and is absolutely continuous w.r.t. Lebesgue measure. We can thus find a piecewise constant function ψ^τ with jumps at points $x_1(\tau) < \bar x_2(\tau) < \cdots < \bar x_N(\tau)$ such that

$$\int \left|\frac{d\mu^{i+}_\tau}{dx} - \psi^\tau\right| dx < \varepsilon\,, \qquad \int_{x_j(\tau)}^{x_{j+1}(\tau)} \left(\frac{d\mu^{i+}_\tau}{dx} - \psi^\tau\right) dx = 0, \qquad j = 1,\dots,N-1\,. \tag{3.7}$$

To prove the lemma in this special case, relying on Proposition 2, it thus suffices to find i-characteristics $t \mapsto x_j(t)$ such that the following hold (Figure 2):

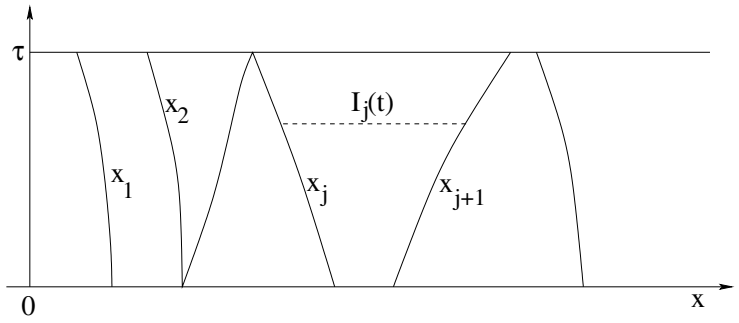

FIG. 2.

(i) For each $j = 1, \dots, N$, the function ψ^τ is constant on the interval $\big]x_j(\tau), x_{j+1}(\tau)\big[$ and (3.7) holds. Moreover, either $x_j(0) = x_{j+1}(0)$, or else the derivative component $\psi^0 \doteq l^i(u)u_x(0,\cdot)$ is constant on the interval $\big]x_j(0), x_{j+1}(0)\big[$.

(ii) An estimate corresponding to (3.3)–(3.4) holds restricted to each subinterval $\big[x_j(\tau),\ x_{j+1}(\tau)\big[$.

We need to explain in more detail this last statement. Define

$$I_j(t) \doteq \big[x_j(t),\, x_{j+1}(t)\big[\,, \qquad\qquad \Delta_j \doteq \big\{(t,x)\,;\quad t\in[0,\tau]\,,\quad x\in I_j(t)\big\}\,.$$

For each j, we denote by Γ_j the total amount of wave interaction within the domain Δ_j. This is defined as in [B], first for a sequence of front tracking approximations u_ν, then taking a limit as $\nu\to\infty$. Furthermore, we define the constant values

$$\psi_j^\tau \doteq \psi^\tau(x), \qquad\qquad x\in I_j(\tau)\,,$$

$$\psi_j^0 \doteq \psi^0(x), \qquad\qquad x\in I_j(0).$$

Call

$$\sigma_j^0 \doteq \lim_{t\to 0+} \mu^{i+}\big(I_j(t)\big)$$

the initial amount of positive i-waves inside the interval I_j.

For each interval I_j, we consider on one hand the function w_j^τ corresponding to (3.2)–(3.3), namely,

$$w_j^\tau(s) \doteq \min\left\{\sigma_j^0\,,\ \frac{s}{\tau + (\psi_j^0)^{-1}}\right\} + \kappa\Gamma_j\cdot \mathrm{sgn}(s)\,.$$

Here $(\psi_j^0)^{-1} \doteq 0$ in the case where $x_j(0) = x_{j+1}(0)$. This may happen when the initial data has a jump at $x_j(0)$, and the corresponding measure μ^{i+} has a Dirac mass (with infinite density) at that point.

On the other hand, we look at the nondecreasing, odd function η_j such that

$$\eta_j(s) \doteq \min\Big\{\psi_j^\tau\, s,\quad \psi_j^\tau\big[x_{j+1}(\tau) - x_j(\tau)\big]\Big\}, \qquad\qquad s>0\,.$$

Our basic goal is to prove that (Figure 3)

$$\eta_j(s) \le w_j^\tau(s) \qquad\qquad \text{for all } s>0\,. \tag{3.8}$$

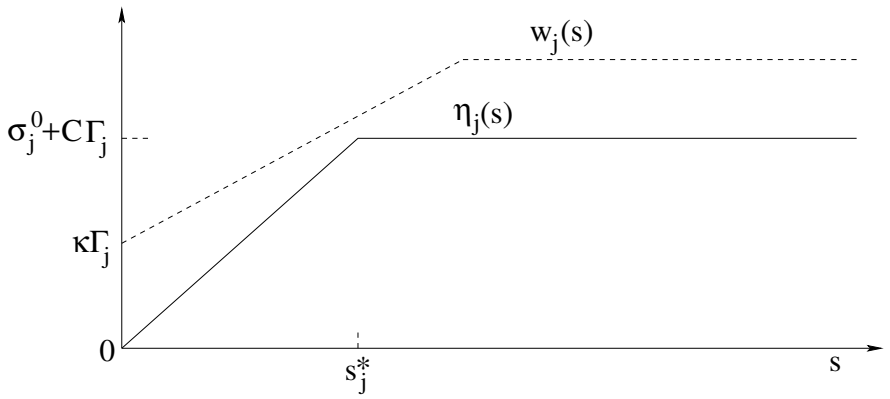

FIG. 3.

Indeed, by (3.7), for $s > 0$ one has

$$\sup_{meas(A)\leq 2s} \frac{\mu_\tau^{i+}\big(A\cap I_j(\tau)\big)}{2} \leq \eta_j(s) + \varepsilon_j$$

with

$$\sum_j \varepsilon_j < \varepsilon\,.$$

Proving (3.8) for each j will thus imply

$$\mu_\tau^{i+} \preceq w(\tau,x) = w(\tau-,\ x) + \kappa\,\mathrm{sgn}(x)\cdot\big[Q(\bar u) - Q\big(u(\tau)\big) + \mathcal{O}(1)\cdot\varepsilon\big]\,.$$

Since $\varepsilon > 0$ was arbitrary, this establishes the lemma under the additional assumptions (H).

4. We now work toward a proof of (3.8), in three cases.

Case 1: $\sigma_j^0 = 0$.

Case 2: $x_j(0) = x_{j+1}(0)$ and $\sigma_j^0 > 0$.

Case 3: $x_j(0) < x_{j+1}(0)$ and $\sigma_j^0 = \big(x_{j+1}(0) - x_j(0)\big)\,\psi_j^0 > 0$.

In Case 1 the proof is easy. Indeed, the total amount of positive i-waves in $I_j(\tau)$ is here bounded by a constant times the total amount of interaction taking place inside the domain Δ_j, i.e.,

$$\mu_\tau^{i+}\big(I_j(\tau)\big) \leq C_0\cdot\Gamma_j$$

for some constant C_0. On the other hand

$$w_j^\tau(s) = \kappa\Gamma_j\cdot\mathrm{sgn}(s)\,.$$

Choosing $\kappa > C_0$ we achieve (3.8).

5. Since Case 2 can be obtained from Case 3 in the limit as $x_{j+1} - x_j \to 0$, we shall only give a proof for Case 3.

We can again distinguish two cases. If the amount of interaction Γ_j is large compared with the initial amount of i-waves, say

$$\Gamma_j \geq \frac{1}{6C_0}\sigma_j^0\,,$$

then the bound (3.8) is readily achieved choosing $\kappa > 8C_0$. Indeed, for $s > 0$ we have

$$\eta_j(s) \leq \frac{1}{2}\,\mu_\tau^{i+}\big(I_j(\tau)\big) \leq C_0\Gamma_j + \sigma_j^0 \leq 7C_0\,\Gamma_j\,.$$

The more difficult case to analyze is when Γ_j is small, say

$$\Gamma_j < \sigma_j^0/6C_0\,. \tag{3.9}$$

Looking at Figure 3, it clearly suffices to prove (3.8) for the single value

$$s = s_j^* \doteq \frac{x_{j+1}(\tau) - x_j(\tau)}{2}\,.$$

Equivalently, calling

$$z_j(t) \doteq x_{j+1}(t) - x_j(t)$$

the length of the interval $I_j(t)$ and

$$\sigma_j^\tau \doteq \mu_\tau^{i+}\big(I_j(\tau)\big) = z_j(\tau)\,\psi_j^\tau$$

the total amount of positive i-waves inside $I_j(\tau)$, we need to show that

$$\sigma_j^\tau \leq 2\kappa\Gamma_j + \min\left\{\sigma_j^0\,,\quad \frac{2s_j^*}{\tau + (\psi_j^0)^{-1}}\right\}\,. \tag{3.10}$$

By the approximate conservation of i-waves over the region Δ_j, we can write

$$\sigma_j^\tau \leq \sigma_j^0 + C_0\Gamma_j\,. \tag{3.11}$$

Using (3.11) in (3.10), our task is reduced to showing that

$$\sigma_j^\tau \leq 2\kappa\Gamma_j + \frac{2s_j^*}{\tau + (\psi_j^0)^{-1}} \tag{3.12}$$

for a suitably large constant κ. Because of (3.11), it suffices to show that

$$\begin{aligned} z_j(\tau) &\geq (\sigma_j^0 - C'\Gamma_j)\big(\tau + (\psi_j^0)^{-1}\big) \\ &= \big[z_j(0) + \tau\sigma_j^0\big] - C'\big(\tau + (\psi_j^0)^{-1}\big)\,\Gamma_j \end{aligned} \tag{3.13}$$

for a suitable constant C'.

6. We now prove (3.13). Notice that, by genuine nonlinearity and the normalization (1.2), if no other waves were present in the region Δ_j we would have $\Gamma_j = 0$ and

$$\frac{d}{dt}z_j(t) \equiv \sigma_j^0\,.$$

In this case, the equality would hold in (3.13).

To handle the general case, we represent the solution u as a limit of front tracking approximations u_ν, where for each $\nu \geq 1$ the function $u_\nu(0,\cdot)$ contains exactly ν rarefaction fronts equally spaced along the interval $I_j(0)$ (Figure 4). Each of these fronts has initial strength $\sigma_\alpha(0) = \sigma_j^0/\nu$. For $\alpha = 1,\ldots,\nu$, let $y_\alpha(t) \in I_j(t)$ be the

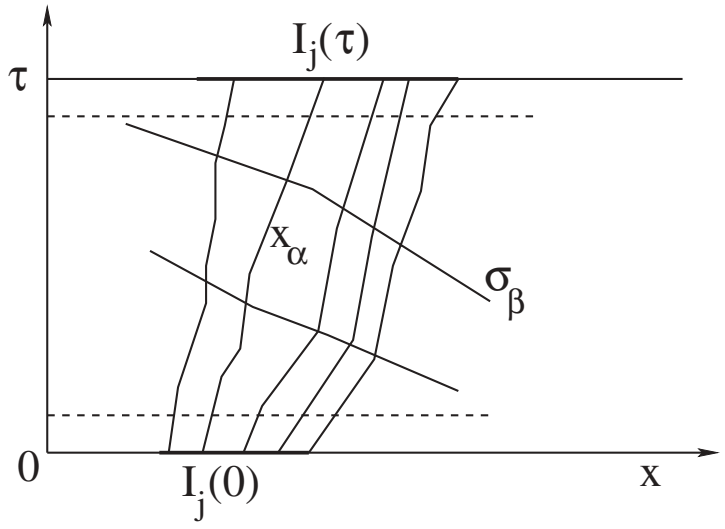

FIG. 4.

location of one of these fronts at time $t \in [0,\tau]$, and let $\sigma_\alpha(t) > 0$ be its strength. Moreover, call

$$J_\alpha(t) \doteq \big[y_\alpha(t)\,,\ y_{\alpha+1}(t)\big]\,, \qquad\qquad \Delta_\alpha \doteq \big\{(t,x)\,;\ \ t\in[0,\tau]\,,\ \ x\in J_\alpha(t)\big\}\,,$$

and let Γ_α be the total amount of interaction in u_ν taking place inside the domain Δ_α.

We define a subset of indices $\mathcal{I} \subseteq \{1,\dots,\nu\}$ by setting $\alpha \in \mathcal{I}$ if

$$5C_0\Gamma_\alpha > \sigma_\alpha(0) = \sigma_j^0/\nu\,. \tag{3.14}$$

Observe that, if $\alpha \notin \mathcal{I}$, then

$$\left|\frac{\sigma_\alpha(t)}{\sigma_\alpha(0)} - 1\right| < \frac{1}{2} \qquad \text{for all } t\in[0,\tau]\,.$$

In particular, if $\alpha, \alpha+1 \notin \mathcal{I}$, then the interval $J_\alpha(t)$ is well defined for all $t\in[0,\tau]$. Its length

$$z_\alpha(t) \doteq y_{\alpha+1}(t) - y_\alpha(t)$$

satisfies the differential inequality

$$\frac{d}{dt} z_\alpha(t) \geq W_\alpha(t) - C_1 \cdot \sum_{\beta\in\mathcal{C}_\alpha(t)} |\sigma_\beta| \tag{3.15}$$

for some constant C_1. Here

$$\begin{aligned} W_\alpha(t) &\doteq \big[\text{amount of } i\text{-waves inside the interval } J_\alpha(t)\big] \\ &\geq \sigma_\alpha(0) - C_0\Gamma_\alpha\,, \end{aligned} \tag{3.16}$$

while $\mathcal{C}_\alpha(t)$ refers to the set of all wave fronts of different families which are crossing the interval J_α at time t. Calling W'_α the total amount of waves of families $\neq i$ which lie inside $J_\alpha(0)$, we can now write

(3.17)
$$\int_0^\tau \left(\sum_{\beta \in \mathcal{C}_\alpha(t)} |\sigma_\beta| \right) dt \leq \Big(\max_{t\in[0,\tau]} z_\alpha(t) \Big) \cdot \frac{2\nu}{\sigma_j^0} \cdot \Gamma_\alpha + \mathcal{O}(1) \cdot \tau \Gamma_\alpha + \mathcal{O}(1) \cdot \left(\frac{z_j(0)+1}{\nu} \right) W'_\alpha.$$

Indeed, by strict hyperbolicity, every front σ_β of a different family can spend at most a time $\mathcal{O}(1) \cdot z_\alpha$ inside J_α. Either it is located inside J_α already at time $t = 0$, or else, when it enters, it crosses y_α or $y_{\alpha+1}$. In this case, since $\alpha, \alpha+1 \notin \mathcal{I}$, by (3.14) it will produce an interaction of magnitude $|\sigma_\beta\, \sigma_\alpha| \geq |\sigma_\beta \cdot \sigma_j^0|/2\nu$. The second term on the right-hand side of (3.17) takes care of the new wave fronts which are generated through interactions inside J_α. The last term takes into account wave fronts of different families that initially lie already inside J_α at time $t = 0$. Integrating (3.15) over the time interval $[0, \tau]$ and using (3.16)–(3.17) one obtains

$$z_\alpha(\tau) \;\geq\; z_\alpha(0) + \tau \frac{\sigma_j^0}{\nu} - \mathcal{O}(1) \cdot \tau \Gamma_\alpha - \mathcal{O}(1) \cdot \Big(\max_{t\in[0,\tau]} z_\alpha(t) \Big) \cdot \frac{2\nu}{\sigma_j^0} \cdot \Gamma_\alpha - \mathcal{O}(1) \cdot \left(\frac{z_j(0)+1}{\nu} \right) W'_\alpha.$$

(3.18)

7. To proceed in our analysis, we now show that

$$\max_{t\in[0,\tau]} z_\alpha(t) \leq 2\, z_\alpha(\tau)\,. \tag{3.19}$$

Indeed, let $\tau' \in [0, \tau]$ be the time where the maximum is attained. If our claim (3.19) does not hold, there would exist a first time $\tau'' \in [\tau', \tau]$ such that $z_\alpha(\tau'') = z_\alpha(\tau')/2$. From (3.15) and the assumption $W_\alpha(t) \geq 0$ it follows that

$$\int_{\tau'}^{\tau''} C_1 \sum_{\beta \in \mathcal{C}_\alpha(t)} |\sigma_\beta|\, dt \geq \frac{z_\alpha(\tau')}{2}\,. \tag{3.20}$$

Using the smallness of the total variation, a contradiction is now obtained as follows. Call

$$\Phi(t) \doteq C_0 Q(t) + \sum_{k_\beta \neq i} \phi_{k_\beta}\big(t, x_\beta(t)\big)\, |\sigma_\beta|\,,$$

where the sum ranges over all fronts of strength σ_β located at x_β, of a family $k_\beta \neq i$. The weight functions ϕ_j are defined as

$$\phi_j(t,x) \doteq \begin{cases} 0 & \text{if} \quad x > y_{\alpha+1}(t)\,, \\ \frac{y_{\alpha+1}(t) - x}{y_{\alpha+1}(t) - y_\alpha(t)} & \text{if} \quad x \in \big[y_\alpha(t),\; y_{\alpha+1}(t)\big]\,, \\ 1 & \text{if} \quad x < y_\alpha(t) \end{cases}$$

in the case $j > i$, while

$$\phi_j(t,x) \doteq \begin{cases} 1 & \text{if} \quad x > y_{\alpha+1}(t)\,, \\ \frac{x - y_\alpha(t)}{y_{\alpha+1}(t) - y_\alpha(t)} & \text{if} \quad x \in \big[y_\alpha(t),\; y_{\alpha+1}(t)\big]\,, \\ 0 & \text{if} \quad x < y_\alpha(t) \end{cases}$$

in the case $j < i$. Because of the term $C_0 Q(t)$, the functional Φ is nonincreasing at times of interactions. Moreover, outside interaction times a computation entirely similar to the one on page 213 of [B] now yields

$$-\frac{d}{dt}\Phi(t) \geq \sum_{\beta\in\mathcal{C}_\alpha(t)} |\sigma_\beta| \cdot \frac{c_0}{z(t)} \tag{3.21}$$

for some small constant $c_0 > 0$ related to the gap between different characteristic speeds. From (3.20) and (3.21), respectively, we now deduce

$$\int_{\tau'}^{\tau''} \sum_{\beta\in\mathcal{C}_\alpha(t)} |\sigma_\beta|\, dt \geq \frac{z_\alpha(\tau')}{2C_1}\,,$$

$$\int_{\tau'}^{\tau''} \sum_{\beta\in\mathcal{C}_\alpha(t)} |\sigma_\beta|\, dt \leq \int_{\tau'}^{\tau''} \left|\frac{d\Phi(t)}{dt}\right| \cdot \frac{z_\alpha(\tau')}{c_0}\, dt \leq \frac{\Phi(\tau')}{c_0}\, z_\alpha(\tau')\,.$$

Since $\Phi(t) = \mathcal{O}(1)\cdot \text{Tot.Var.}\big\{u(t)\big\}$, by the smallness of the total variation we can assume $\Phi(\tau') < 2C_1/c_0$. In this case, the two above inequalities yield a contradiction.

8. Using (3.19), from (3.18) we obtain

$$\begin{aligned}
z_j(\tau) &= \sum_{1\leq\alpha\leq\nu} z_\alpha(\tau) \geq \sum_{\alpha\notin\mathcal{I}} z_\alpha(\tau) \\
&\geq \sum_{\alpha\notin\mathcal{I}} \left\{ \frac{z_\alpha(0) + \tau\sigma_j^0/\nu}{1 + C_2(\nu/\sigma_j^0)\Gamma_\alpha} - \mathcal{O}(1)\cdot\tau\Gamma_j - \mathcal{O}(1)\cdot\left(\frac{z_j(0)+1}{\nu}\right) W'_\alpha \right\} \\
&\geq \sum_{\alpha\notin\mathcal{I}} \left(z_\alpha(0) + \tau\frac{\sigma_j^0}{\nu}\right)\left(1 - C_2\frac{\nu}{\sigma_j^0}\Gamma_\alpha\right) - \mathcal{O}(1)\cdot\tau\Gamma_j - \mathcal{O}(1)\cdot\frac{z_j(0)+1}{\nu} \\
&\geq \sum_{\alpha\notin\mathcal{I}} \left(z_\alpha(0) + \tau\frac{\sigma_j^0}{\nu}\right) - C_2\frac{z_j(0)}{\sigma_j^0}\Gamma_j - \mathcal{O}(1)\cdot\tau\Gamma_j - \mathcal{O}(1)\cdot\frac{z_j(0)+1}{\nu}\,. \qquad (3.22)
\end{aligned}$$

By (3.14) the cardinality of the set $\mathcal{I}$ satisfies

$$\#\mathcal{I}\cdot\frac{\sigma_j^0}{5C_0\nu} \leq \sum_{\alpha\in\mathcal{I}} \Gamma_\alpha \leq \Gamma_j;$$

hence

$$\frac{\#\mathcal{I}}{\nu} \leq \frac{5C_0}{\sigma_j^0}\Gamma_j\,.$$

In turn, this implies

$$\begin{aligned}
\sum_{\alpha\notin\mathcal{I}} \left(z_\alpha(0) + \tau\frac{\sigma_j^0}{\nu}\right) &\geq \big(z_j(0) + \tau\sigma_j^0\big)\left(1 - \frac{\#\mathcal{I}}{\nu}\right) \\
&\geq \big(z_j(0) + \tau\sigma_j^0\big) - 5C_0\Gamma_j\frac{z_j(0)}{\sigma_j^0}\Gamma_j - 5C_0\tau\Gamma_j\,. \qquad (3.23)
\end{aligned}$$

Using (3.23) in (3.22), observing that

$$\frac{z_j(0)}{\sigma_j^0} = \frac{x_{j+1}(0) - x_j(0)}{\sigma_j^0} = (\psi_j^0)^{-1},$$

and letting $\nu \to \infty$, we conclude

$$z_j(\tau) \geq \big(z_j(0) + \tau\sigma_j^0\big) - \mathcal{O}(1) \cdot (\psi_j^0)^{-1}\Gamma_j - \mathcal{O}(1) \cdot \tau\Gamma_j\,.$$

This establishes (3.13) for a suitable constant C'.

9. In the general case, without the assumptions (H), the lemma is proved by an approximation argument. We construct a convergent sequence of initial data $\bar u_\nu \to \bar u$ which satisfy (H) and such that

$$\bar u_\nu \to \bar u\,, \qquad Q(\bar u_\nu) \to Q(\bar u)\,, \qquad \big|\mu^{i+}_{\nu,0} - \mu^{i+}_0\big| \to 0\,.$$

Calling w_ν the solution of (3.1) with initial data

$$w_\nu(0,x) = \mathrm{sgn}(x) \cdot \sup_{meas(A)\leq 2|x|} \frac{\mu^{i+}_{\nu,0}(A)}{2}\,,$$

by the previous analysis we have

$$\mu^{i+}_{\nu,\tau_\nu} \preceq D_x\Big[w_\nu(\tau_\nu-) + \mathrm{sgn}(x) \cdot \big[Q(\bar u_\nu) - Q(u_\nu(\tau_\nu))\big]\Big]\,.$$

Observe that $w_\nu(\tau_\nu-) \to w(\tau-)$ in $\mathbf{L}^1_{\rm loc}$. Choosing $\kappa \geq C_0$, by the lower semicontinuity result stated in Lemma 1 we now conclude

$$\mu^{i+}_\tau \preceq D_x\Big[w(\tau-) + \kappa\,\mathrm{sgn}(x) \cdot \big[Q(\bar u) - Q(u(\tau))\big]\Big]\,. \qquad \square$$

4. Proof of the main theorem. Using the previous lemmas, we now give a proof of Theorem 1. For a given interval $[0,\tau]$, the solution of the impulsive Cauchy problem (1.17)–(1.18) can be obtained as follows. Consider a partition $0 = t_0 < t_1 < \cdots < t_N = \tau$. Construct an approximate solution by requiring that $w(0,x) = \hat v_i(x)$,

$$w_t + (w^2/2)_x = 0 \tag{4.1}$$

on each subinterval $[t_{k-1}, t_k[$, while

$$w(t_k, x) = w(t_k-\,,\ x) + \kappa\,\mathrm{sgn}(x) \cdot \big[Q(t_{k-1}) - Q(t_k)\big]\,. \tag{4.2}$$

We then consider a sequence of partitions $0 = t^\nu_0 < t^\nu_1 < \cdots < t^\nu_{N_\nu} = \tau$, and the corresponding solutions w_ν. If the mesh of the partitions approaches zero, i.e.,

$$\lim_{\nu\to\infty} \sup_k |t^\nu_k - t^\nu_{k-1}| = 0\,,$$

then the approximate solutions w_ν converge to a unique limit, which yields the solution of (1.17)–(1.18).

Call $\mathcal{F}$ the set of nondecreasing odd functions, concave for $x > 0$. This set is positively invariant for the flow of Burgers's equation (4.1). Moreover, this flow is order preserving. Namely, if $w, w' \in \mathcal{F}$ are solutions of (4.1) with initial data such that $w(0,x) \leq w'(0,x)$ for all $x > 0$, then also

$$w(t,x) \leq w'(t,x) \qquad\qquad \text{for all}\ \ t, x > 0\,.$$

Equivalently,

$$D_x w(0) \preceq D_x w'(0) \qquad \Longrightarrow \qquad D_x w(t) \preceq D_x w'(t)$$

for every $t > 0$. For each fixed ν, we can apply Lemma 2 on each subinterval $[t^\nu_{k-1}, t^\nu_k]$ and obtain

$$\mu^{i+}_{t^\nu_k} \preceq D_x w_\nu(t^\nu_k) \qquad \Longrightarrow \qquad \mu^{i+}_{t^\nu_{k+1}} \preceq D_x w_\nu(t^\nu_{k+1})\,.$$

By induction on k, this yields

$$\mu^{i+}_\tau \preceq D_x w_\nu(\tau)\,, \tag{4.3}$$

where w_ν is the approximate solution constructed according to (4.1)–(4.2). Letting $\nu \to \infty$ and using Lemma 1, we achieve a proof of Theorem 1. □

5. Examples.

Example 1. Consider first the scalar case. Let $u = u(t,x)$ be a solution of Burgers's equation with smooth initial data

$$u_t + (u^2/2)_x = 0\,, \qquad\qquad u(0,x) = \bar u(x)\,.$$

Define the function w as the solution to

$$w_t + (w^2/2)_x = 0\,, \qquad\qquad w(0,x) = \mathrm{sgn}(x) \cdot \sup_{meas(A)<2|x|} \int_A \frac{\bar u_x(y)}{2}\, dy\,.$$

This corresponds to (1.16)–(1.17) with $Q \equiv 0$. According to Oleinik's estimate we now have $u_x(t,x) \le 1/t$ for all $t > 0$, and a.e. $x \in \mathbb{R}$. Of course, this reflects the fact that, along each characteristic with $\dot x = u\big(t, x(t)\big)$, the gradient satisfies

$$\frac{d}{dt} u_x\big(t, x(t)\big) = -\frac{1}{u_x^2\big(t, x(t)\big)}\,. \tag{5.1}$$

A better estimate on $u_x\big(t, x(t)\big)$ when it is positive, based on (5.1), is

$$u_x\big(t, x(t)\big) \le \frac{1}{t + \big[\bar u_x(x(0))\big]^{-1}}\,. \tag{5.2}$$

According to (1.18), for every $t > 0$ we have the relation

$$\mu^+_t \preceq D_x w(t)\,, \tag{5.3}$$

which includes the additional information (5.2). This relation is sharp in the sense that the converse inequality

$$D_x w(t) \preceq \mu^+_t$$

also holds, as long as no positive waves are cancelled by interacting with shocks.

Analogous results are valid for scalar equations with more general flux:

$$u_t + f(u)_x = 0, \qquad\qquad u(0,x) = \bar u(x)\,,$$

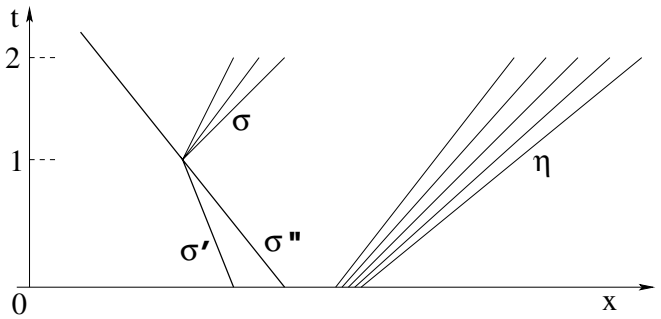

FIG. 5.

under the assumption of genuine nonlinearity $f'' > 0$. Notice that in this case the normalization (1.2) yields $r(u) = 1/f''(u)$, $l(u) = f''(u)$. As long as the solution $u(t,\cdot)$ remains smooth, the corresponding wave measure is thus defined as

$$\mu_t(A) = \int_A f''\big(u(t,x)\big) \cdot u_x(t,x)\, dx\,.$$

Example 2. Consider the p-system in Lagrangean coordinates (see [Sm])

$$v_t - u_x = 0\,, \qquad u_t + (K/v^\gamma)_x = 0\,.$$

Here $K > 0$ and $\gamma \geq 1$ are constants. Consider a solution which initially contains two approaching 1-shocks and a 2-rarefaction (Figure 5). Assume that at time $t = 1$ the two shocks interact. As shown in [Sm], the Riemann problem is then solved in terms of a 1-shock and a 2-rarefaction. Let us look at the measure of positive 2-waves in the solution. Let η be the strength of the centered rarefaction. During the interval $t \in [0,1[$ the density of rarefaction waves decays, as in the scalar case. At time $t = 1$ a new centered rarefaction is created by the interaction. Calling σ', σ'' the strengths of the incoming shocks, the strength σ of this new rarefaction will satisfy

$$\sigma \leq \kappa \cdot |\sigma'\sigma''| \doteq \tilde\sigma \tag{5.4}$$

for a suitable constant $\kappa > 0$. Notice that the decrease in the interaction potential at time $t = 1$ is $\Delta Q = -|\sigma'\sigma''|$. The values of the corresponding function $w = w(t,x)$ in (1.16)–(1.18) at various times are illustrated in Figure 6. For $t \in [0,1[$ the estimate (1.18) is sharp, in the sense that

$$D_x w(t) \preceq \mu_t^{2+} \preceq D_x w(t)\,. \tag{5.5}$$

The first relation in (5.5) will fail for $t \geq 1$ if $\tilde\sigma < \sigma$. The accuracy of our estimate in this case depends essentially on the careful choice of the constant κ in (5.4). In particular, if we could choose κ so that $\sigma = \kappa|\sigma'\sigma''|$, then both relations in (5.5) would remain valid for all times $t \geq 0$.

Remark. Concerning compression waves, an estimate of the form (1.18) could be derived also for the negative part of the measures μ_t^i. In this case, μ_t^{i-} can be compared with the gradient $D_x w$ of an odd, nonincreasing solution of a perturbed Burgers's equation. However, the result does not appear to be very interesting. Indeed, as time progresses negative waves become ever more singular and a bound such as (1.18)

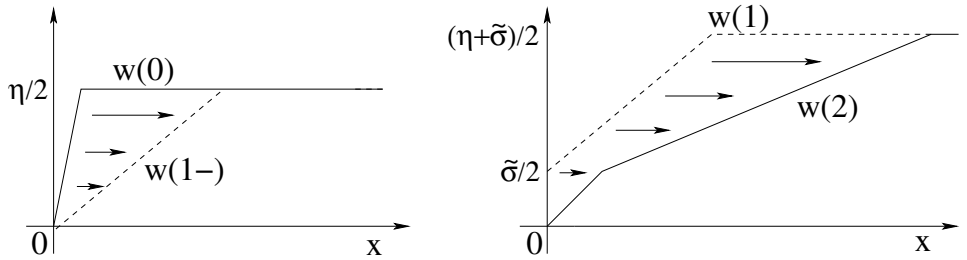

FIG. 6.

retains little content. For large times, a much better way to estimate negative waves is to analyze their cancellation with positive waves of the same family, as in [L1], [L2].

REFERENCES

[BaB] P. BAITI AND A. BRESSAN, *Lower semicontinuity of weighted path length in BV*, in Geometrical Optics and Related Topics, F. Colombini and N. Lerner, eds., Birkhäuser, Boston, 1997, pp. 31–58.

[B] A. BRESSAN, *Hyperbolic Systems of Conservation Laws. The One Dimensional Cauchy Problem*, Oxford University Press, New York, 2000.

[BC] A. BRESSAN AND R. M. COLOMBO, *Decay of positive waves in nonlinear systems of conservation laws*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 26 (1998), pp. 133–160.

[BG] A. BRESSAN AND P. GOATIN, *Oleinik type estimates and uniqueness for $n \times n$ conservation laws*, J. Differential Equations, 156 (1999), pp. 26–49.

[BY] A. BRESSAN, AND T. YANG, *On the convergence rate of vanishing viscosity approximations*, Comm. Pure Appl. Math., submitted.

[GL] J. GLIMM AND P. LAX, *Decay of solutions of systems of nonlinear hyperbolic conservation laws*, Mem. Amer. Math. Soc., 101 (1970).

[L1] T. P. LIU, *Decay to N-waves of solutions of general systems of nonlinear hyperbolic conservation laws*, Comm. Pure Appl. Math., 30 (1977), pp. 586–611.

[L2] T. P. LIU, *Linear and nonlinear large-time behavior of solutions of general systems of hyperbolic conservation laws*, Comm. Pure Appl. Math., 30 (1977), pp. 767–796.

[L3] T. P. LIU, *Admissible solutions of hyperbolic conservation laws*, Mem. Amer. Math. Soc., 30 (1981).

[O] O. OLEINIK, *Discontinuous solutions of nonlinear differential equations*, Amer. Math. Soc. Transl. (2), 26 (1963), pp. 95–172.

[Sm] J. SMOLLER, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1983.

SIAM J. MATH. ANAL.
Vol. 36, No. 2, pp. 678–690

ATTRACTION DOMAINS OF DEGENERATE SINGULAR EQUILIBRIA IN QUASI-LINEAR ODES*

RICARDO RIAZA†

Abstract. The present paper addresses stability properties of singular equilibria arising in a given family of quasi-linear ODEs. These ODEs model continuous-time methods for root-finding problems and their singular equilibria are *degenerate* in the sense that the linearization of the system at equilibrium yields a singular matrix pencil. The analysis is based upon a normal form defined by a codimension-one regular system coupled with a singular scalar equation. The key step is the formulation of certain codimension-one Lyapunov matrix equations which incorporate the relevant singular information and allow for the construction of Lyapunov functions supporting the stability analysis. This approach makes it possible to state precisely the asymptotic stability of such degenerate equilibria, and provides a local estimation of the corresponding attraction domains. An application to the computation of singular DC operating points in nonlinear circuits is discussed.

Key words. implicit ODE, singularity, normal form, stability of equilibria, matrix equation, Lyapunov function, root-finding, nonlinear circuit

AMS subject classifications. 15A24, 34A09, 34C20, 37B25, 65H20, 94C05

DOI. 10.1137/S0036141003425003

1. Introduction.

Consider the implicit differential system

$$y' = \xi(y,z) \tag{1a}$$

$$zz' = \zeta(y,z), \tag{1b}$$

where $y \in \mathbb{R}^{n-1}$, $z \in \mathbb{R}$, and the functions $\xi(y,z)$, $\zeta(y,z)$ are sufficiently smooth.

System (1) arises as a local normal form for quasi-linear ODEs $A(u)u' = f(u)$ (where A and f are sufficiently smooth matrix- and vector-valued functions) around a *singular point* satisfying $\det A(u^*) = 0$ and $(\det A)'(u^*)v \neq 0$, $\forall v \in \mathrm{Ker}A(u^*) - \{0\}$ [11, 16, 19, 22]. It is also closely related to certain singular differential-algebraic equations (DAEs) [3, 4, 17]. These singular equations are relevant in problems arising in magnetohydrodynamics, electrical circuits, or power system theory, to name a few [5, 6, 15, 23].

Singularities of the normal form (1) are located in the hyperplane $z = 0$, and can be classified (see [3, 4, 6, 12, 13, 17, 19, 23] and references therein) into *pseudo-equilibria* (defined by the condition $\zeta(y,0) = 0$), *forward impasse points* ($\zeta(y,0) < 0$), or *backward impasse points* ($\zeta(y,0) > 0$). Smooth solutions may be defined through pseudoequilibrium points, whereas a pair of trajectories collapse when reaching an impasse point, either in forward or backward time direction.

The present work is focused on stability issues concerning *singular equilibria* of (1), characterized by the pair of conditions $\xi(y,0) = 0$, $\zeta(y,0) = 0$. Within the context of semiexplicit DAEs, these singular equilibria have been analyzed in [3] under an assumption of regularity on the matrix pencil (see [18] and references therein for

*Received by the editors April 23, 2003; accepted for publication (in revised form) October 24, 2003; published electronically August 27, 2004. This research was supported by Proyecto I+D 14583, Universidad Politécnica de Madrid.

http://www.siam.org/journals/sima/36-2/42500.html

†Departamento de Matemática Aplicada a las Tecnologías de la Información, Escuela Técnica Superior de Ingenieros de Telecomunicación, Universidad Politécnica de Madrid, Ciudad Universitaria s/n - 28040, Madrid, Spain (rrr@mat.upm.es).

background on this topic) describing the linearization of the problem. For system (1), this matrix pencil would read $\mu C - D$, μ being a complex parameter, whereas $C = \mathrm{diag}\{I_{n-1}, 0\}$ and D is the Jacobian matrix of the right-hand side of (1) evaluated at equilibrium; regularity of the matrix pencil means that there exists a μ_0 such that $\mu_0 C - D$ is invertible, allowing for the definition of a Kronecker index for the pencil.

In contrast, the present paper will address *degenerate* singular equilibria, displaying a singular pencil in the linearization. Specifically, we will consider ODEs of the form

$$y' = Hy + \beta(y, z), \tag{2a}$$

$$zz' = \lambda z^2 + y^T G y + \gamma(y, z), \tag{2b}$$

where $H,\ G \in \mathbb{R}^{(n-1)\times(n-1)}$, G being symmetric, and $\lambda \in \mathbb{R}$. The functions $\beta(y, z)$ and $\gamma(y, z)$ are $O(\|(y, z)\|)^2$ and $O(\|(y, z)\|)^3$, respectively. Singularity of the matrix pencil arising in the linearization at the origin would in this situation follow from the vanishing of the last row of $\mu C - D$ for any μ, with the above-explained notation.

System (2), with $H = -I_{n-1},\ \lambda = -1/2$, is proved in [19] to describe a normal form around a singular equilibrium of the so-called continuous Newton method (see [18, 20, 21] and the bibliography therein)

$$-J(u)u' = f(u) \tag{3}$$

for sufficiently smooth $f : \mathbb{R}^n \to \mathbb{R}^n$, J being the Jacobian matrix of f. Euler discretization of (3) yields the classical Newton iteration for root-finding and optimization problems. The interest of a continuous-time scheme in this setting stems from its better properties regarding global issues and singular problems, together with the fact that a unique continuous system may lead to different iterative techniques, including damped and accelerated versions of basic methods, through the use of different integration schemes. Therefore, the same continuous-time study may be of interest for a wide family of discrete-time techniques [21].

A singular equilibrium of (3) is defined by the pair of conditions $f(u^*) = 0$ and $\mathrm{rk}J(u^*) < n$, together with the assumption that u^* is a limit point of the set where J is invertible. Singular roots arise for instance in predictor-corrector continuation methods [1], and are mapped into the origin in the normal form (2) [19]. In the discrete-time context of the classical Newton method, several results concerning the existence of locally cone-shaped regions of attraction for these singular zeroes were proved in [8, 10, 14] and references therein. Continuous-time extensions have been addressed in [18, 20], and applications of this approach to the original discrete-time setting can be found in [21].

Nevertheless, several issues in this direction remain open. In section 2, we analyze the actual local shape of attraction domains of degenerate equilibria of (2) and the dynamic phenomena which are responsible for these local shapes. Under certain assumptions, the attraction domain comprises a cone-shaped region, but the actual domain may be larger, which may have important implications regarding the use of (3) and related discretizations in singular root-finding problems. This local shape will be shown to be intimately linked with the nature of singularities surrounding the equilibrium and, in particular, with the backward or forward nature of nearby impasse points. The analysis will be supported on the use of several Lyapunov functions, based in turn on certain $(n-1)$-dimensional Lyapunov matrix equations constructed from the "regular" part (2a) but incorporating the relevant information from the singular

equation (2b). An application to the computation of singular DC operating points in nonlinear circuits is discussed in section 3.

The reader is referred to [2] for background on semiflows, invariance, and Lyapunov functions. Useful facts coming from matrix analysis and involving, in particular, the Lyapunov matrix equation can be found in [9].

2. Dynamics around degenerate equilibria. System (2) and, in particular, the continuous Newton method, define a (possibly not complete) flow Φ on the set of regular points

$$\mathcal{X} = \{(y,z) \in \mathbb{R}^{n-1} \times \mathbb{R} \ / z \neq 0\}. \tag{4}$$

Nevertheless, in order to analyze the behavior of this flow near the origin, the reader should avoid simply considering an "extension" of this flow to $\mathcal{X} \cup \{0\}$ by adding $\Phi(t,0) = 0$ for $t \in \mathbb{R}$, since the resulting extension of Φ may not be well defined as a flow. Therefore, our present approach will be based on the fact that well-defined, complete semiflows are induced by (2) on certain positively invariant subsets of $\mathcal{X}$; asymptotic stability of the origin may then be precisely addressed for such semiflows.

Inspired on convergence results for singular equilibria of Newton's method [8, 10, 14, 18, 20], one may conjecture if these asymptotic stability results for (2) may follow from the assumption that λ and the spectral abscissa

$$\alpha = \max_{\mu \in \sigma(H)} \operatorname{Re} \mu \tag{5}$$

are negative. Note that, for the continuous Newton method, it is $\alpha = -1,\ \lambda = -1/2$. System (7) in section 2.1 will prove that this conjecture is false in general. It is shown in section 2.2 that, at least, the conditions $\alpha < 0,\ \lambda < 0$ make it possible to prove that nearby trajectories remain on a given neighborhood of the origin as long as they are defined.

If the matrix G is positive definite, then the assumptions $\alpha < 0,\ \lambda < 0$ suffice indeed to guarantee the asymptotic convergence to the origin of all trajectories emanating from *regular* points within a given neighborhood of the equilibrium. This is illustrated by the case $\eta > 0$ in section 2.1, and proved in general in section 2.3. The local attraction domains of such singular roots are therefore significantly larger than a cone-shaped region, making those solutions more easily computable through Newton-based techniques.

Nevertheless, without the positive definiteness of G, this nice behavior will not be displayed. It will be shown in section 2.4 that, under the dominance condition $\alpha < \lambda < 0$, there exists a cone-shaped region with vertex in the origin which is positively invariant and asymptotically convergent to the degenerate equilibrium. Compared with previous approaches in this direction (see [18] and references therein), the Lyapunov function method used to prove this result will additionally give a hint for the estimation of the actual local shape of the attraction domain. An example is provided by the case $\eta < 0$ in section 2.1.

Particularization of these results for the continuous Newton method and an application in circuit theory can be found in sections 2.5 and 3, respectively.

2.1. A glimpse. Consider the vector field $f(y,z) = (y, z^2 + 2\eta y^2)$, with $(y,z) \in \mathbb{R}^2,\ \eta \in \mathbb{R}$. The continuous Newton method (3) reads $y' = -y,\ 4\eta y y' + 2zz' = -z^2 - 2\eta y^2$ or, equivalently,

$$y' = -y, \tag{6a}$$

$$zz' = -(1/2)z^2 + \eta y^2, \tag{6b}$$

which has the form depicted in (2), with $Hy = -y,\ \lambda = -1/2,\ y^T G y = \eta y^2, \beta = \gamma = 0$. The origin may be easily shown to be a degenerate singular equilibrium regardless of the value of $\eta \in \mathbb{R}$. Nevertheless, this parameter strongly influences the dynamic behavior around the origin, as discussed below.

Note that the particular value $\eta = 0$ yields a removable singularity, since (6) would amount in this case to $y' = -y,\ z' = -z/2$, leading to an asymptotically stable equilibrium in the classical sense. This nongeneric phenomenon has been analyzed in [18, 20, 21] and will not be considered further here.

Case $\eta > 0$. Solutions of (6) read

$$\begin{aligned} y(t) &= y_0 e^{-t}, \\ z(t) &= \mathrm{sg}(z_0)\sqrt{z_0^2 e^{-t} + 2\eta y_0^2 (e^{-t} - e^{-2t})}. \end{aligned}$$

Trajectories with $z_0 \neq 0$ are well defined for all $t \geq 0$, since the radicand is always positive. Furthermore, every initial point with $z_0 \neq 0$ converges to the origin. This means that the domain of attraction of the origin is the set of regular points, showing that of singular roots in Newton-based techniques may be significantly larger than a locally cone-shaped set.

Concerning general systems of the form (2), this behavior will be shown in section 2.3 to follow from the positive definiteness of G, together with the assumptions $\lambda < 0$, $\alpha < 0$. The first condition forces all singularities in a neighborhood of the equilibrium to behave as backward impasse points, and amounts in this example to $\eta > 0$. Trajectories are then repelled by the singular manifold and must evolve towards the equilibrium.

Case $\eta < 0$. In this case, if we rewrite the radicand as $(z_0^2 + 2\eta y_0^2)e^{-t} - 2\eta y_0^2 e^{-2t}$, it is not difficult to check that initial points (y_0, z_0) in the cone-shaped region

$$\{(y_0, z_0) \in \mathbb{R}^2 : z_0^2 + 2\eta y_0^2 \geq 0\} \ \equiv\ \{(y_0, z_0) \in \mathbb{R}^2 : |y_0| \leq |z_0|/\sqrt{2|\eta|}\}$$

guarantee that the radicand remains positive and, therefore, trajectories are well defined for all positive t. It is also easy to check that all these solutions converge to the origin.

On the contrary, the condition $z_0^2 + 2\eta y_0^2 < 0 \equiv |y_0| > |z_0|/\sqrt{2|\eta|}$ yields a positive collapse-time

$$t^* = \ln\left(\frac{2\eta y_0^2}{z_0^2 + 2\eta y_0^2}\right),$$

beyond which trajectories are not defined. Note that $2\eta y_0^2 < z_0^2 + 2\eta y_0^2 < 0$ and, therefore, $t^* > 0$. This means that trajectories outside the cone evolve towards a forward impasse point, where they cease to exist. This directional convergence phenomenon will be shown in section 2.4 to be a general property of systems of the form (2) with $\alpha < \lambda < 0$ and, in particular, of the continuous Newton method. The actual local shape of the attraction domain can be estimated with the Lyapunov function approach discussed there.

A generalization of (6). If we finally consider the system

$$y' = \alpha y, \tag{7a}$$

$$z z' = \lambda z^2 + \eta y^2, \tag{7b}$$

with $\eta < 0$, $\lambda < \alpha < 0$, it may be shown that solutions are

$$y(t) = y_0 e^{\alpha t},$$
$$z(t) = \mathrm{sg}(z_0)\sqrt{\left(z_0^2 + \frac{\eta}{\lambda-\alpha}y_0^2\right)e^{2\lambda t} - \frac{\eta}{\lambda-\alpha}y_0^2 e^{2\alpha t}}.$$

Now, any $y_0 \neq 0$ yields a positive escape time

$$t^* = \frac{\ln\left(1 + \frac{(\lambda-\alpha)z_0^2}{\eta y_0^2}\right)}{2(\alpha-\lambda)},$$

where it is to be noted that both $(\lambda-\alpha)z_0^2/\eta y_0^2$ and $\alpha - \lambda$ are positive. This means that only the z-coordinate curve is convergent to the origin and suggests that the dominance condition $\alpha < \lambda < 0$ is a key requirement in the phenomenon of directional stability.

2.2. Positive invariance with $\boldsymbol{\alpha < 0}$, $\boldsymbol{\lambda < 0}$. As a preliminary result let us recall that, writing as $\mu_1, \dots, \mu_n$ the n (not necessarily distinct) real eigenvalues of an $n \times n$ symmetric matrix A, and denoting

$$\eta_A = \min\{\mu_1, \dots, \mu_n\}, \tag{8a}$$
$$\kappa_A = \max\{\mu_1, \dots, \mu_n\}, \tag{8b}$$

then $\eta_A|x|^2 \leq x^\tau A x \leq \kappa_A|x|^2$ for any vector $x \in \mathbb{R}^n$, $|\ |$ standing for the Euclidean norm.

PROPOSITION 1. *Consider a quasi-linear ODE of the form* (2) *with* $\lambda < 0$. *Assume that the spectral abscissa defined in* (5) *verifies* $\alpha < 0$, *and denote* $\tilde{\kappa} = \max\{\kappa_G, 0\} \geq 0$, *where* κ *is defined as in* (8b). *Let* P *be the positive definite solution of the* $(n-1)$*-dimensional Lyapunov matrix equation*

$$PH + H^T P = 2(-\tilde{\kappa} + \lambda)I_{n-1}. \tag{9}$$

Then, there exists $r_0 > 0$ *such that*

$$V(y,z) = y^T P y + z^2 \tag{10}$$

satisfies $V' \leq 0$ *on* $\{x = (y,z) \in \mathcal{X} \ / \ |x| \leq r_0\}$. *Hence, for* $0 < V_0 < \min_{|x|=r_0} V(y,z)$, *the level sets* $\{x = (y,z) \in \mathcal{X} \ / \ V(y,z) \leq V_0\}$ *are positively invariant.*

Proof. The derivative of V along trajectories of (2) reads

$$V' = y^T(PH + H^T P)y + 2[y^T P\beta(y,z) + \lambda z^2 + y^T G y + \gamma(y,z)].$$

Now, using (9) and the fact that $y^T G y \leq \tilde{\kappa}|y|^2$, we have

$$\begin{aligned} V' &\leq 2[(-\tilde{\kappa} + \lambda)|y|^2 + y^T P\beta(y,z) + \lambda z^2 + \tilde{\kappa}|y|^2 + \gamma(y,z)] \\ &= 2[\lambda(|y|^2 + z^2) + y^T P\beta(y,z) + \gamma(y,z)]. \end{aligned}$$

Let $r_\beta,\ c_\beta,\ r_\gamma,\ c_\gamma$ be positive constants such that

$$|x| \leq r_\beta \Rightarrow |\beta(x)| \leq c_\beta|x|^2,$$
$$|x| \leq r_\gamma \Rightarrow |\gamma(x)| \leq c_\gamma|x|^3,$$

and define

$$r_0 = \min\left\{ r_\beta,\ r_\gamma,\ \frac{|\lambda|}{\kappa_P c_\beta + c_\gamma} \right\}.$$

Then, $|x| \le r_0$ implies that

$$|y^T P\beta(y,z) + \gamma(y,z)| \le |y|\kappa_P c_\beta |x|^2 + c_\gamma |x|^3 \le r_0(\kappa_P c_\beta + c_\gamma)|x|^2 \le |\lambda||x|^2$$

and, therefore, $V' \le 0$. Positive invariance of the level sets $\{x = (y,z) \in \mathcal{X} \ / \ V(y,z) \le V_0\}$ follows from [2, Theorem 18.2]. □

Denoting

$$\tilde{P} = \begin{pmatrix} P & 0 \\ 0 & 1 \end{pmatrix},$$

the level sets $\{x = (y,z) \in \mathcal{X} \ / \ V(y,z) \le V_0\}$ can be alternatively described in terms of the Hilbert norm

$$\|x\| = \sqrt{x^T \tilde{P} x} \equiv \sqrt{y^T P y + z^2} = \sqrt{V(y,z)}. \tag{11}$$

To this end, let us define

$$\mathcal{B}(0,\rho) \ = \{x \in \mathbb{R}^n \ / \ \|x\| \le \rho\}, \tag{12a}$$

$$\mathcal{B}^+(0,\rho) = \{x = (y,z) \in \mathbb{R}^n \ / \ \|x\| \le \rho,\ z > 0\}, \tag{12b}$$

$$\mathcal{B}^-(0,\rho) = \{x = (y,z) \in \mathbb{R}^n \ / \ \|x\| \le \rho,\ z < 0\}, \tag{12c}$$

$$\mathcal{B}^\pm(0,\rho) = \mathcal{B}^+(0,\rho) \cup \mathcal{B}^-(0,\rho) = \mathcal{B}(0,\rho) \cap \mathcal{X}. \tag{12d}$$

From the relations $\eta_{\tilde{P}}|x|^2 \le \|x\|^2 \le \kappa_{\tilde{P}}|x|^2$, it is easy to check that $\min_{|x|=r_0} V(y,z) = \eta_{\tilde{P}} r_0^2$. Defining $\rho_0 = \sqrt{\eta_{\tilde{P}}}\, r_0$, the following result can be immediately derived from Proposition 1.

COROLLARY 1. *If $\alpha < 0$ and $\lambda < 0$, system (2) induces a semiflow Φ on $\mathcal{B}^\pm(0,\rho)$ for all positive $\rho < \rho_0$.*

Note that $\mathcal{B}^\pm(0,\rho)$ excludes points in the hyperplane $z = 0$. From the result above the reader should not conclude that, in general, trajectories evolve towards the origin; note that there might be forward impasse points on $z = 0$ attracting solutions in finite time. In this situation, the additional dominance condition $\alpha < \lambda < 0$ (verified, in particular, by the continuous Newton method), allows one to prove the existence of a locally cone-shaped region which is positively invariant and asymptotically convergent to the origin, as shown in section 2.4. Nevertheless, in the particular case in which the origin is entirely surrounded by backward impasse points, then the origin is actually a stable attractor for the dynamics on $\mathcal{B}^\pm(0,\rho)$, without the need for the above-mentioned dominance condition. This simpler case, associated with the positive definiteness of G, is considered in section 2.3.

2.3. Completeness and asymptotic stability with positive definite G.

THEOREM 1. *Consider the quasi-linear ODE (2). Assume that $\alpha < 0$, $\lambda < 0$, and that G is positive definite. Then, there exists a positive $\rho_1 < \rho_0$ such that the semiflow Φ induced by (2) on $\mathcal{B}^\pm(0,\rho_1)$ is complete; that is, all solutions are defined on the time interval $[0,\infty)$. Furthermore, $\lim_{t\to\infty} \Phi(t,x_0) = 0$ for all x_0 in $\mathcal{B}^\pm(0,\rho_1)$.*

Proof. The key aspect here is that the smallness of ρ_1 must guarantee that all singularities in $\mathcal{B}(0,\rho_1)-\{0\}$ are backward impasse points. To achieve this, choose any positive r_1 satisfying

$$r_1 < \min\left\{r_0, \frac{\eta_G}{c_\gamma}\right\}, \tag{13}$$

where $\eta_G > 0$ since G is positive definite, and define $\rho_1 = \sqrt{\eta_{\tilde{P}}}r_1$. Hence, if $\|(y,0)\| \leq \rho_1 = \sqrt{\eta_{\tilde{P}}}r_1$, then $|y| \leq \|(y,0)\|/\sqrt{\eta_{\tilde{P}}} \leq r_1$ and $|\gamma(y,0)| \leq c_\gamma|y|^3 \leq c_\gamma r_1|y|^2 < \eta_G|y|^2$ (note that $y=0$ is excluded). This implies that $\gamma(y,0) > -\eta_G|y|^2$ and, since $y^TGy \geq \eta_G|y|^2$,

$$y^TGy + \gamma(y,0) > 0,\ (y,0) \in \mathcal{B}(0,\rho_1)-\{0\},$$

showing that all singularities in $\mathcal{B}(0,\rho_1)-\{0\}$ are backward impasse points. This fact will suffice to prove that all trajectories are well defined in $[0,\infty)$.

Suppose that $x_0 = (y_0,z_0) \in \mathcal{B}^+(0,\rho_1)$ is such that the trajectory emanating from this point is defined for a maximal positive time $t^+(x_0) < \infty$ (the reasoning in $\mathcal{B}^-(0,\rho_1)$ would be entirely analogous). In this situation, [2, Proposition 10.12] shows that for every compact set $\mathcal{C}$ on $\mathcal{B}^+(0,\rho_1)$ there would exist a time $t_0 < t^+(x_0)$ such that $\Phi(t,x_0) \notin \mathcal{C}$ for $t > t_0$. If we define $\mathcal{C}_\varepsilon = \{(y,z) \in \mathcal{B}^+(0,\rho_1) \ / \ z \geq \varepsilon\}$, for every $\varepsilon > 0$ there would exist a $t_0(\varepsilon)$ such that $z(t) < \varepsilon$ if $t > t_0(\varepsilon)$. This shows that, under the assumption $t^+(x_0) < \infty$, it would be $\lim_{t\to t^+(x_0)} z(t) = 0$.

Nevertheless, the positive definiteness of G precludes $z(t)$ from reaching the set $z=0$ in finite time, as shown below. Since $|\gamma(y,z)| \leq c_\gamma r_1|x|^2 \leq \eta_G|x|^2$, we have $\gamma \geq -\eta_G|x|^2$ and, therefore,

$$\lambda z^2 + y^TGy + \gamma(y,z) \geq \lambda z^2 + \eta_G|y|^2 - \eta_G(|y|^2+z^2) = (\lambda-\eta_G)z^2.$$

This means that the real-valued function $(\lambda z^2 + y^TGy + \gamma(y,z))/z$ is bounded below by $(\lambda-\eta_G)z$ on $\mathcal{B}^+(0,\rho_1)$. Since orbits do not leave $\mathcal{B}^+(0,\rho_1)$ due to Proposition 1, the z-component of the trajectory emanating from (y_0,z_0), with $z_0 > 0$, satisfies $z(t) \geq z_0 e^{(\lambda-\eta_G)t} > 0$ for all finite t, in contradiction with $\lim_{t\to t^+(x_0)} z(t) = 0$. This proves that $t^+(x_0) = \infty$ and, since x_0 is arbitrary, the semiflow is complete on $\mathcal{B}^\pm(0,\rho_1)$.

In this situation, the function V defined in (10) behaves as a classical Lyapunov function, since limit points verifying $z=0$, $y \neq 0$ are ruled out by the backward nature of impasse points in $\mathcal{B}(0,\rho_1)$. Hence, $\Phi(t,x_0) \to 0$ as $t \to \infty$, for all x_0 in $\mathcal{B}^\pm(0,\rho_1)$. ∎

Under the hypotheses of Theorem 1, the semiflow Φ may be safely extended to $\mathcal{B}^\pm(0,\rho_1)\cup\{0\}$ by adding $\Phi(t,0) = 0$ for $t \in [0,\infty)$, the origin being an asymptotically stable equilibrium of the resulting semiflow. In applications concerning the continuous Newton method, this situation yields a domain of attraction for a singular root which comprises all regular points within a usual ball about the degenerate equilibrium, as compiled in item 1 of Theorem 3 in section 2.5. An example of this nice behavior is given by the case $V_0 = 0$ in the nonlinear circuit presented in section 3.

2.4. Directional convergence with $\alpha < \lambda < 0$ and arbitrary G. If G is not positive definite, forward impasse points of system (2) may (and actually will, if G is not positive semidefinite) be displayed around the origin, precluding the application of the results discussed in section 2.3. Nevertheless, Proposition 1 and Corollary 1

can be strengthened under the additional dominance condition $\alpha < \lambda < 0$, which is satisfied in particular by the continuous Newton method, for which $\alpha = -1$, $\lambda = -1/2$.

In this direction, note that the condition $\alpha < \lambda$ implies that $H - \lambda I_{n-1}$ is an asymptotically stable matrix, since $\sigma(H-\lambda I_{n-1}) = \{\mu-\lambda \,/\, \mu \in \sigma(H)\}$, and, therefore, the condition $\alpha = \max_{\mu\in\sigma(H)} \operatorname{Re}\mu < \lambda$ yields $\max_{\mu\in\sigma(H)} \operatorname{Re}(\mu-\lambda) < 0$. This fact guarantees that the matrix equation (14) below has indeed a positive definite solution.

THEOREM 2. *Consider a quasi-linear ODE* (2) *with* $\alpha < \lambda < 0$, *and let* Q *be the positive definite solution of the* $(n-1)$*-dimensional Lyapunov matrix equation*

$$Q(H - \lambda I_{n-1}) + (H - \lambda I_{n-1})^T Q = -I_{n-1}. \tag{14}$$

Then, for every $\theta > 0$ *(satisfying additionally* $\theta < 1/\sqrt{2|\eta_G|}$ *if* $\eta_G < 0$, *where* η_G *is defined as in* (8a)), *there exists a positive* $\rho(\theta) < \rho_0$ *such that*

$$U(y,z) = y^T Q y - \theta^2 z^2 \tag{15}$$

satisfies $U' \le 0$ *on* $\partial\mathcal{K}(0,\theta) \cap \mathcal{B}^{\pm}(0,\rho(\theta))$, *where*

$$\mathcal{K}(0,\theta) = \{(y,z) \in \mathbb{R}^{n-1}\times\mathbb{R} : y^T Q y \le \theta^2 z^2\}, \tag{16}$$

and $\partial\mathcal{K}(0,\theta)$ *stands for the boundary* $\{(y,z) \in \mathbb{R}^{n-1}\times\mathbb{R} : y^T Q y = \theta^2 z^2\}$. *Hence, the set* $\mathcal{K}(0,\theta) \cap \mathcal{B}^{\pm}(0,\rho(\theta))$ *is positively invariant for the semiflow* Φ *defined in Corollary* 1. *Furthermore, the restriction of* Φ *to* $\mathcal{K}(0,\theta) \cap \mathcal{B}^{\pm}(0,\rho(\theta))$ *is complete, and* $\lim_{t\to\infty} \Phi(t,x_0) = 0$ *for all* x_0 *in* $\mathcal{K}(0,\theta) \cap \mathcal{B}^{\pm}(0,\rho(\theta))$.

Proof. The derivative of U along trajectories of (2) reads

$$U' = y^T(QH + H^T Q)y + 2[y^T Q\beta(y,z) - \theta^2(\lambda z^2 + y^T G y + \gamma(y,z))].$$

In the boundary $\partial\mathcal{K}(0,\theta)$, it is $U = 0$ or, equivalently, $\theta^2 z^2 = y^T Q y$. Therefore, in $\partial\mathcal{K}(0,\theta)$ we have

$$\begin{aligned} y^T(QH + H^T Q)y - 2\lambda\theta^2 z^2 &= y^T(QH + H^T Q)y - 2\lambda y^T Q y \\ &= y^T[Q(H - \lambda I_{n-1}) + (H - \lambda I_{n-1})^T Q]y = -|y|^2, \end{aligned}$$

yielding

$$\begin{aligned} U' &= -|y|^2 - 2\theta^2 y^T G y + 2[y^T Q\beta(y,z) - \theta^2\gamma(y,z)] \\ &\le -|y|^2 - 2\theta^2\eta_G|y|^2 + 2[y^T Q\beta(y,z) - \theta^2\gamma(y,z)]. \end{aligned}$$

In light of the restriction imposed on θ when $\eta_G < 0$, we always have $-1 - 2\theta^2\eta_G < 0$. The property $U' \le 0$, together with the other claims in the theorem, follows then from the choice of $\rho(\theta)$ in a way such that $2|y^T Q\beta(y,z) - \theta^2\gamma(y,z)| \le (1 + 2\theta^2\eta_G)|y|^2$ whenever $U = 0$, $\|x\| \le \rho(\theta)$. Details are straightforward and are left to the reader. ∎

2.5. The continuous Newton method at singular roots: Attraction domains.

THEOREM 3. *Consider the normal form* (2) *for the continuous Newton method around a singular equilibrium, for which* $H = -I_{n-1}$, $\lambda = -1/2$. *Then, the solution of the Lyapunov matrix equation* (14) *reads* $Q = I_{n-1}$. *Depending on the inertia of* G *in* (2b), *the following sets are included in the attraction domain of the origin,* $\mathcal{A}(0)$, *according to Theorems* 1 *and* 2*:*

1. *If $\eta_G > 0$, that is, if G is positive definite, then there exists a positive ρ_1 such that*

$$\mathcal{B}^{\pm}(0,\rho_1) \subset \mathcal{A}(0),$$

   $\mathcal{B}^{\pm}$ being defined in (12b).
2. *If $\eta_G \leq 0$, that is, if G is not positive definite, then for every positive $\theta < 1/\sqrt{2|\eta_G|}$ there exists a positive $\rho(\theta)$ such that the Euclidean cone $\mathcal{K}(0,\theta) = \{(y,z) \in \mathbb{R}^{n-1} \times \mathbb{R} : |y|^2 \leq \theta^2 z^2\}$ verifies that*

$$\mathcal{K}(0,\theta) \cap \mathcal{B}^{\pm}(0,\rho(\theta)) \subset \mathcal{A}(0).$$

   In the particular case $\eta_G = 0$, the result holds for any $\theta > 0$.

Note that the cone-shaped regions described in the second item of Theorem 3 are also included in the attraction domain of cases with positive definite G, but in this situation the first item yields a wider estimation of the local domain. It is also worth remarking that weak problems, for which these results may be improved, are not considered here (see [18, 20] and references therein). In cases with nonpositive definite G, the limit value $\theta = 1/\sqrt{2|\eta_G|}$ may provide a rather nice estimation of the actual extension of the cone-shaped region convergent to a singular root in the continuous Newton method. This value will also be shown to play a role in discrete-time counterparts of this method, particularly in the classical Newton iteration. This is illustrated, together with several additional features of these techniques, in a nonlinear circuit example addressed in section 3.

3. Bifurcation points of nonlinear circuits. The circuit displayed in Figure 1 includes an independent current source I_0, a linear capacitor C, two linear resistors R_1 and R_2, a nonlinear voltage-controlled current source (VCCS) with a quadratic characteristic $i = v^2$ (v being the voltage drop across the resistor R_1), a Josephson junction, and an independent voltage source V_0. The Josephson junction consists of two superconductors separated by an oxide barrier [7], and can be considered as a nonlinear inductor characterized by the differential relation $\phi' = v_J$, where ϕ denotes the magnetic flux in the junction, together with a (simplified) sinusoidal current-flux relation $i_J = \sin\phi$. As shown in Figure 1, most parameters in the circuit have been normalized for simplicity.

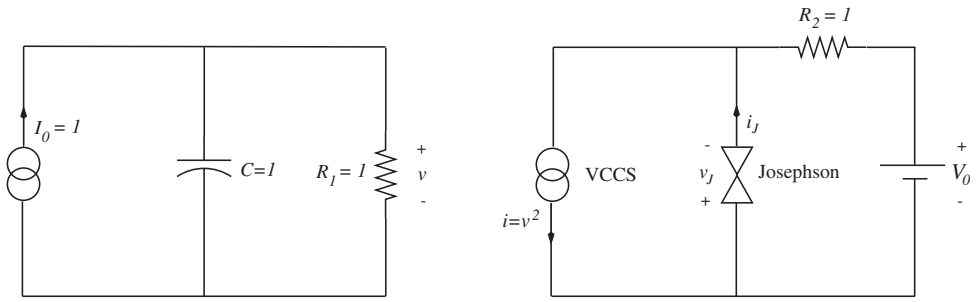

FIG. 1. *Nonlinear circuit.*

The dynamics of this circuit can be described in terms of the capacitor charge q and the flux ϕ in the Josephson junction through the ODE

$$q' = -q + 1, \tag{17a}$$

$$\phi' = q^2 - \sin\phi - V_0, \tag{17b}$$

whereas the continuous Newton method for the right-hand side of (17) can be written as

$$q' = -q + 1, \tag{18a}$$

$$\cos\phi\ \phi' = -q^2 + 2q - \sin\phi - V_0. \tag{18b}$$

Singularities of this quasi-linear ODE are defined by the condition $\cos\phi = 0$, which yields $\phi = \pi/2 + k\pi,\ k \in \mathbb{Z}$. Note that the location of singularities does not depend on the value of V_0. Singular equilibria will be displayed only if $\sin\phi = \pm 1$ at equilibrium (defined by $q = 1$, $\sin\phi = 1 - V_0$), that is, if $V_0 = 0$ or $V_0 = 2$. It can be checked that these values yield saddle-node bifurcations for (17) at $q = 1$, $\phi = \pi/2 + 2k\pi,\ k \in \mathbb{Z}$, for $V_0 = 0$, and $q = 1$, $\phi = -\pi/2 + 2k\pi,\ k \in \mathbb{Z}$, for $V_0 = 2$.

Let us consider the behavior of the continuous Newton method (18) regarding these singular operating points, focusing on

(a) $q = 1$, $\phi = \pi/2$ for $V_0 = 0$;
(b) $q = 1$, $\phi = -\pi/2$ for $V_0 = 2$.

In both cases, the normal form (2) can be easily computed through the coordinate change

$$y = q - 1, \tag{19a}$$

$$z = \cos\phi, \tag{19b}$$

with $\phi \in (0, \pi)$ for (a) and $\phi \in (-\pi, 0)$ for (b). Some simple computations lead to

$$y' = -y, \tag{20a}$$

$$zz' = -(1/2)z^2 \pm y^2 + \text{h.o.t.} \tag{20b}$$

The "+" sign in (20b) corresponds to (a), whereas the "−" case is obtained in (b). Note that the quadratic terms of (20b) are those of (6b) with $\eta = 1$ and $\eta = -1$, respectively.

Case (a). Singularities near the equilibrium $q = 1$, $\phi = \pi/2$, for $V_0 = 0$, are backward impasse points and, in light of the results in section 2.3 (see also item 1 in Theorem 3), every regular point sufficiently close to the singular equilibrium must evolve towards the origin. Computer simulations indicate that this is actually the case: Figure 2(a) displays an estimation of the attraction domain of this singular operating point. Note that the boundary of the attraction domain is partially defined by the straight lines $\phi = -\pi/2$ and $\phi = 3\pi/2$, which correspond to singularities of the quasi-linear ODE (18).

Case (b). Concerning the equilibrium $q = 1$, $\phi = -\pi/2$, for $V_0 = 2$, it follows from the results discussed in section 2.4 and compiled in item 2 of Theorem 3 that, for all $\theta < 1/\sqrt{2}$, there must exist a $\rho(\theta)$ defining a region of the form $y^2 \le \theta^2 z^2$ positively invariant and convergent to the origin. In the limit case $\theta = 1/\sqrt{2}$, the set

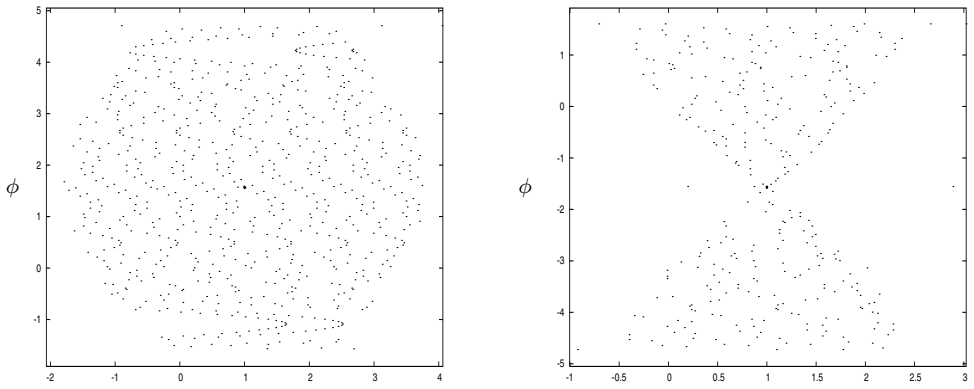

FIG. 2. *Domains of attraction of the continuous Newton method:* (a) *equilibrium at* $(1, \pi/2)$*,* $V_0 = 0$*;* (b) *equilibrium at* $(1, -\pi/2)$*,* $V_0 = 2$*.*

$y^2 = (1/2)z^2$ reads, in the original coordinates q, ϕ and using (19),

$$(q-1)^2 = (1/2)\cos^2\phi, \quad -\pi < \phi < 0. \tag{21}$$

This curve is plotted in Figure 2(b), together with a computer estimation of the attraction domain of the singular equilibrium. The figure clearly indicates that (21) provides, in this case, an accurate estimation of the boundary of the attraction domain near the degenerate solution. It is worth remarking that, as we move away from the singularity, the incidence of higher order terms becomes more significant and the divergence between the curve representing (21) and the boundary of the attraction domain is more apparent. Equation (21) is not plotted for $\phi < -\pi$ and $\phi > 0$ since the coordinate change (19) remains valid only for $-\pi < \phi < 0$. attraction domain is clearly delimited by the straight lines

With illustrative purposes, let us briefly address this directional convergence phenomenon in the discrete-time setting. Figure 3(a) displays the set of points which converge in the classical Newton iteration (obtained as the Euler discretization of (18) with stepsize 1) to the degenerate equilibrium without crossing the singular manifold; that is, the iteration is truncated for any orbit which jumps from one side of the singular manifold to the other (a jump which is allowed by the discrete-time nature of the method). The curve (21) again provides an accurate estimation of the boundary of the attraction domain for this truncated iteration.

The computer estimation of the actual local domain of attraction of the singular equilibrium for the discrete-time method is shown in Figure 3(b). This domain comprises not only the one shown in (a), but also those points which converge to the solution after crossing the singular manifold. For instance, the initial point $q = 2.5,\ \phi = -0.5$ may be shown to "jump over" the singular manifold in the first iteration step, reaching $q = 1.0,\ \phi = -3.6571$ and then evolving towards the singular root without additional jumps. Note that the first iteration step is exact in the q-component due to the decoupled and linear nature of (18a).

Although additional details in this direction are beyond the purposes of the present paper, let us finally remark here that this approach provides a linearly con-

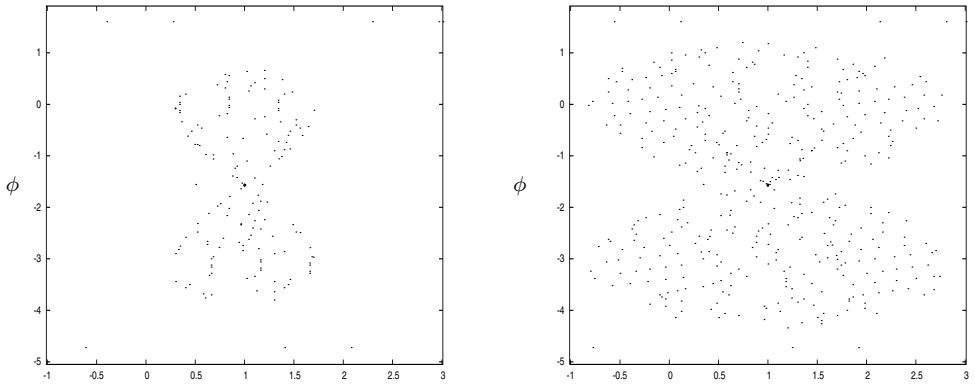

FIG. 3. *Local domains of attraction of* $(1, -\pi/2)$*:* (a) *truncated and* (b) *standard discrete-time Newton method,* $V_0 = 2$.

vergent iteration to the singular root. This follows from the value $\lambda = -1/2$ in the normal form (2b): Euler discretization with stepsize 1 places an eigenvalue at $1/2$ in the linearized discrete-time system. Quadratic convergence to singular roots may be recovered through the use of certain Runge–Kutta discretizations; see [21].

Acknowledgment. The author gratefully acknowledges several stimulating discussions with Professor Rafael Ortega from Universidad de Granada, Spain.

REFERENCES

[1] E. L. ALLGOWER AND K. GEORG, *Numerical Continuation Methods. An Introduction*, Springer-Verlag, Berlin, 1990.
[2] H. AMANN, *Ordinary Differential Equations*, Walter de Gruyter, Berlin, 1990.
[3] R. E. BEARDMORE AND R. LAISTER, *The flow of a DAE near a singular equilibrium*, SIAM J. Matrix Anal. Appl., 24 (2002), pp. 106–120.
[4] R. E. BEARDMORE, R. LAISTER, AND A. PEPLOW, *Trajectories of a DAE near a pseudo-equilibrium*, Nonlinearity, 17 (2004), pp. 253–279.
[5] S. L. CAMPBELL AND W. MARSZALEK, *DAEs arising from traveling wave solutions of PDEs*, J. Comput. Appl. Math., 82 (1997), pp. 41–58.
[6] L. O. CHUA AND A. D. DENG, *Impasse points,* I*: Numerical aspects*, Internat. J. Circuit Theory Appl., 17 (1989), pp. 213–235.
[7] L. O. CHUA, C. A. DESOER, AND E. S. KUH, *Linear and Nonlinear Circuits*, McGraw-Hill, New York, 1987.
[8] A. O. GRIEWANK, *On solving nonlinear equations with simple singularities or nearly singular solutions*, SIAM Rev., 27 (1985), pp. 537–563.
[9] R. A. HORN AND CH. R. JOHNSON, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.
[10] C. T. KELLEY AND Z. Q. XUE, *Inexact Newton methods for singular problems*, Optim. Methods Softw., 2 (1993), pp. 249–267.
[11] M. MEDVED, *Normal forms of implicit and observed implicit differential equations*, Riv. Mat. Pura Appl., 10 (1992), pp. 95–107.
[12] P. J. RABIER, *Implicit differential equations near a singular point*, J. Math. Anal. Appl., 144 (1989), pp. 425–449.
[13] P. J. RABIER AND W. C. RHEINBOLDT, *On impasse points of quasi-linear differential-algebraic equations*, J. Math. Anal. Appl., 181 (1994), pp. 429–454.

[14] G. W. Reddien, *On Newton's method for singular problems*, SIAM J. Numer. Anal., 15 (1978), pp. 993–996.
[15] G. Reiszig, *Differential-algebraic equations and impasse points*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 43 (1996), pp. 122–133.
[16] G. Reiszig and H. Boche, *A normal form for implicit differential equations near singular points*, in Proceedings of ECCTD'97, Budapest, Hungary, Vol. 2, 1997, pp. 1048–1053.
[17] G. Reiszig and H. Boche, *On singularities of autonomous implicit ordinary differential equations*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 50 (2003), pp. 922–931.
[18] R. Riaza, *Stability issues in regular and noncritical singular DAEs*, Acta Appl. Math., 73 (2002), pp. 301–336.
[19] R. Riaza, *Preliminary normal forms for a class of singular equilibria in implicit ODEs*, in Nonlinear Analysis and Applications, R. Agarwal and D. O'Regan, eds., Kluwer, 2003, pp. 831–850.
[20] R. Riaza and P. J. Zufiria, *Stability of singular equilibria in quasilinear implicit differential equations*, J. Differential Equations, 171 (2001), pp. 24–53.
[21] R. Riaza and P. J. Zufiria, *Discretization of implicit ODEs for singular root-finding problems*, J. Comput. Appl. Math., 140 (2002), pp. 695–712.
[22] J. Sotomayor and M. Zhitomirskii, *Impasse singularities of differential systems of the form* $A(x)x' = F(x)$, J. Differential Equations, 169 (2001), pp. 567–587.
[23] V. Venkatasubramanian, *A Taxonomy of the Dynamics of Large Differential Algebraic Systems such as the Power System*, Ph.D. thesis, Washington University, St. Louis, 1992.

SIAM J. MATH. ANAL.
Vol. 36, No. 3, pp. 691–716

RESULTS ON PARABOLIC EQUATIONS RELATED TO SOME CAFFARELLI–KOHN–NIRENBERG INEQUALITIES*

ANDREA DALL'AGLIO†, DANIELA GIACHETTI†, AND IRENEO PERAL‡

Abstract. In this paper problem

$$\begin{cases} u_t - \operatorname{div}(|x|^{-p\gamma}|\nabla u|^{p-2}\nabla u) = \lambda \dfrac{u^{p-2}u}{|x|^{p(\gamma+1)}} & \text{in } \Omega \times (0,\infty),\ 0 \in \Omega, \\ u(x,t) = 0 \quad \text{on } \partial\Omega \times (0,\infty), \\ u(x,0) = \psi(x) \geq 0 \end{cases} \tag{0.1}$$

is studied when $1 < p < N$, $-\infty < (\gamma+1) < \frac{N}{p}$, and under hypotheses on the initial data.

Key words. nonlinear parabolic equations, p-laplacian, existence, behavior of solutions, critical problems, Caffarelli–Kohn–Nirenberg inequalities

AMS subject classifications. 35K25, 35K55, 35K57, 35K65, 46E30, 46E35

DOI. 10.1137/S0036141003432353

1. Introduction. The results by Baras and Goldstein in [7] concerning a blow-up for the solution to the heat equation with a critical potential of the type

$$\begin{cases} u_t - \Delta u = \lambda \dfrac{u}{|x|^2} & \text{in } \Omega \times (0,\infty),\ 0 \in \Omega, \\ u(x,t) = 0 \quad \text{on } \partial\Omega \times (0,\infty), \\ u(x,0) = \psi(x) \geq 0 \end{cases} \tag{1.1}$$

have attracted in recent years the interest of research on some related problems. Roughly speaking, apparently, the main ingredient of the problem studied by Baras and Goldstein is a classical Hardy inequality,

$$\int_{\mathbb{R}^n} \frac{|u|^2}{|x|^2}\,dx \leq C_N \int_{\mathbb{R}^n} |\nabla u|^2\,dx, \tag{1.2}$$

where $C_N = (\frac{2}{N-2})^2$ is the optimal constant that is not achieved in the Sobolev space $\mathcal{D}^{1,2}(\mathbb{R}^n)$. For problem (1.1) Baras and Goldstein have proved that if $\lambda \leq C_N^{-1}$, then there exists a global solution if the initial datum is in a convenient class, while if $\lambda > C_N^{-1}$, there is no solution in the sense that if we consider the solutions u_n of the problems with truncated potential $W_n(x) = \min\{n, |x|^{-2}\}$, then

$$\lim_{n\to\infty} u_n(x,t) = +\infty \quad \text{for all } (x,t) \in \Omega \times \mathbb{R}^+.$$

*Received by the editors July 21, 2003; accepted for publication (in revised form) February 20, 2004; published electronically September 24, 2004. This work was realized while the third author was a visitor of *Dipartimento di Modelli e Metodi Matematici per le Scienze Applicate*, Università di Roma La Sapienza, Italy; later the first and second authors visited *Departamento de Matemáticas*, Universidad Autónoma de Madrid, Spain. The authors would like to thank both institutions for their hospitality and support.

http://www.siam.org/journals/sima/36-3/43235.html

†Dipartimento di Metodi e Modelli Matematici per le Scienze Applicate, Università di Roma La Sapienza, Via A. Scarpa 16 I-00161, Rome, Italy (aglio@dmmm.uniroma1.it, giachett@dmmm.uniroma1.it).

‡Departamento de Matemáticas, Universidad Autónoma de Madrid, Campus de Cantoblanco, 28049 Madrid, Spain (ireneo.peral@uam.es). This author was partially supported by Project BFM2001-0183, MCYT (Spain).

We will call this behavior *spectral instantaneous complete blow-up.* On the other hand, we have the following extension of Hardy's inequality:

$$\int_{\mathbb{R}^n} \frac{|u|^p}{|x|^{(\gamma+1)p}}\,dx \le C_{n,p,\gamma} \int_{\mathbb{R}^n} \frac{|\nabla u|^p}{|x|^{\gamma p}}\,dx, \quad -\infty < \gamma < \frac{N-p}{p}. \tag{1.3}$$

This is a particular limit case of the following Caffarelli–Kohn–Nirenberg inequalities which are proven in [13] (see also [14], [4], and [11]).

PROPOSITION 1.1. *Assume that $1 < p < N$. Then there exists a positive constant $C_{N,p,\gamma,q}$ such that, for every $u \in C_0^\infty(\mathbb{R}^N)$,*

$$\left(\int_{\mathbb{R}^n} \frac{|u|^q}{|x|^{\delta q}}\,dx\right)^{p/q} \le C_{N,p,\gamma,q} \int_{\mathbb{R}^n} \frac{|\nabla u|^p}{|x|^{\gamma p}}\,dx, \tag{1.4}$$

where p, q, γ, δ are related by

$$\frac{1}{q} - \frac{\delta}{N} = \frac{1}{p} - \frac{\gamma+1}{N}, \quad \gamma \le \delta \le \gamma + 1, \tag{1.5}$$

and $\delta q < N$, $\gamma p < N$.

Remark 1.2.

(i) Inequality (1.3) holds a fortiori in every open set Ω.

(ii) One can take

$$C_{n,p,\gamma} = \left(\frac{p}{N - p(\gamma+1)}\right)^p \tag{1.6}$$

in (1.3). This choice of $C_{n,p,\gamma}$ is optimal in every open set Ω containing 0. (The arguments are similar to those in [19] for $\gamma = 0$.)

(iii) If $0 \in \Omega$, the optimal constant is never attained in (1.3).

Remark 1.3. The other limit case for inequality (1.4) is for $\delta = \gamma$, and then one obtains a weighted Sobolev inequality

$$\left(\int_{\mathbb{R}^n} \frac{|u|^{p^*}}{|x|^{\gamma p^*}}\,dx\right)^{p/p^*} \le S_{n,p,\gamma} \int_{\mathbb{R}^n} \frac{|\nabla u|^p}{|x|^{\gamma p}}\,dx, \tag{1.7}$$

where $p^* = \frac{pN}{N-p}$.

It is quite natural to study the parabolic equations associated to inequality (1.3); namely, for the same values of p and γ we consider the problem

$$\text{(P)} \qquad \begin{cases} u_t - \operatorname{div}\left(\dfrac{|\nabla u|^{p-2}\nabla u}{|x|^{\gamma p}}\right) = \lambda \dfrac{|u|^{p-2}u}{|x|^{(\gamma+1)p}}, & (x,t) \in \Omega \times (0,T), \\ u(x,t) = 0, & (x,t) \in \partial\Omega \times (0,T), \\ u(x,0) = \psi(x), & x \in \Omega, \end{cases}$$

where we assume that Ω is a bounded domain in $\mathbb{R}^n$ such that $0 \in \Omega$ and $\partial\Omega$ is a $\mathcal{C}^1$ submanifold.

It is clear that the constant (1.6) will play an essential role in what follows, since the behavior of the problem (P) will deeply depend on whether the parameter λ is smaller or greater than the value

$$\lambda_{n,p,\gamma} = \frac{1}{C_{n,p,\gamma}} = \left(\frac{N - p(\gamma+1)}{p}\right)^p. \tag{1.8}$$

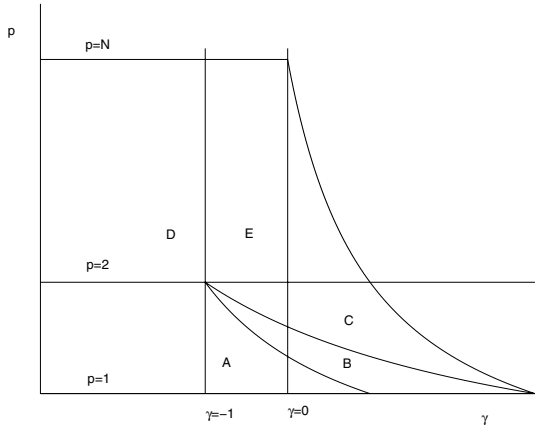

FIG. 1.1. *Summary of the existence and nonexistence results for* $\lambda > \lambda_{N,p,\gamma}$:
Region A: *Global existence of energy solutions.*
Region B: *Global existence of entropy solutions.*
Region C: *Global existence of very weak solutions.*
Region D: *Local existence of solutions.*
Region E: *Instantaneous complete blow-up.*

It could be expected that the behavior for problem (P) should be similar to the one obtained by Baras and Goldstein for (1.1). This conjecture is not completely true. Actually, there is another property which plays an important role in the spectral instantaneous and complete blow-up: a Harnack inequality for the homogeneous parabolic equation. This property is verified if $p \geq 2$ and $(1+\gamma) > 0$. The case $p = 2$ was proved by Chiarenza and Serapioni in [15], while the case $p > 2$ was proved by Abdellaoui and Peral in [1].

The main contribution of this paper is *to show that in the complementary range of the parameters* p *and* γ *we find solutions, even for large values of* λ. The case $p = 2$, $\gamma = 0$ has been studied in [7] and recently in [26]. The case $p \neq 2$, $\gamma = 0$ has been studied in [19] and [5].

The plan of this work is as follows. We begin with section 2, where some notation is provided and appropriate function spaces are defined. Section 3 is devoted to the existence results. In subsection 3.1 we obtain the existence of a global solution in the case $\lambda < \lambda_{N,p,\gamma}$ for all $1 < p < N$. This is the content of Theorem 3.1. In this case the solution belongs to the space $L^p(0,T;\mathcal{D}^{1,p}_{0,\gamma}(\Omega))$, which is naturally related to (P) (see section 2 for the definition). For this reason we will refer to this function u as an *energy solution*. In the proof of Theorem 3.1 we give the details of some convergence results that will be used thereafter. Subsection 3.2 deals with the case $\lambda > \lambda_{N,p,\gamma}$ and $1 < p \leq 2$. The existence of solutions according to the values of γ and p is investigated, and the main results are stated in Theorems 3.3, 3.6, and 3.8. Roughly speaking, as γ and p become larger, we find solutions which are less and less regular. More precisely, we show the following.

1. If $1 < p \leq 2$ and $\gamma + 1 < \frac{N(2-p)}{2p}$ (see region A in Figure 1.1), then we show the existence of energy solutions (see Theorem 3.3).

2. If $1 < p \leq 2$ and $\frac{N(2-p)}{2p} \leq \gamma + 1 < \frac{N(2-p)}{p}$ (see region B), we show the existence of a solution of (P) in the sense of distributions; however, this solution does not belong to the energy space (see Theorem 3.6). We will show that this is an *entropy*

solution in the sense introduced in [8], [22], and [23] for equations with L^1 data (see Definition 3.5 below).

3. If $1 < p \le 2$ and $N(2-p)/(p) < \gamma + 1 < N/p$ (see region C), we show the existence of solutions of (P) in a very weak sense (see Theorem 3.8). We would like to point out that in this case we solve a problem where the right-hand side is not bounded in L^1.

Notice that, comparing the existence results with those contained in [3] for the case $p = 2$, we find that in the nonlinear case (i.e., $p \neq 2$) a very much different behavior of the solutions appears, depending on the parameters; namely, the behavior in cases 2 and 3 above is typical of the nonlinear setting and does not appear in the linear case.

In subsection 3.3, for completeness, we include the elementary local existence result for $p \ge 2$ and $\gamma \le -1$ (see region D in Figure 1.1) in Theorem 3.10, which is also stated in [2].

In section 4 we study the blow-up when $p > 2$, $0 < 1 + \gamma < \frac{N}{p}$, and $\lambda > \lambda_{N,p,\gamma}$ (see region E in Figure 1.1), extending and improving the result of [19] for $\gamma = 0$. (See also [12].)

The case $p = 2$ is obtained in [3] by different kinds of techniques. The main result is Theorem 4.4 and its consequences. The results in Theorems 4.5 and 4.7 have also been stated in [2] and are included here for completeness. With regard to the proof of instantaneous blow-up that we give, it is interesting to point out that for $p > 2$ the blow-up is stronger than that obtained for $p = 2$. Indeed, even the solutions u_n of the problems with truncated potential, $W_n(x) = \min\{n, |x|^{-p(\gamma+1)}\}$, blow up in finite time, and the blow-up time tends to zero as $n \to \infty$.

Finally, in section 5 we study the extinction in finite time of the solution in the case $1 < p < 2$, according to the relation between λ and $\lambda_{N,p,\gamma}$. Roughly speaking, the role that $\lambda_{N,p,\gamma}$ plays in the case $p > 2$ for the blow-up is changed to be a threshold for the finite time extinction property in the case $1 < p < 2$.

2. Notation and function spaces. For $1 < p < \infty$ and $\gamma < \frac{N-p}{p}$, we define the weighted space

$$L^p_\gamma(\Omega) = \left\{ u : \Omega \to \mathbb{R} \text{ measurable, such that } \frac{u(x)}{|x|^\gamma} \in L^p(\Omega) \right\},$$

equipped with the norm

$$\|u\|_{L^p_\gamma(\Omega)} = \left(\int_\Omega \frac{|u(x)|^p}{|x|^{\gamma p}}\, dx \right)^{1/p}.$$

It is easy to check that the dual space $(L^p_\gamma(\Omega))'$ of $L^p_\gamma(\Omega)$ is the space $L^{p'}_{-\gamma}(\Omega)$, where p' is defined by $\frac{1}{p} + \frac{1}{p'} = 1$. Moreover, we define $\mathcal{D}^{1,p}_{0,\gamma}(\Omega)$ as the closure of $C^\infty_0(\Omega)$ in the norm

$$\|u\|_{\mathcal{D}^{1,p}_{0,\gamma}(\Omega)} = \|\nabla u\|_{L^p_\gamma(\Omega)} = \left(\int_\Omega \frac{|\nabla u(x)|^p}{|x|^{\gamma p}}\, dx \right)^{1/p}.$$

As $1 < p < \infty$, $\mathcal{D}^{1,p}_{0,\gamma}(\Omega)$ is reflexive, and we can define the dual space of $\mathcal{D}^{1,p}_{0,\gamma}(\Omega)$, which we will denote by $\mathcal{D}^{-1,p'}_{-\gamma}(\Omega)$, as

$$\mathcal{D}^{-1,p'}_{-\gamma}(\Omega) = \{ G \in \mathcal{D}'(\Omega) : G = \operatorname{div} F,\ F \in L^{p'}_{-\gamma}(\Omega; \mathbb{R}^N) \}.$$

Let us point out that functions in $L^p_\gamma(\Omega)$ do not need to be distributions since they do not belong necessarily to $L^1(\Omega)$. If $\gamma+1 \le -\frac{(p-1)N}{p}$, $\mathcal{D}^{1,p}_{0,\gamma} \not\subset L^1(\Omega)$. The meaning of the gradient in this case is understood as follows. If $u \in \mathcal{D}^{1,p}_{0,\gamma}$ and $\{\phi_n\}_{n\in\mathbb{N}} \subset \mathcal{C}^\infty_0(\Omega)$ is an approximating sequence, then we obtain

$$\nabla\phi_n \to \mathcal{V} \text{ in } L^p_\gamma(\Omega; R^N);$$

in fact, by density and duality we can justify the integration by parts, namely,

$$\int_\Omega \langle \mathcal{V}, \psi\rangle dx = \lim_{n\to\infty} \int_\Omega \langle \nabla\phi_n, \psi\rangle dx = -\int_\Omega u \operatorname{div}(\psi) dx \quad \text{for all } \psi \in \mathcal{D}^{1,p'}_{0,-(\gamma+1)}.$$

As a consequence we define $\operatorname{grad}(u) := \mathcal{V}$. On the other hand, Theorem 1.18 in [20] shows that if $u \in \mathcal{D}^{1,p}_{0,\gamma}$, then the truncature $T_k(u) \in \mathcal{D}^{1,p}_{0,\gamma}(\Omega)$, where $T_k(u)$ is defined by $T_k(u) = u$ if $|u| < k$ and $T_k(u) = k\frac{u}{|u|}$ if $|u| \ge k$. Since $T_k(u) \in L^\infty(\Omega)$, we can define $\nabla T_k(u)$ as a distribution and by Theorem 1.20 in [20] we have

$$\nabla T_k(u) = \operatorname{grad}(u)\chi_{\{|u|<k\}}. \tag{2.1}$$

Hereafter we will denote $\nabla u = \operatorname{grad}(u)$. Notice the relation of this concept of gradient with the one in Lemma 2.1 in [8].

Therefore, inequality (1.4) implies the continuous imbedding

$$\mathcal{D}^{1,p}_{0,\gamma}(\Omega) \subset L^q_\delta(\Omega) \quad \text{for } p, q, \gamma, \delta \text{ satisfying (1.5).} \tag{2.2}$$

This implies, by duality,

$$L^{q'}_{-\delta}(\Omega) \subset \mathcal{D}^{-1,p'}_{-\gamma}(\Omega) \quad \text{for } p, q, \gamma, \delta \text{ satisfying (1.5).} \tag{2.3}$$

We now define the following "evolution" spaces which will be useful in what follows.

$$L^p(0,T;\mathcal{D}^{1,p}_{0,\gamma}(\Omega)) = \{u(x,t) : \Omega\times(0,T) \to \mathbb{R} \text{ measurable}:$$
$$u(\cdot,t) \in \mathcal{D}^{1,p}_{0,\gamma}(\Omega) \text{ for a.e. } t \in (0,T),\ \|u(\cdot,t)\|_{\mathcal{D}^{1,p}_{0,\gamma}(\Omega)} \in L^p(0,T)\},$$

endowed with the norm

$$\|u\|_{L^p(0,T;\mathcal{D}^{1,p}_{0,\gamma}(\Omega))} = \left(\int_0^T \|u(\cdot,t)\|^p_{\mathcal{D}^{1,p}_{0,\gamma}(\Omega)} dt\right)^{1/p} = \left(\iint_{Q_T} \frac{|\nabla u|^p}{|x|^{p\gamma}}\, dx\right)^{1/p}.$$

The dual space of $L^p(0,T;\mathcal{D}^{1,p}_{0,\gamma}(\Omega))$ is $L^{p'}(0,T;\mathcal{D}^{-1,p'}_{-\gamma}(\Omega))$. Let us point out that

$$\mathcal{D}^{1,p}_{0,\gamma}(\Omega) \subset L^q_\delta(\Omega) \quad \text{compactly}$$

for every p, q, γ, δ satisfying $\frac{1}{q} - \frac{\delta}{N} > \frac{1}{p} - \frac{\gamma+1}{N}$ with $\gamma \le \delta \le \gamma+1$ and $\delta q < N$, $\gamma p < N$.

Indeed, a sequence $\{u_n\}$ which is bounded in $\mathcal{D}^{1,p}_{0,\gamma}(\Omega)$ has a subsequence, again denoted by $\{u_n\}$, which converges almost everywhere in Ω to a function $u \in L^q_\delta(\Omega)$. Moreover, by Hölder's inequality and (1.7), for every measurable subset $E \subset \Omega$,

$$\int_E \frac{|u_n-u|^q}{|x|^{\delta q}}\, dx \le \left(\int_\Omega \frac{|u_n-u|^{p^*}}{|x|^{\gamma p^*}}\, dx\right)^{q/p^*} \left(\int_E \frac{1}{|x|^{(\delta-\gamma)\frac{qp^*}{p^*-q}}}\, dx\right)^{(p^*-q)/p^*}$$
$$\le c \left(\int_E \frac{1}{|x|^{(\delta-\gamma)\frac{qp^*}{p^*-q}}}\, dx\right)^{(p^*-q)/p^*}.$$

Since the function in the last integral is an L^1 function, we get the compactness result by Vitali's theorem.

It is easy to see that the operator defined by

$$-\Delta_{p,\gamma}u = -\mathrm{div}\left(\frac{|\nabla u|^{p-2}\nabla u}{|x|^{p\gamma}}\right)$$

maps $\mathcal{D}^{1,p}_{0,\gamma}(\Omega)$ into its dual $\mathcal{D}^{-1,p'}_{-\gamma}(\Omega)$ and is hemicontinuous, coercive, and monotone. (See [21].)

In what follows, we will often use the following result, which is an easy application of Theorem 1.2 of [21] and the reference [24] for the continuity with respect to the time of the L^2-norm.

PROPOSITION 2.1. *If $f \in L^{p'}(0,T;\mathcal{D}^{-1,p'}_{-\gamma}(\Omega))$, $\psi \in L^2(\Omega)$, then there exists a unique solution in the distribution sense, $u \in L^p(0,T;\mathcal{D}^{1,p}_{0,\gamma}(\Omega)) \cap \mathcal{C}^0(0,T;L^2(\Omega))$, of the following problem:*

$$\begin{cases} u_t - \Delta_{p,\gamma}u = f & \text{in } \Omega\times(0,T)\,, \\ u(x,t) = 0 & \text{in } \partial\Omega\times(0,T)\,, \\ u(x,0) = \psi(x) & \text{in } \Omega. \end{cases}$$

We have the following result about the boundedness of the solutions.

LEMMA 2.2. *Let $u \in L^p(0,T;\mathcal{D}^{1,p}_{0,\gamma}(\Omega)) \cap \mathcal{C}^0(0,T;L^2(\Omega))$ be a distributional solution of* (F) *(see section 2), with $\psi \in L^\infty(\Omega)$, and assume that there exist two constants q and β_0 such that*

$$q > \frac{N}{p}, \qquad \beta_0 < p\gamma, \qquad \operatorname*{ess\,sup}_t \int_\Omega |f(x,t)|^q |x|^{\beta_0 q}\,dx < +\infty. \tag{2.4}$$

Then $u \in L^\infty(Q_T)$.

The proof is a slight modification of the classical arguments and is omitted.

3. Existence results. We start with the simpler case $\lambda < \lambda_{N,p,\gamma}$, where $\lambda_{N,p,\gamma}$ is defined by (1.8).

3.1. The case $\boldsymbol{\lambda < \lambda_{N,p,\gamma}}$: Global existence. As usual we denote by $T_n(s)$ the truncation function, i.e., $T_n(s) = s$ if $|s| < n$, $T_n(s) = n\,\mathrm{sign}\,s$ if $|s| > n$. Let us observe that in this range for λ the operator $-\Delta_{p,\gamma} - \lambda\frac{|u|^{p-2}u}{|x|^{p(\gamma+1)}}$ is coercive in the space $\mathcal{D}^{1,p}_{0,\gamma}(\Omega)$. This essentially justifies the following.

THEOREM 3.1. *If $1 < p < N$, $\gamma < \frac{N-p}{p}$, $\lambda < \lambda_{N,p,\gamma}$, $\psi(x) \in L^2(\Omega)$, there exists one distributional solution u for problem* (P). *Moreover, $u \in L^p(0,T;\mathcal{D}^{1,p}_{0,\gamma}(\Omega)) \cap \mathcal{C}^0(0,T;L^2(\Omega))$.*

Proof. Define

$$w_n(x) = \begin{cases} |x|^{-p\gamma} & \text{if } \gamma \geq 0, \\ |x|^{-p\gamma} + \dfrac{1}{n} & \text{if } \gamma < 0, \end{cases}$$

$$f_n(x,u) = \begin{cases} \dfrac{T_n(|u|^{p-2}u)}{|x|^{p(\gamma+1)} + \frac{1}{n}} & \text{if } \gamma \geq 0, \\ \dfrac{T_n(|u|^{p-2}u)}{|x|^{p\gamma}(|x|^p + \frac{1}{n})} & \text{if } \gamma < 0. \end{cases}$$

Let us first consider the following approximate problems:

$$(\mathrm{P}_n)\quad \begin{cases} (u_n)_t - \operatorname{div}\left(w_n(x)|\nabla u_n|^{p-2}\nabla u_n\right) = \lambda f_n(x,u_n), & (x,t)\in\Omega\times(0,T),\\ u_n(x,t)=0, & (x,t)\in\partial\Omega\times(0,T),\\ u_n(x,0)=T_n(\psi(x)), & x\in\Omega. \end{cases}$$

By Proposition 2.1 of section 2 and Schauder's fixed point theorem, it is quite easy to get existence of a solution $u_n \in W_0^{1,p}(\Omega)\cap L^\infty(Q_T)$. Let us multiply (P_n) by $u_n(x,t)$. Using inequality (1.4), one obtains

$$\begin{aligned}\iint_{Q_T}\frac{\partial u_n}{\partial t}u_n + \iint_{Q_T} w_n(x)|\nabla u_n|^p &\le \lambda\iint_{Q_T} f_n(x,u_n)u_n\\ &\le \lambda\iint_{Q_T}\frac{|u_n|^p}{|x|^{p(\gamma+1)}} \le \frac{\lambda}{\lambda_{N,p,\gamma}}\iint_{Q_T}\frac{|\nabla u_n|^p}{|x|^{p\gamma}},\end{aligned}$$

where the first integral is understood as a duality product. Since $\lambda<\lambda_{N,p,\gamma}$, we get the estimates

$$\|u_n\|_{L^\infty(0,T;L^2(\Omega))}\le c_1, \tag{3.1}$$

$$\iint_{Q_T}\frac{|\nabla u_n|^p}{|x|^{p\gamma}}\,dx\,dt\le c_2, \tag{3.2}$$

that is,

$$\|u_n\|_{L^p(0,T;\mathcal{D}^{1,p}_{0,\gamma}(\Omega))}\le c_3. \tag{3.3}$$

Therefore, there exist a function $u\in L^p(0,T;\mathcal{D}^{1,p}_{0,\gamma}(\Omega))\cap L^\infty(0,T;L^2(\Omega))$ and a subsequence (still denoted by u_n) such that $u_n \rightharpoonup u$ weakly in $L^p(0,T;\mathcal{D}^{1,p}_{0,\gamma}(\Omega))$ and $*$-weakly in $L^\infty(0,T;L^2(\Omega))$.

Moreover, if B_ε is the sphere centered in the origin with radius ε, we also have

$$\|u_n\|_{L^p(0,T;W^{1,p}(\Omega\setminus B_\varepsilon))}\le c_4(\varepsilon) \tag{3.4}$$

for every $\varepsilon>0$. By (P_n) we also deduce

$$\left\|\frac{\partial u_n}{\partial t}\right\|_{L^{p'}(0,T;W^{-1,p'}(\Omega\setminus B_\varepsilon))}\le c_5(\varepsilon). \tag{3.5}$$

Using a compactness Aubin-type result (see, for instance, [24]), by (3.4) and (3.5) we can assume that $u_n\to u$ strongly in $L^p((\Omega_\varepsilon)\times(0,T))$ for every $\varepsilon>0$, and therefore, up to a subsequence,

$$u_n\to u \quad \text{a.e. and in measure in } Q_T. \tag{3.6}$$

Let us now prove that, for every $\varepsilon>0$, if we define

$$Q_T^{(\varepsilon)} = (\Omega\setminus B_\varepsilon)\times(0,T),$$

then

$$\nabla u_n\to\nabla u \quad \text{in measure on } Q_T^{(\varepsilon)}. \tag{3.7}$$

To do this, we follow a technique similar to the one introduced by Boccardo and Murat in [10]. Let us define, for $h > 0$, the set

$$H_h = H_{h,m,n} = \{(x,t) \in Q_T^{(\varepsilon)} : |\nabla u_n - \nabla u_m| > h\}.$$

We are going to prove that, for every $\delta > 0$, one has $\operatorname{meas} H_h < \delta$ for m and n large enough. Then, if we set, for positive A, k,

$$\begin{aligned}
\Gamma(n,A) &= \{(x,t) \in Q_T^{(\varepsilon)} : |\nabla u_n| > A\},\\
\Lambda(k) &= \{(x,t) \in Q_T^{(\varepsilon)} : |u_n - u_m| > k\},\\
D(A,k,h) &= \{(x,t) \in Q_T^{(\varepsilon)} : |\nabla u_n - \nabla u_m| > h,\\
&\qquad |\nabla u_n| \le A,\ |\nabla u_m| \le A,\ |u_n - u_m| \le k\},
\end{aligned}$$

then

$$H_h \subset \Gamma(n,A) \cup \Gamma(m,A) \cup \Lambda(k) \cup D(A,k,h).$$

For every $n \in \mathbb{N}$, $\operatorname{meas}\Gamma(n,A)$ is small for A large enough, uniformly in n, since $|\nabla u_n|^q$ is bounded in $L^1(Q_T)$ for every $q < Np/(N-\gamma p)$. Indeed

$$\iint_{Q_T} |\nabla u_n|^q = \iint_{Q_T} \frac{|\nabla u_n|^q}{|x|^{\gamma q}} |x|^{\gamma q} \le \left(\iint_{Q_T} \frac{|\nabla u_n|^p}{|x|^{\gamma p}} |x|^{\gamma q} \right)^{\frac{q}{p}} \left(\iint_{Q_T} |x|^{\frac{\gamma p q}{p-q}} \right)^{\frac{p-q}{p}}, \tag{3.8}$$

and the last integral is finite. Moreover, by (3.6), for every fixed k, $\operatorname{meas}\Lambda(k)$ is small if n, m are large enough. We now consider the set $D(A,k,h)$. By multiplying by $\varphi(x) T_k(u_n - u_m)$ the equations satisfied by u_n and u_m, respectively, where $\varphi(x) \in C_0^\infty(\Omega)$, $\varphi(x) \equiv 0$ for $|x| \le \varepsilon/2$, and $\varphi(x) \equiv 1$ for $|x| \ge \varepsilon$, one obtains, since the integral involving the time-derivative is positive,

$$\begin{aligned}
&\iint_{Q_T^{(\varepsilon/2)}} \frac{|\nabla u_n|^{p-2}\nabla u_n - |\nabla u_m|^{p-2}\nabla u_m}{|x|^{p\gamma}} \nabla T_k(u_n - u_m)\varphi(x)\\
&\qquad \le \lambda k \iint_{Q_T^{(\varepsilon/2)}} \frac{|u_n|^{p-1} + |u_m|^{p-1}}{|x|^{p(\gamma+1)}} + k \iint_{Q_T^{(\varepsilon/2)}} \frac{|\nabla u_n|^{p-1} + |\nabla u_m|^{p-1}}{|x|^{p\gamma}} |\nabla \varphi|.
\end{aligned} \tag{3.9}$$

Using Hölder's inequality, (1.4), and (3.3), one checks that the right-hand side of (3.9) is bounded by $c_6 k$, where c_6 is a constant which only depends on λ, ε, p, N. Since the left-hand side is greater than

$$\varepsilon^{-p\gamma} \iint_{Q_T^{(\varepsilon)} \cap \{|u_n - u_m| \le k\}} \left(|\nabla u_n|^{p-2}\nabla u_n - |\nabla u_m|^{p-2}\nabla u_m\right) \cdot \nabla(u_n - u_m),$$

we have proved that this last integral is small (uniformly in n and m) if k is sufficiently small. Observe now that by the monotonicity and continuity of $|\xi|^{p-2}\xi$, for every $h > 0$, there exists $\mu > 0$ such that

$$\begin{aligned}
D(A,k,h) \subset G(A,k,\mu) = \{(x,t) \in Q_T^{(\varepsilon)} &: |\nabla u_n| \le A,\ |\nabla u_m| \le A,\ |u_n - u_m| \le k,\\
&(|\nabla u_n|^{p-2}\nabla u_n - |\nabla u_m|^{p-2}\nabla u_m) \cdot \nabla(u_n - u_m) > \mu\}.
\end{aligned}$$

It follows that

$$\text{meas}\, D(A,k,h) \le \frac{1}{\mu} \iint_{Q_T^{(\varepsilon)} \cap \{|u_n - u_m| \le k\}} \left(|\nabla u_n|^{p-2}\nabla u_n - |\nabla u_m|^{p-2}\nabla u_m\right) \cdot \nabla(u_n - u_m),$$

so that meas $D(A,k,h)$ is small (uniformly in n and m) if k is sufficiently small. This proves (3.7). We can now pass to the limit in (P_n) in the sense of distributions. Indeed, if we multiply (P_n) by $\varphi(x,t) \in C_0^\infty(Q_T)$, we obtain

$$-\iint_{Q_T} u_n \frac{\partial \varphi}{\partial t} + \iint_{Q_T} \frac{|\nabla u_n|^{p-2}\nabla u_n}{|x|^{p\gamma}} \nabla \varphi = \lambda \iint_{Q_T} T_n\left(\frac{|u_n|^{p-2}u_n}{|x|^{p(\gamma+1)}}\right)\varphi. \tag{3.10}$$

One can easily pass to the limit in each term using the convergences (3.6) and (3.7), the estimates (3.1) and (3.3), the inequality (1.4), and Vitali's theorem. ∎

3.2. The case $\lambda > \lambda_N$, , $p \le 2$: Global existence. In this section we will suppose $\lambda > \lambda_{N,p,\gamma}$ and $p \le 2$. We will show the existence of solutions with different behaviors (see Theorems 3.3, 3.6, and 3.8 in subsections 3.2.1, 3.2.2, and 3.2.3 below), depending on the range for the parameters γ and p.

More precisely, we will find solutions which become weaker and weaker (from the point of view of regularity) as γ and p increase (see Figure 1.1).

First, let us prove the following lemma which will be useful in what follows. It gives the existence of self-similar solutions $S(x,t)$ of the equation in problem (P) for this range of the parameters.

LEMMA 3.2. *If $\lambda > \lambda_{N,p,\gamma}$ and $p < 2$, the function*

$$S(x,t) = A \cdot \left(\frac{t}{|x|^{p(\gamma+1)}}\right)^{\frac{1}{2-p}}, \tag{3.11}$$

where $A = A(\lambda,\gamma) > 0$, is such that

$$A^{p-2} = \frac{1}{(2-p)[(p-1)\delta^p - (N - p(\gamma+1))\delta^{p-1} + \lambda]} \quad \text{and} \quad \delta = \frac{p(\gamma+1)}{2-p} \tag{3.12}$$

satisfy the following:

1. *If $\gamma + 1 < \frac{N(2-p)}{2p}$, then $S(\cdot,t) \in \mathcal{D}^{1,p}_\gamma(\Omega)$ and verifies* (P) *in the sense of distributions.*
2. *If $\frac{N(2-p)}{2p} \le \gamma + 1 < \frac{N(2-p)}{p}$, then*
   - (i) *$S(\cdot,t) \in L^q(\Omega)$ for every q such that $1 < q < \frac{N(2-p)}{p(\gamma+1)}$;*
   - (ii) *$\nabla S(\cdot,t) \in L^{q_1}(\Omega)$ for every q such that $0 < q_1 < \frac{N(2-p)}{2+p\gamma}$;*
   - (iii) *$\nabla S(\cdot,t) \in L^q_\gamma(\Omega)$ for every q such that $0 < q < \frac{N(2-p)}{2(\gamma+1)}$;*
   - (iv) *$\frac{|\nabla S(\cdot,t)|^{p-1}}{|x|^{p\gamma}}$, $\frac{S(\cdot,t)^{p-1}}{|x|^{p(\gamma+1)}} \in L^1(\Omega)$;*
   - (v) *S solves* (P) *in the sense of distributions.*
3. *If $N\frac{(2-p)}{p} \le (\gamma+1) < \frac{N}{p}$, then S solves* (P) *in $\mathcal{D}'(\mathbb{R}^N \setminus \{0\} \times (0,\infty))$ (and in some weighted Sobolev spaces that will be made precise later).*

Proof. We start by looking for solutions of (P) of the form

$$S(x,t) = t^\alpha f(r), \quad \text{with } r = |x|.$$

Choosing the exponent $\alpha = 1/(2-p)$, one can cancel the variable t from the equation, getting the following ordinary differential equation for $f(r)$:

$$\alpha f = (p-1)r^{-p\gamma}|f'|^{p-2}f'' + r^{-(p\gamma+1)}(N - (p\gamma+1))|f'|^{p-2}f' + \lambda r^{-p(\gamma+1)}|f|^{p-2}f. \tag{3.13}$$

Next we look for solutions $f(r)$ of the form

$$f(r) = Ar^{-\delta}, \quad A > 0.$$

It is easy to check that if we choose δ as in (3.12), we can cancel the terms involving powers of r in (3.13), getting solutions of the form (3.11), provided the constant A is defined as in (3.12) and is positive. This last assertion is true if

$$\lambda > \left(\frac{p(\gamma+1)}{2-p}\right)^p (s-1) = \mu_{p,\gamma},$$

where

$$s = \frac{N(2-p)}{p(\gamma+1)}.$$

Let us observe that the critical value $\lambda_{N,p,\gamma}$ can be rewritten as

$$\lambda_{N,p,\gamma} = \left(\frac{p-2+s}{2-p}(\gamma+1)\right)^p.$$

Moreover, if we regard the constants $\lambda_{N,p,\gamma}$ and $\mu_{N,p,\gamma}$ as functions of the variable s,

$$\lambda_{N,p,\gamma}(2) = \mu_{N,p,\gamma}(2), \qquad \lambda'_{N,p,\gamma}(2) = \mu'_{N,p,\gamma}(2), \qquad \lambda''_{N,p,\gamma}(s) > 0 \quad \text{for } s \geq 2-p,$$

which implies $\lambda_{N,p,\gamma} \geq \mu_{N,p,\gamma}$, since $s > 2-p$. Therefore, for $\lambda \geq \lambda_{N,p,\gamma}$ we have $A > 0$, and we obtain the existence of a positive solution $S(x,t)$. The regularity of S stated in the lemma is an easy calculation from the explicit expression of S. It is also easy to see that, if $\gamma + 1 < N(2-p)/p$, then $S(x,t)$ is a solution of (P) in the sense of distributions. □

We can summarize the results about S for $1 < p < 2$ as follows.

(a) If $\gamma+1 < \frac{N(2-p)}{2p}$, $S(x,t)$ is an energy solution; i.e., $S(x,t) \in L^p(0,T;\mathcal{D}^{1,p}_{0,\gamma}(\Omega)) \cap \mathcal{C}^0(0,T;L^2(\Omega))$.

(b) If $\frac{N(2-p)}{2p} \leq \gamma + 1 < \frac{N(2-p)}{p}$, $S(x,t)$ is an *entropy solution* (see Definition 3.5 in subsection 3.2.2).

(c) If $\frac{N(2-p)}{p} \leq \gamma + 1 < \frac{N}{p}$, $S(x,t)$ is a *very weak solution* (see Theorem 3.8, below).

We will prove that the regularity of the self-similar solution gives the behavior of the solutions for the initial value problem in each interval of the parameters. Notice that behavior means that, if $1 < p < 2$, then, for all $\gamma \in (-\infty, \frac{N-p}{p})$, the *spectral instantaneous and complete blow-up* as in Baras–Goldstein does not occur. Namely, there exist solutions with different meanings for all λ.

Let us point out that, if $p = 2$, all the previous critical values collapse to $1+\gamma = 0$, and we will find that for $1 + \gamma \leq 0$ there exist solutions in the energy sense. Note that in this case, by linearity, we obtain *global solutions*. Hence, also in this case, the *spectral instantaneous and complete blow-up* does not occur.

Moreover, if $p > 2$ and $1+\gamma \leq 0$, an argument of comparison allows us to conclude that there exists at least a local (in time) solution.

The remaining question about the behavior in the case $p \geq 2$, $\frac{N}{p} > 1+\gamma > 0$ will be discussed in section 4.

3.2.1. The case $\lambda > \lambda_N$, , , $p \le 2$, $\gamma + 1 < N(2-p)/(2p)$: Global existence of solutions with finite energy.

THEOREM 3.3. *If $\lambda > \lambda_{N,p,\gamma}$, $1 < p \le 2$, $\gamma + 1 < \frac{N(2-p)}{2p}$, $\psi(x) \in L^2(\Omega)$, then there exists a distributional solution u of problem* (P) *such that*

$$u \in L^p(0,T;\mathcal{D}^{1,p}_{0,\gamma}(\Omega)) \cap L^\infty(0,T;L^2(\Omega)).$$

Proof. Let us consider the approximate problems (P_n) defined in the proof of Theorem 3.1. Using $u_n(x,t)$ as test function in (P_n), we get

$$\frac{1}{2}\int_\Omega u_n^2(x,\tau)\,dx + \iint_{Q_\tau} \frac{|\nabla u_n|^p}{|x|^{p\gamma}} \le \lambda \iint_{Q_\tau} \frac{|u_n|^p}{|x|^{p(\gamma+1)}} - \frac{1}{2}\int_\Omega \psi^2(x)\,dx.$$

If $p < 2$, one has

$$\iint_{Q_\tau} \frac{|u_n|^p}{|x|^{p(\gamma+1)}} \le \iint_{Q_\tau} u_n^2 + c_1 T \int_\Omega \frac{dx}{|x|^{2p(\gamma+1)/(2-p)}},$$

where $c_1 = c_1(p)$. The last integral is finite by the hypotheses on γ. If $p = 2$, then necessarily $\gamma + 1 < 0$, and therefore

$$\iint_{Q_\tau} \frac{|u_n|^p}{|x|^{p(\gamma+1)}} \le c_2 \iint_{Q_\tau} u_n^2,$$

with $c_2 = c_2(\Omega,\gamma)$. In both cases, by Gronwall's lemma, we obtain the estimates (3.1)–(3.3), and we can conclude the proof exactly as for Theorem 3.1. ∎

Remark 3.4. Note that actually, in the proof of this theorem, λ can be any real number, since the principal part of the operator is never used to obtain estimates.

3.2.2. The case $\lambda > \lambda_N$, , , $p \le 2$, $N(2-p)/(2p) < \gamma + 1 < N(2-p)/p$: Global existence of entropy solutions. We will specify the sense in which we consider solutions in this case.

DEFINITION 3.5. *Assume that $\psi \in L^1(\Omega)$. We say that $u \in C([0,T];L^1(\Omega))$ is an entropy solution to problem* (P) *if $\frac{|u|^{(p-1)}}{|x|^{p(\gamma+1)}} \in L^1(Q_T)$, $T_k(u) \in L^p(0,T;\mathcal{D}^{1,p}_{0,\gamma}(\Omega))$ for all $k > 0$, and*

$$\int_\Omega \Theta_k(u(T) - v(T))\,dx + \int_0^T \langle v_t, T_k(u-v)\rangle\,dt + \iint_{Q_T} \frac{|\nabla u|^{p-2}}{|x|^{p\gamma}} \nabla u \cdot \nabla(T_k(u-v))$$
$$\le \int_\Omega \Theta_k(\psi - v(0))\,dx + \lambda \iint_{Q_T} \frac{|u|^{p-2}u}{|x|^{p(\gamma+1)}} T_k(u-v) \tag{3.14}$$

for all $k > 0$ and $v \in L^p((0,T),\mathcal{D}^{1,p}_{0,\gamma}(\Omega)) \cap L^\infty(Q_T) \cap C([0,T];L^1(\Omega))$ such that $v_t \in L^{p'}((0,T);\mathcal{D}^{-1,p'}_{-\gamma}(\Omega))$, where Θ_k is given by

$$\Theta_k(s) = \int_0^s T_k(t)dt. \tag{3.15}$$

For a general definition and basic properties of entropy solutions, see, for instance, the references [9], [23], and [22].

THEOREM 3.6. *If $\lambda \geq \lambda_{N,p,\gamma}$, $1 < p < 2$, $\frac{N(2-p)}{2p} \leq \gamma + 1 < \frac{N(2-p)}{p}$, while the initial datum $\psi(x)$ satisfies*

$$\psi \in L^q(\Omega) \quad \text{for every } q \text{ such that } 1 < q < \frac{N(2-p)}{p(\gamma+1)},$$

then there exists a distributional solution u of problem (P) *such that*

$$u \in L^\infty(0,T;L^q(\Omega)) \quad \text{for every } q \text{ such that } 1 < q < \frac{N(2-p)}{p(\gamma+1)}, \tag{3.16}$$

$$\frac{|\nabla u|^{q_1}}{|x|^{\gamma q_1}} \in L^1(Q_T) \quad \text{for every } q_1 \text{ such that } 0 < q_1 < \frac{N(2-p)}{2(\gamma+1)}, \tag{3.17}$$

$$\frac{|\nabla u|^{p-1}}{|x|^{p\gamma}},\ \frac{u^{p-1}}{|x|^{p(\gamma+1)}} \in L^1(Q_T). \tag{3.18}$$

Moreover, u is an entropy solution to problem (P).

Proof. Once again, we consider the approximate problems (P_n), and we multiply them by the test function $\Phi(u_n) = [(1+|u_n|)^{1-\mu} - 1]\operatorname{sign} u_n$, with $\mu \in (0,1)$ to be chosen hereafter. If we define

$$\Psi(s) = \int_0^s \Phi(\sigma)\,d\sigma = \frac{(1+|s|)^{2-\mu}-1}{2-\mu} - |s|,$$

we have

$$\Psi(s) \geq c_1(\mu)|s|^{2-\mu} - c_2(\mu). \tag{3.19}$$

Therefore,

$$\begin{aligned}
&\int_\Omega \Psi(u(x,\tau))\,dx + (1-\mu)\iint_{Q_\tau} \frac{|\nabla u_n|^p}{|x|^{\gamma p}} \frac{1}{(1+|u_n|)^\mu} \\
&\qquad \leq \int_\Omega \Psi(\psi(x))\,dx + \lambda \iint_{Q_\tau} \frac{|u_n|^{p-1}}{|x|^{p(\gamma+1)}} (1+|u_n|)^{1-\mu} \\
&\qquad \leq \int_\Omega \Psi(\psi(x))\,dx + c_3 \iint_{Q_\tau} \frac{|u_n|^{p-\mu}+1}{|x|^{p(\gamma+1)}},
\end{aligned} \tag{3.20}$$

where c_3 depends on λ, μ, p. Note that $\Psi(\psi)$ is integrable by the hypothesis on the initial datum. Since $p < 2$, we can estimate the last integral as

$$\iint_{Q_\tau} \frac{|u_n|^{p-\mu}+1}{|x|^{p(\gamma+1)}} \leq c_4 \iint_{Q_\tau} |u_n|^{2-\mu} + c_5 \iint_{Q_\tau} \frac{1}{|x|^{p(\gamma+1)(2-\mu)/(2-p)}}, \tag{3.21}$$

where c_4 and c_5 depend on μ and p. Now we choose μ in such a way that

$$2 - \frac{(2-p)N}{p(\gamma+1)} < \mu < 1. \tag{3.22}$$

This implies that the last integral in (3.21) converges. Using (3.19)–(3.22) and Gronwall's lemma, we obtain the following estimates:

$$\|u_n\|_{L^\infty(0,T;L^q(\Omega))} \le c_6 \quad \text{for every } q \text{ such that } 1 < q < \frac{(2-p)N}{p(\gamma+1)}, \tag{3.23}$$

$$\iint_{Q_T} \frac{|\nabla u_n|^p}{|x|^{\gamma p}} \frac{1}{(1+|u_n|)^\mu} \le c_7 \quad \text{for every } \mu \text{ such that (3.22) holds}, \tag{3.24}$$

$$\iint_{Q_T} \frac{|\nabla u_n|^{q_1}}{|x|^{\gamma q_1}} \le c_8 \quad \text{for every } q_1 \text{ such that } 0 < q_1 < \frac{(2-p)N}{2(\gamma+1)}, \tag{3.25}$$

$$\iint_{Q_T} \frac{|u_n|^{p-\mu}}{|x|^{p(\gamma+1)}} \le c_9 \quad \text{for every } \mu \text{ such that (3.22) holds}. \tag{3.26}$$

Indeed,

$$\iint_{Q_T} \frac{|\nabla u_n|^{q_1}}{|x|^{\gamma q_1}} \le \left(\iint_{Q_T} \frac{|\nabla u_n|^p}{|x|^{\gamma p}} \frac{1}{(1+|u_n|)^\mu} \right)^{q_1/p} \left(\iint_{Q_T} (1+|u_n|)^{\mu q_1/(p-q_1)} \right)^{(p-q_1)/p}.$$

The estimate (3.25) follows from (3.24) and (3.22).

We now show that the sequence $\{u_n\}$ satisfies

$$\iint_{Q_T} \frac{|u_n|^{(p-1)r}}{|x|^{p(\gamma+1)}} \le c_{10} \quad \text{for all } r \text{ such that } 1 \le r < \frac{2-p}{p-1}\left[\frac{N}{p(\gamma+1)} - 1\right], \tag{3.27}$$

$$\iint_{Q_T} \frac{|\nabla u_n|^{(p-1)s}}{|x|^{p\gamma}} \le c_{11} \quad \text{for all } s \text{ such that } 1 \le s < \frac{(N-p\gamma)(2-p)}{(p-1)(2+p\gamma)}. \tag{3.28}$$

Inequality (3.27) follows from (3.26) and (1.4), while (3.28) follows easily from (3.25). We can now pass to the limit in the distributional formulation, as we have done in the proof of Theorem 3.1, using the estimate in $L^{q_1}(0,T;W^{1,q_1}(\Omega \setminus B_\varepsilon))$, which follows from (3.17), for every $\varepsilon > 0$.

The function u is an entropy solution. Indeed, it is easy to prove (taking $T_k(u_n)$ as test function in (P_n)) that $T_k(u_n)$ is bounded in $L^p(0,T;\mathcal{D}^{1,p}_{0,\gamma}(\Omega))$ and (using Vitali's theorem and (3.28)) that $f_n(x,u_n)$ converges to $\frac{u^{p-1}}{|x|^{p(\gamma+1)}}$ strongly in $L^1(Q_T)$.

Then, if we take $T_k(u_n - v)$ as test function in (P_n), with v as in Definition 3.5, we can easily pass to the limit and get the result with the same techniques as in [22]. □

Remark 3.7. As far as the sharpness of the regularity of the solutions found in Theorem 3.6, let us observe that any function of the form $S_{t_0}(x,t) = S(x,t+t_0)$, where S is defined by (3.11), is a solution in the distribution sense of problem (P), with initial data $\psi(x) = S(x,t_0)$, and its regularity is exactly the one we quoted in Theorem 3.6.

3.2.3. The case $\lambda > \lambda_N$, , $p \le 2$, $N(2-p)/(p) < \gamma+1 < N/p$: Global existence of very weak solutions. We point out that for every $t > 0$ the singular solution $S(x,t)$ is continuous with respect to t with values in $L^2_{-\alpha p/2}(\Omega)$ for every α such that

$$\alpha > \frac{2(\gamma+1)}{2-p} - \frac{N}{p}. \tag{3.29}$$

This, together with the previous estimates on S, suggests the definition of the following space:

(3.30)
$$\mathcal{Y}_\alpha = \{u \in L^p(0,T;\mathcal{D}^{1,p}_{0,\gamma-\alpha}(\Omega)) \cap C^0([0,T];L^2_{-\frac{\alpha p}{2}}(\Omega)) : u' \in L^{p'}(0,T;\mathcal{D}^{-1,p'}_{-\beta}(\Omega))\},$$

where

$$\beta = \gamma + \alpha(p-1). \tag{3.31}$$

The following theorem specifies the meaning of a very weak solution.

THEOREM 3.8. *Assume that $\lambda > \lambda_{N,p,\gamma}$, $1 < p < 2$, $\frac{N(2-p)}{p} \le \gamma + 1 < \frac{N}{p}$, and that the initial data $\psi(x)$ belongs to $L^2_{-\frac{\alpha p}{2}}(\Omega)$ for some α satisfying* (3.29). *Then there exists a function $u \in \mathcal{Y}_\alpha$ which is a distributional solution of* (P) *away from the origin (that is, in $\mathcal{D}'((\Omega \setminus \{0\}) \times (0,T))$). Moreover, u is a solution of* (P) *in the following sense:*

$$\begin{aligned} &-\int_0^\tau \langle v', |x|^{\alpha p}u\rangle\, dt + \int_\Omega u(\tau)v(\tau)|x|^{\alpha p}\,dx - \int_\Omega \psi v(0)|x|^{\alpha p}\,dx \\ &+ \iint_{Q_\tau} \frac{|\nabla u|^{p-2}\nabla u\cdot\nabla(v|x|^{\alpha p})}{|x|^{\gamma p}}\,dx\,dt = \iint_{Q_\tau} \frac{|u|^{p-2}uv|x|^{\alpha p}}{|x|^{(\gamma+1)p}}\,dx\,dt \end{aligned} \tag{3.32}$$

for every $\tau \in [0,T]$ and for every $v \in \mathcal{Y}_\alpha$.

Proof. Step 1: *A priori estimate.* Let u_n be a solution of problem (P_n). We use $|x|^{\alpha p}u_n(x,t)$ as test function in (P_n). Then, by Young's inequality,

$$\begin{aligned} &\int_\Omega u_n^2(x,T)|x|^{\alpha p}\,dx + \iint_{Q_T} |\nabla u_n|^p |x|^{(\alpha-\gamma)p} \\ &\quad\le c_1 \iint_{Q_T} |\nabla u_n|^{p-1}|x|^{(\alpha-\gamma)p-1} + \lambda \iint_{Q_T} \frac{|u_n|^p}{|x|^{p(\gamma+1-\alpha)}} + \frac{1}{2}\int_\Omega \psi^2(x)|x|^{\alpha p}\,dx \\ &\quad\le \frac{1}{2}\iint_{Q_T} |\nabla u_n|^p|x|^{(\alpha-\gamma)p} + c_3 \iint_{Q_T} |u_n|^2|x|^{\alpha p} + c_3\int_\Omega |x|^{p\left(\alpha-\frac{2(\gamma+1)}{2-p}\right)} \\ &\qquad+ \frac{1}{2}\int_\Omega \psi^2(x)|x|^{\alpha p}\,dx. \end{aligned}$$

Under the hypotheses on α and on the initial datum, the last two integrals are finite. Therefore, we can use Gronwall's lemma to conclude that

$$u_n \text{ is bounded in } L^p(0,T;\mathcal{D}^{1,p}_{0,\gamma-\alpha}(\Omega)) \cap C^0([0,T];L^2_{-\frac{\alpha p}{2}}(\Omega)).$$

By (P_n), one can easily check that

$$u_n' \text{ is bounded in } L^{p'}(0,T;\mathcal{D}^{-1,p'}_{-\beta}(\Omega)).$$

Step 2: *Passage to the limit.* By weak convergence, and following the same argument as in the proof of Theorem 3.1 for the pointwise convergence of the gradients, we obtain a function $u \in L^p(0,T;\mathcal{D}^{1,p}_{0,\gamma-\alpha}(\Omega)) \cap L^\infty(0,T;L^2_{-\frac{\alpha p}{2}}(\Omega))$, with

$u' \in L^{p'}(0,T;\mathcal{D}^{-1,p'}_{-\beta}(\Omega))$, such that

$$u_n \rightharpoonup u \quad \text{weakly in } L^p(0,T;\mathcal{D}^{1,p}_{0,\gamma-\alpha}(\Omega)),$$
$$u_n \rightharpoonup u \quad *\text{-weakly in } L^\infty(0,T;L^2_{-\frac{\alpha p}{2}}(\Omega)),$$
$$\nabla u_n \to \nabla u \quad \text{almost everywhere in } Q_T,$$
$$u_n(\cdot,\tau) \to u(\cdot,\tau) \quad \text{a.e. in } \Omega \text{ and weakly in } L^2_{-\frac{\alpha p}{2}}(\Omega) \text{ for every } \tau \in [0,T].$$

Using these convergences, one can take $|x|^{\alpha p}v$ as test function in (P_n) and pass to the limit as $n \to \infty$, obtaining the weak formulation (3.32). Since the functions of the form $|x|^{\alpha p}v$ include smooth test functions in $\mathcal{D}(Q_T)$ which are zero in a neighborhood of the origin, we have also proved that u is a solution in the distributional sense far from the origin.

We now prove that $u \in C^0([0,T];L^2_{-\frac{\alpha p}{2}}(\Omega))$. According to the uniform estimates for the approximate solutions, we find that $u_n(\cdot,t)$ is an equicontinuous sequence in $L^2_{-\frac{\alpha p}{2}}(\Omega))$. By the Ascoli–Arzelà lemma, we conclude. □

Remark 3.9.

(i) The previous result, in the case where $\gamma = 0$, improves the result contained in [19] and specifies the meaning of the solution given in that paper; more precisely, it gives us that the solution is in $L^p(0,T;\mathcal{D}^{1,p}_{-\alpha}(\Omega))$ for some $\alpha > 2/(2-p) - N/p$.

(ii) If we define the operator $\Gamma v = |x|^{\alpha p}v$, then Γ is an isomorphism from $\mathcal{D}^{1,p}_{0,\gamma-\alpha}(\Omega)$ to $\mathcal{D}^{1,p}_{0,\beta}(\Omega)$, where $\beta = (p-1)\alpha + \gamma$. Therefore, the weak formulation (3.32) could be rewritten as

$$-\int_0^\tau \langle w', u\rangle\, dt + \int_\Omega u(\tau)w(\tau)\, dx - \int_\Omega \psi w(0)\, dx + \iint_{Q_\tau} \frac{|\nabla u|^{p-2}\nabla u \cdot \nabla w}{|x|^{\gamma p}}\, dx\, dt = \iint_{Q_\tau} \frac{|u|^{p-2}uw}{|x|^{(\gamma+1)p}}\, dx\, dt$$

for every $\tau \in [0,T]$ and for every $w \in L^p(0,T;\mathcal{D}^{1,p}_{0,\beta}(\Omega)) \cap C^0([0,T];L^2_{\frac{\alpha p}{2}}(\Omega))$ such that $w' \in L^{p'}(0,T;\mathcal{D}^{-1,p'}_{\alpha-\gamma}(\Omega))$.

(iii) In the case where the initial data $\psi(x)$ is nonnegative and satisfies

$$\psi(x) \le S(x,t+t_0) \quad \text{for some positive } t_0,$$

it is possible to obtain an alternative (constructive) proof by a monotone iteration argument, using $S(x,t+t_0)$ as a supersolution and solving, by induction, the sequence of problems

$$(\tilde{\mathrm{P}}_n) \qquad \begin{cases} \dfrac{\partial u_n}{\partial t} - \Delta_{p,\gamma} u_n = \lambda T_n\left(\dfrac{1}{|x|^{p(\gamma+1)}}\right) u_{n-1}^{p-1}, & (x,t) \in \Omega \times (0,T), \\ u_n(x,t) = 0, & (x,t) \in \partial\Omega \times (0,T), \\ u_n(x,0) = \psi(x), & x \in \Omega, \end{cases}$$

with $u_0 \equiv 0$.

(iv) The solution found in Theorem 3.10 satisfies the equation in a very weak sense because the right-hand side of the equation does not even belong to L^1.

3.3. The case $\lambda > \lambda_N$, , , $p \geq 2$, $\gamma \leq -1$: Existence for small times. This subsection deals with existence for small values of t in the case $\lambda > \lambda_{N,p,\gamma}$, $p > 2$, $\gamma \leq -1$. The result of this subsection can be compared with the ones of section 4: an instantaneous blow up will occur for the solutions of the approximate problems for the same values of λ and p when $\gamma > -1$.

THEOREM 3.10. *If $\lambda > \lambda_{N,p,\gamma}$, $p \geq 2$, $\gamma \leq -1$, while the initial data $\psi(x)$ satisfies $\psi(x) \in L^\infty(Q_T)$ and $\psi(x) \geq 0$, then there exist $T^* = T^*(N, p, \gamma, \lambda, \|\psi\|_{L^\infty(\Omega)}) > 0$ and a distributional solution u in Q_{T^*} of our problem with $u \in L^p(0, T; \mathcal{D}^{1,p}_{0,\gamma}(\Omega)) \cap L^\infty(0, T; L^2(\Omega))$ for every $T < T^*$. Moreover, if $p = 2$, T^* is any positive value.*

Proof. Let us define the problems $(\tilde{\mathrm{P}}_n)$ as in the previous subsection and let $y(t)$ be the solution of the ordinary differential equation

$$\begin{cases} y'(t) = d\, y^{p-1}, \\ y(0) = \|\psi\|_{L^\infty(\Omega)}, \end{cases}$$

where

$$d \geq \lambda \sup_{x \in \Omega} |x|^{-p(\gamma+1)}. \tag{3.33}$$

An immediate calculation shows the following.

(α) If $p > 2$, the solution is

$$y(t) = \frac{\|\psi\|_{L^\infty(\Omega)}}{(1 - (p-2)d\|\psi\|^{p-2}_{L^\infty(\Omega)} t)^{1/(p-2)}},$$

which blows up in $t = T^* = \frac{1}{(p-2)d\|\psi\|^{p-2}_{L^\infty(\Omega)}}$.

(β) If $p = 2$, then the global solution is

$$y(t) = \|\psi\|_{L^\infty(\Omega)} e^{dt}.$$

Since $y(t)$ is a supersolution of (P), by the comparison principle we have

$$u_1 \leq u_2 \leq \cdots \leq u_n \leq \cdots \leq y.$$

If we multiply problem $(\tilde{\mathrm{P}}_n)$ by $u_n \chi_{(0,\tau)}$, we obtain

$$\frac{1}{2} \int_\Omega u_n^2(x, \tau)\, dx + \iint_{Q_T} \frac{|\nabla u|^p}{|x|^{p\gamma}} \leq \lambda \iint_{Q_T} |y|^{p-1} |x|^{-p(\gamma+1)} + \frac{1}{2} \int_\Omega \psi^2(x)\, dx.$$

By condition (3.33),

$$\lambda \iint_{Q_T} |y|^{p-1} |x|^{-p(\gamma+1)} \leq \operatorname{meas} \Omega (y(\tau) - \|\psi\|_{L^\infty(\Omega)}).$$

Therefore, we get the estimates

$$\|u_n\|_{L^\infty(0,\tau;L^2(\Omega))} \leq c_1, \quad \|u_n\|_{L^p(0,\tau;\mathcal{D}^{1,p}_{0,\gamma}(\Omega))} \leq c_2 \quad \text{for every } \tau < T^*.$$

In the case $p = 2$, we can fix any $T^* > 0$ to get the same estimates. Now the conclusion follows exactly as in the proof of Theorem 3.1. ∎

4. Blow-up: $p > 2$, $N/p > (1+\gamma) > 0$, and $\lambda > \lambda$, , . We consider in this section the *spectral, instantaneous, and complete blow-up* in the case $p > 2$ and $(1+\gamma) > 0$. The case $p = 2$ has been obtained in [3] and requires a different method. We would like to point out that in the case $p > 2$ a stronger result than in the linear case is obtained. This behavior is given because even the problem with the truncated potential blows up in finite time. We will assume that the initial data verifies that $\psi \in L^2(\Omega)$ and there exists $\delta > 0$ such that $\psi > 0$ in $B_\delta(0)$. Notice that for the equation

$$u_t - \Delta_{p,\gamma} u = 0 \tag{4.1}$$

and by direct calculations we can find Barenblatt-type solutions; precisely,

$$\mathcal{B}(x,t) = t^{-N\beta(N,p,\gamma)} \left[M - \frac{(p-2)\beta(N,p,\gamma)^{\frac{1}{p-1}}}{p(\gamma+1)} \xi^{\frac{p(\gamma+1)}{p-1}} \right]_+^{\frac{(p-2)}{(p-1)}}, \tag{4.2}$$

where M is a positive arbitrary constant,

$$\beta(N,p,\gamma) = \frac{1}{N(p-2)+p(\gamma+1)}, \quad \text{and} \quad \xi = \frac{|x|}{t^{\beta(N,p,\gamma)}}.$$

This property could be understood as some kind of *finite speed of propagation* for the equation with zero right-hand side. It is necessary to point out that if $\gamma \neq 0$, the equation is not invariant by translation, and then the corresponding translated Barenblatt functions are not solutions to the equation.

The lack of homogeneity in (4.1) provides the following weak Harnack inequality.

LEMMA 4.1. *Let u be a nonnegative weak solution to* (4.1)*, and assume that $u(x_0,t_0) > 0$ for some $(x_0,t_0) \in \Omega_T$; then there exists $B(N,p,\gamma) > 1$ such that, for all $\theta, \rho > 0$ satisfying $B_{4\rho}(x_0) \times (t_0 - 4\theta, t_0 + 4\theta) \subset \Omega_T$, we have*

$$\begin{aligned} &\frac{1}{|B_\rho(x_0)|} \int_{B_\rho(x_0)} u(x,t_0)\,dx \\ &\quad \le B \left[\left(\frac{\rho^{p(\gamma+1)}}{\theta} \right)^{\frac{1}{p-2}} + \left(\frac{\theta}{\rho^{p(\gamma+1)}} \right)^{\frac{N}{p(\gamma+1)}} \left(\inf_{B_\rho(x_0)} u(.,t_0+\theta) \right)^{\frac{\lambda_\gamma}{p(\gamma+1)}} \right], \end{aligned} \tag{4.3}$$

where $\lambda_\gamma = N(p-2) + p(\gamma+1) = \frac{1}{\beta(N,p,\gamma)}$.

The proof is similar to the one by DiBenedetto in [17] for the case $\gamma = 0$. The details can be found in [1] in the case $(1+\gamma) > 0$, where some counterexamples to the Harnack inequality if $(1+\gamma) \le 0$ are shown.

We consider problem (P), and we make the following assumptions:

(H1) $p > 2$, $0 < 1 + \gamma < N/p$, and $\lambda > \lambda_{n,p,\gamma}$.

(H2) $\psi \in L^\infty(\Omega)$, $\psi(x) \ge 0$, and moreover, there exists $\rho, \delta > 0$ such that $\psi(x) > \delta$ for every $x \in B_\rho(0)$.

We will prove that problem (P) has no solution. We start by studying, for $n \in \mathbb{N}$, the approximate problems

$$\begin{cases} (u_n)_t - \Delta_{p,\gamma} u_n = \lambda W_n(x) |u_n|^{p-2} u_n & \text{in } Q_T, \\ u(x,t) = 0 & \text{on } \partial\Omega \times (0,T), \\ u(x,0) = \psi(x) & \text{in } \Omega, \end{cases} \tag{4.4}$$

where $W_n(x) = T_n(\frac{1}{|x|^{p(\gamma+1)}})$. Note that for every fixed n, problem (4.4) has a solution at least for small times (depending on n and λ), as one can easily see using a convenient supersolution independent of x.

By separation of variables we look for solutions of (4.4) of the form $\Phi(x,t) = \Theta(t)X(x)$, to use as a subsolution. The equation becomes

$$\Theta' X - \Theta^{p-1}\Delta_{p,\gamma} X = \lambda W_n(x)\Theta^{p-1}X^{p-1}.$$

We take the $\Theta(t)$ solution of

$$\begin{cases} \Theta'(t) = \mu\Theta^{p-1}(t), \\ \Theta(0) = A, \end{cases} \tag{4.5}$$

that is,

$$\Theta(t) = \frac{A}{[1-(p-2)\mu A^{p-2}t]^{1/(p-2)}}$$

with $\mu, A > 0$ to be chosen. Note that $\lim_{t\to\tau}\Theta(t) = \infty$ for $\tau = \frac{1}{\mu(p-2)A^{p-2}}$.

On the other hand, $X(x)$ must solve the elliptic problem

$$\begin{cases} -\Delta_{p,\gamma} X = \lambda W_n(x)X^{p-1} - \mu X & \text{in } \Omega, \\ X(x) = 0 & \text{on } \partial\Omega. \end{cases} \tag{4.6}$$

Defining $\alpha X = Y$ with $\mu\alpha^{p-2} = \lambda$ the problem above becomes

$$\begin{cases} -\Delta_{p,\gamma} Y = \lambda(W_n(x)Y^{p-1} - Y) & \text{in } \Omega, \\ Y(x) = 0 & \text{in } \partial\Omega. \end{cases} \tag{4.7}$$

Problem (4.7) fails in the hypotheses for bifurcation from infinity as in [6]; see [16] for details.

Let $\lambda_1(n)$ be the first eigenvalue for the problem

$$\begin{cases} -\Delta_{p,\gamma}\varphi = \lambda W_n(x)|\varphi|^{p-2}\varphi & \text{in } \Omega, \\ \varphi(x) = 0 & \text{in } \partial\Omega. \end{cases}$$

Then (i) $\lambda_1(n) > 0$; (ii) $\lambda_1(n)$ is isolated and simple; (iii) the first eigenfunction does not change sign; (iv) $\lambda_1(n)$ is decreasing in n, and $\lambda_1(n) \searrow \lambda_{N,p,\gamma}$. The properties (i), (ii), and (iii) are similar to the p-*laplacian* case and are detailed in [16]; (iv) is easily checked following the proof for the p-laplacian in [19].

THEOREM 4.2. *If $\lambda > \lambda_{N,p,\gamma}$, then there exists n_0 such that, for every $n > n_0$, there exists a bounded positive solution $Y(x)$ to* (4.7).

Proof. As $\lambda > \lambda_{N,p,\gamma}$ there exists n_0 such that, for $n > n_0$, $\lambda > \lambda_1(n)$. Now $\lambda_1(n)$ is the unique bifurcation point of positive solutions from infinity for problem (4.7). Moreover, as $(1+\gamma) > 0$, the solutions in the branch are bounded; see [16] and [6]. Moreover, if $Y > 0$ is a solution to (4.7), then $\|Y\|_\infty \geq R_n > 0$ for some constant R_n, because if a positive solution Y is such that $\|Y\|_\infty < \varepsilon$, then we have $-\Delta_{p,\gamma} Y \leq \lambda Y(n\varepsilon^{p-2} - 1) < 0$, and for ε small we reach a contradiction with the maximum principle. □

As a consequence we can find a subsolution to problem (4.4) that shows the finite time blow-up. Precisely, we have the following result.

LEMMA 4.3. *Let u be a solution to problem* (4.4)*, where $\lambda > \lambda_1(n)$ and $\psi(x) > 0$ in every $x \in \Omega$. Then there exists $T > 0$ depending on the data and there exists a subsolution Φ such that $u(x,t) \geq \Phi(x,t)$ and $\lim_{t\to T} \Phi(x,t) = \infty$ for every $x \in \Omega$.*

Proof. The solution u is positive and, by regularity (see [1]), is bounded for small times. Therefore, we fix a small time $\tau > 0$, and we look for a subsolution of the form $\Phi(x,t) = X(x)\Theta(t)$, with $X(x)$ the solution of (4.6), and

$$\Theta(t) = \epsilon(1 - (p-2)\varepsilon^{p-2}(t-\tau))^{-1/(p-2)},$$

with $\varepsilon > 0$ such that $\varepsilon X(x) \leq u(x,\tau)$. By the weak comparison principle we conclude. □

In order to show the instantaneous complete blow-up, we need to rescale the problem, using the following property. Define

$$Z_n(x) = \left(\frac{n_0}{n}\right)^{\frac{1}{p-2}} X\left(\left(\frac{n}{n_0}\right)^{\frac{1}{p(\gamma+1)}} x\right). \tag{4.8}$$

Then Z_n solves

$$\begin{cases} -\Delta_{p,\gamma} Z_n = \lambda W_n(x) Z_n^{p-1} - \mu Z_n & \text{if } |x| < \left(\frac{n_0}{n}\right)^{\frac{1}{p(\gamma+1)}}, \\ Z_n(x) = 0 & \text{if } |x| = \left(\frac{n_0}{n}\right)^{\frac{1}{p(\gamma+1)}} \end{cases}$$

since $(\frac{n}{n_0}) W_{n_0}((\frac{n}{n_0})^{\frac{1}{p(\gamma+1)}} x) = W_n(x)$. Moreover, the radius of the ball goes to zero and $\|Z_n\|_\infty \to 0$ as $n \to \infty$. Therefore, for prescribed $R, \eta > 0$ we can choose n such that

$$\left(\frac{n_0}{n}\right)^{\frac{1}{p(\gamma+1)}} < R, \qquad Z_n(x) \leq \eta \quad \text{on } B_R. \tag{4.9}$$

THEOREM 4.4. *Assume that* (H1), (H2) *hold. Then for every $\varepsilon > 0$ there exist $r(\varepsilon) > 0$ and n_ε such that if u_n is the minimal solution to* (4.4) *$\forall n > n_\varepsilon$*

$$u_n(x,t) \equiv +\infty \quad \textit{for} \quad t > \varepsilon \textit{ and } |x| < r(\varepsilon).$$

Proof. Take n_0 such that $\lambda > \lambda_1(n_0)$. We prescribe the blow-up time $T = \varepsilon$ and choose $\mu = [(p-2)\varepsilon]^{-1}$. For such μ and $n > n_0$, the scaled solution (4.8) to (4.4), X_n, satisfies (4.9) with $R = \rho$ and $\eta = \delta$. Consider $\Theta(t)$ solution to (4.5) with μ as above and A = 1. Then $\phi_n(x,t) = \Theta(t) X_n(x)$ blows up in $T = \varepsilon$. By weak comparison in the ball $|x| < (\frac{n_0}{n})^{\frac{1}{p(\gamma+1)}}$, the minimal solution to (4.4) blows up in $T_0 < \varepsilon$. □

We point out that in order to obtain blow-up in a prescribed small time we have to take the index n large enough. We will use the concept of entropy solution introduced in Definition 3.5 and a straightforward modification of the comparison arguments for entropy solutions (see [23]).

THEOREM 4.5. *Assume that* (H1), (H2) *hold. Then problem* (P) *has no entropy solution, even for small times, and moreover, if $u_n(x,t)$ is the minimal solution to* (4.4)*, we have that $\lim_{n\to\infty} u_n(x,t) = +\infty$ for all $(x,t) \in \Omega \times (0,\infty)$.*

Proof. By contradiction, assume that there exists an entropy solution $u(x,t) > 0$ of problem (P). Then u is a supersolution to problem (4.4) for all n. As a consequence the minimal solution to (4.4) satisfies $u_n(x,t) \leq u(x,t)$; hence $u(x,t)$ blows up at least in the time in which u_n blows up, so we conclude.

By using Theorem 4.4 we obtain a region E_∞ such that

$$E_\infty \supset \{|x| < r(t)\} \times (0, \infty),$$

such that

$$\lim_{n\to\infty} u_n(x,t) = +\infty \quad \text{for all } (x,t) \in E_\infty.$$

Next we use the Harnack inequality (4.3), assume that there exists a point $(x_0, t_0) \in \Omega \times (0,\infty)$ such that $0 \le u_n(x_0,t_0) \le M < \infty$, and call

$$\rho(x_0,t_0) = \operatorname{dist}\{x_0, \partial\Omega\} > 0.$$

Then, if $B_r(x_0) \times \{t = t_1\} \cap E_\infty$ has N-dimensional positive measure for some $r < \rho(x_0,t_0)$ and $t_1 < t_0$, we consider the problem

$$\begin{cases} (v_n)_t - \Delta_{p,\gamma} v_n = 0 & \text{in } B_r(x_0) \times (t_1, t_0), \\ v_n(x,t) = 0 & \text{on } \partial B_r(x_0) \times (t_1, t_0), \\ v_n(x,t_1) = u_n(x,t_1) & \text{in } B_r(x_0); \end{cases} \tag{4.10}$$

then $v_n(x,t) \le u_n(x,t)$, and this is a contradiction to the Harnack inequality (4.3). If for all $r < \rho(x_0)$ and all $t_1 < t_0$, $|B_r(x_0) \times \{t = t_1\} \cap E_\infty| = 0$, then for all $\delta > 0$ we can find in a finite number of steps a point $(x_1, t_0 - \delta) \in \Omega \times (0, t_0)$ such that

$$|B_r(x_0) \times \{t = t_1\} \cap E_\infty| > 0,$$

and then we reach a contradiction as above. ∎

Remark 4.6. Notice that this result is stronger, in some sense, than the result by Baras and Goldstein (see [7]) for the heat equation; if $p > 2$, even the solution to the equation with truncated potential blows up in finite time.

Next we will prove that even if we truncate the whole nonlinearity, we find spectral instantaneous complete blow-up. More precisely, we have the following result.

THEOREM 4.7. *Consider the truncated problem*

$$\begin{cases} (v_n)_t - \Delta_{p,\gamma} v_n = \lambda\, W_n(x)\, T_n(v_n^{p-1}) & \text{in } \Omega \times \mathbb{R}^+, \\ v(x,t) = 0 & \text{on } \partial\Omega \times \mathbb{R}^+, \\ v(x,0) = \psi(x) & \text{in } \Omega, \end{cases} \tag{4.11}$$

where (H1) *and* (H2) *hold. Then*

$$\lim_{n\to\infty} v_n(x,t) = +\infty \quad \text{for every } (x,t) \in \Omega \times \mathbb{R}^+.$$

Proof. Using the same argument as in [2], we find that if $B_{4r}(0) \subset \Omega$, then

$$\lim_{n\to\infty} \int_{B_r(0)} v_n(x,t)\, dx = +\infty \quad \text{for every } t > 0.$$

Then by the Harnack inequality and a strategy which is similar to the one in Theorem 4.5, we obtain the complete blow-up. ∎

Remark 4.8.

(i) An alternative method to the one described above can be seen in [1]. The separation of variables gives a more transparent view of the behavior but uses in a strong way the presence of exactly two homogeneities. In the linear case (see [3]), or if the second member is not eigenvalues-like (see [2]), different arguments are needed.

(ii) If instantaneous and complete blow-up happens without hypothesis (H2), this seems to be an open problem. If $\gamma = 0$, we can take as a subsolution a convenient scaled and translated Barenblatt function that allows us to conclude that there exists a $T^* > 0$ such that for $t > T^*$ the same result as in Theorem 4.7 holds.

5. Behavior of solutions in the case $1 < p < 2$ and $\lambda < \lambda_N$, , . In this section we will try to explain how the optimal constant in the Hardy inequality becomes the threshold for extinction in finite time of the solution.

5.1. Finite time extinction.

THEOREM 5.1. *Assume that*

$$\max\left\{\frac{2N}{N+2}, \frac{2N}{N+2(\gamma+1)}\right\} < p < 2,$$

$\lambda < \lambda_{N,p,\gamma}$, *and* $\psi \in L^2(\Omega)$. *Then there exists a constant*

$$T^* = T^*(N,p,\gamma,\lambda,\Omega) \le c_1(N,p,\gamma,\lambda,\Omega)\|\psi\|_{L^2(\Omega)}^{2-p}$$

such that any solution of problem (P) *satisfies*

$$u(\cdot,t) \equiv 0 \quad \text{for } t \ge T^*. \tag{5.1}$$

Proof. Taking u as a test function in (P), and using inequalities (1.3) and (1.7), we get

$$\frac{1}{2}\frac{d}{dt}\int_\Omega u^2(t)\,dx + \frac{1}{S_{N,p,\gamma}}\left(1 - \frac{\lambda}{\lambda_{N,p,\gamma}}\right)\left[\int_\Omega \frac{|u(t)|^{p^*}}{|x|^{\gamma p^*}}\,dx\right]^{\frac{p}{p^*}} \le 0.$$

Using the assumptions on p and γ, by Hölder's inequality we obtain

$$\int_\Omega u^2(t)\,dx \le \left[\int_\Omega \frac{|u(t)|^{p^*}}{|x|^{\gamma p^*}}\,dx\right]^{\frac{2}{p^*}}\left[\int_\Omega |x|^{\frac{2\gamma p^*}{p^*-2}}\,dx\right]^{\frac{p^*-2}{\gamma p^*}} \le c_1\left[\int_\Omega \frac{|u(t)|^{p^*}}{|x|^{\gamma p^*}}\,dx\right]^{\frac{2}{p^*}},$$

where $c_1 = c_1(N,p,\gamma,\Omega)$ is a positive constant. Therefore, setting

$$\phi(t) = \int_\Omega u^2(t)\,dx,$$

one has

$$\phi'(t) + c_2[\phi(t)]^{\frac{p}{2}} \le 0,$$

with $c_2 > 0$. Since $p < 2$, this implies

$$\phi(t) \le \left([\phi(0)]^{\frac{2-p}{2}} - c_3 t\right)_+^{\frac{2}{2-p}},$$

from which the statement follows. ∎

THEOREM 5.2. *Assume that*

$$\gamma \geq 0, \qquad 1 < p < \frac{2N}{N+2},$$
$$\lambda < \eta_{N,p,\gamma} = \left(\frac{N(2-p)}{p} - 1\right)\left(\frac{[N - p(\gamma+1)]p}{(2-p)(N-p)}\right)^p,$$

and

$$\int_\Omega |\psi|^{\frac{N(2-p)}{p}}\, dx < \infty.$$

Then there exists a constant

$$T^* = T^*(N,p,\gamma,\lambda,\Omega) \leq c_1(N,p,\gamma,\lambda,\Omega)\|\psi\|^{2-p}_{L^{\frac{N}{p}(2-p)}(\Omega)}$$

such that any solution of problem (P) *found by approximation as in Theorem* 3.1 *satisfies*

$$u(\cdot,t) \equiv 0 \quad \text{for } t \geq T^*. \tag{5.2}$$

Proof. We take $v_n = |u_n|^{\alpha-2}u_n$ as test function in (P_n), with $\alpha \geq 2$ to be chosen hereafter. We obtain

$$\frac{1}{\alpha}\frac{d}{dt}\int_\Omega u_n^\alpha(t)\,dx + (\alpha-1)\int_\Omega \frac{|\nabla u_n(t)|^p |u_n(t)|^{\alpha-2}}{|x|^{\gamma p}}\,dx = \lambda\int_\Omega \frac{|u_n(t)|^{\alpha-(2-p)}}{|x|^{(\gamma+1)p}}\,dx.$$

Since

$$\int_\Omega \frac{|\nabla u_n(t)|^p |u_n(t)|^{\alpha-2}}{|x|^{\gamma p}}\,dx = \left(\frac{p}{\alpha-(2-p)}\right)^p \int_\Omega \frac{\left|\nabla\left(|u_n(t)|^{\frac{\alpha-(2-p)}{p}}\right)\right|^p}{|x|^{\gamma p}}\,dx$$

and, by Hardy's inequality,

$$\int_\Omega \frac{|u_n(t)|^{\alpha-(2-p)}}{|x|^{(\gamma+1)p}}\,dx \leq \lambda_{N,p,\gamma}^{-1}\int_\Omega \frac{\left|\nabla\left(|u_n(t)|^{\frac{\alpha-(2-p)}{p}}\right)\right|^p}{|x|^{\gamma p}}\,dx,$$

we obtain

$$\frac{1}{\alpha}\frac{d}{dt}\int_\Omega u_n^\alpha(t)\,dx + c_1\int_\Omega \frac{\left|\nabla\left(|u_n(t)|^{\frac{\alpha-(2-p)}{p}}\right)\right|^p}{|x|^{\gamma p}}\,dx \leq 0,$$

where

$$c_1 = (\alpha-1)\left(\frac{p}{\alpha-(2-p)}\right)^p - \lambda\left(\frac{p}{N-p(\gamma+1)}\right)^p > 0.$$

Therefore, by (1.7),

$$\frac{1}{\alpha}\frac{d}{dt}\int_\Omega u_n^\alpha(t)\,dx + c_1 S_{N,p,\gamma}\left[\int_\Omega \frac{|u_n(t)|^{\frac{[\alpha-(2-p)]p^*}{p}}}{|x|^{\gamma p^*}}\,dx\right]^{\frac{p}{p^*}} \leq 0. \tag{5.3}$$

Choosing

$$\alpha = \frac{N(2-p)}{p},$$

the two powers of u_n become equal. Since $\gamma \geq 0$, if we define

$$\phi(t) = \int_\Omega u_n^\alpha(t)\,dx,$$

we obtain

$$\phi'(t) + c_2[\phi(t)]^{\frac{p}{p^*}} \leq 0,$$

where $c_2 = c_2(N, p, \gamma, \Omega) > 0$, and we obtain the result for the approximate solutions u_n as in the previous theorem. The result on u follows by taking the limit on n. ∎

Remark 5.3. Note that $\eta_{N,p,\gamma} = \lambda_{N,p,\gamma}$ for $p = \frac{2N}{N+2}$.

THEOREM 5.4. *Assume that*

$$\begin{gathered} 0 < \gamma + 1 < \frac{N(2-p)}{2p}, \\ \lambda < \mu_{N,p,\gamma} = \left(\frac{N(2-p)}{p(\gamma+1)} - 1\right)\left(\frac{p(\gamma+1)}{2-p}\right)^p, \end{gathered} \tag{5.4}$$

and that there exists

$$\bar{\alpha} > \frac{(2-p)N}{p(\gamma+1)}$$

such that $\psi \in L^{\bar{\alpha}}(\Omega)$. *Then there exists a constant*

$$T^* = T^*(N, p, \gamma, \lambda, \Omega, \bar{\alpha}, \psi) \leq c_1(N, p, \gamma, \lambda, \Omega, \bar{\alpha})\|\psi\|_{L^\alpha(\Omega)}^{2-p}$$

such that any solution of problem (P) *found by approximation as in Theorem* 3.1 *satisfies*

$$u(\cdot, t) \equiv 0 \quad \text{for } t \geq T^*. \tag{5.5}$$

Proof. We use $|u_n|^{\alpha-2}u_n$ as a test function in (P_n), where α is such that

$$\frac{(2-p)N}{p(\gamma+1)} < \alpha \leq \bar{\alpha} \tag{5.6}$$

and

$$\lambda < (\alpha - 1)\left(\frac{p(\gamma+1)}{\alpha - (2-p)}\right)^p. \tag{5.7}$$

Note that this is always possible, since assumption (5.4) implies that (5.7) is true for $\alpha = \frac{(2-p)N}{p(\gamma+1)}$. As in the previous proof, we obtain inequality (5.3), where the constant c_1 is positive by (5.7). Now observe that condition (5.6) implies

$$\alpha > \frac{N(2-p)}{p}$$

and

$$\frac{\gamma\alpha pp^*}{p^*[\alpha-(2-p)]-\alpha p} > -N;$$

therefore, by Hölder's inequality,

$$\int_\Omega u_n^\alpha(t)\,dx \le \left[\int_\Omega |x|^{\frac{\gamma\alpha pp^*}{p^*[\alpha-(2-p)]-\alpha p}}\,dx\right]^{\frac{p^*[\alpha-(2-p)]-\alpha p}{p^*[\alpha-(2-p)]}} \left[\int_\Omega \frac{|u_n(t)|^{\frac{[\alpha-(2-p)]p^*}{p}}}{|x|^{\gamma p^*}}\,dx\right]^{\frac{\alpha p}{p^*[\alpha-(2-p)]}}$$
$$\le c_2(N,p,\gamma,\alpha,\Omega)\left[\int_\Omega \frac{|u_n(t)|^{\frac{[\alpha-(2-p)]p^*}{p}}}{|x|^{\gamma p^*}}\,dx\right]^{\frac{\alpha p}{p^*[\alpha-(2-p)]}}.$$

Hence one has

$$\frac{d}{dt}\int_\Omega u_n^\alpha(t)\,dx + c_3\left[\int_\Omega u_n^\alpha(t)\,dx\right]^{\frac{\alpha-(2-p)}{\alpha}} \le 0,$$

with $c_3 > 0$. Since $\frac{\alpha-(2-p)}{\alpha} < 1$, we conclude as before. □

Remark 5.5. Note that condition $0 < \gamma + 1 < \frac{N(2-p)}{2p}$ in Theorem 5.4 means that $1 < p < \frac{2N}{N+2(\gamma+1)}$, which implies, for $\gamma \ge 0$, that p also satisfies $1 < p < \frac{2N}{N+2}$. Therefore, we can compare the results of Theorems 5.2 and 5.4 in the region where $1 < p < \frac{2N}{N+2}$ and $\gamma \ge 0$. An easy calculation shows that in that region we have $\eta_{N,p,\gamma} < \mu_{N,p,\gamma}$, where $\eta_{N,p,\gamma}$ and $\mu_{N,p,\gamma}$ are given in the statements of Theorems 5.2 and 5.4, respectively. Since $\frac{N(2-p)}{p} > \frac{N(2-p)}{p(\gamma+1)}$, Theorem 5.4 gives a better result than Theorem 5.2 in the above region. Let us also point out that the value $\mu_{N,p,\gamma}$ is the same value we find in Lemma 3.2, which gives the esistence of self-similar solutions of the equation in problem (P).

5.2. Nonextinction results. If $p > 2$ and ψ verifies the hypothesis (H2), by using the Barenblatt-type functions one can easily prove that there is no extinction in finite time. Indeed, for any fixed time $T > 0$, consider the function $B(x, t+1)$, where B is the function defined in (4.2). One can easily check that, if the constant M in (4.2) is sufficiently small, then this function is a subsolution of problem (P). Since T is arbitrary, the result follows.

In this section we will prove that solutions to problem (P) with $1 < p < 2$, $\gamma + 1 \ge 0$, and $\lambda > \lambda_{n,p,\gamma}$ are nonzero for all time. The key of the proof is the construction of a nonnegative subsolution to the problem

$$\begin{cases} u_t - \Delta_{p,\gamma}(u) = \lambda\dfrac{|u|^{p-2}u}{|x|^{(\gamma+1)p}}, & (x,t) \in \Omega\times(0,T),\\ u(x,t) = 0, & (x,t)\in\partial\Omega\times(0,T),\\ u(x,0) = 0, & x\in\Omega, \end{cases} \tag{5.8}$$

following the ideas in [18] (see also [19]). Consider the eigenvalue problem

$$\begin{cases} -\Delta_{p,\gamma}(\phi_1) = \mu_1(n)W_n(x)\phi_1^{p-1}, & x\in\Omega,\\ u(x) = 0, & x\in\partial\Omega, \end{cases} \tag{5.9}$$

where $W_n(x) = \min\{n, |x|^{-(\gamma+1)p}\}$. The principal eigenvalue is isolated and simple. Moreover, it is easy to check that the sequence of principal eigenvalues, $\{\mu_1(n)\}$, is decreasing, that $\lim_{n\to\infty} \mu_1(n) = \lambda_{n,p,\gamma}$, and that the corresponding eigenfunction ϕ_1 has constant sign (see, for instance, [16]). In this way, if $\lambda > \lambda_{n,p,\gamma}$, there exists n_0 such that for $n > n_0$, one has $\lambda > \mu_1(n)$. Hence, for $n > n_0$, let $\Theta(t)$ be the positive solution to the problem $\Theta'(t) = \Theta^{p-1}(t)$, $\Theta(0) = 0$.

Define

$$v(x,t) = \Theta(\varepsilon t)\phi_1(x),$$

where $\varepsilon > 0$ will be chosen later, and ϕ_1 is a positive eigenfunction of (5.9) such that $\|\phi_1\|_\infty = 1$. We have that

$$\frac{v_t - \Delta_{p,\gamma}(v)}{\lambda v(x,t)^{p-1}} < \frac{\varepsilon \phi_1^{2-p}}{\lambda} + \frac{\mu_1(n)}{\lambda} W_n(x);$$

hence, as $2 - p > 0$, $\gamma + 1 \geq 0$, and $\frac{\mu_1(n)}{\lambda} < 1$, for a suitable $\varepsilon > 0$ we obtain that

$$\frac{v_t - \Delta_{p,\gamma}(v)}{\lambda v(x,t)^{p-1}} < W_n(x).$$

Then $v(x,t)$ is a subsolution to the truncated problem obtained from (5.8) and therefore to problem (5.8) with $1 < p < 2$, $\psi(x) \geq 0$, $(1+\gamma) > 0$, and $\lambda > \lambda_{n,p,\gamma}$. For the truncated equation we obtain a flat supersolution by solving the ordinary differential equation $y'(t) = n\lambda[y(t)]^{p-1}$, $1 < p < 2$, with data $y(0) = a$, whose solution is $y(t) = [a^{2-p} + n\lambda(2-p)t]^{1/(2-p)}$. Given a $T > 0$ we find a value of a for which $v(x,t) < y(t)$ in $\Omega \times (0,T)$ and $y(0) \geq \psi(x)$. Iterating from v, we obtain as a conclusion that in these hypotheses the minimal solution to the truncated equation of (5.8) has no finite time extinction. And therefore the same result holds for (5.8).

Remark 5.6. If $1 + \gamma < 0$, the weights are flat at the origin. If we use the eigenvalue analysis as in [16], i.e., for $\beta_n = (1+\gamma) - \frac{1}{n}$, then we define, for instance,

$$\alpha_n(x) = \begin{cases} |x|^{-p\beta_n} & \text{if } x \in \Omega \cap B_1(0), \\ |x|^{-p(\gamma+1)} & \text{if } x \in \Omega \setminus B_1(0). \end{cases}$$

In this way $\alpha_n(x) \leq |x|^{-p(\gamma+1)}$ for all $x \in \Omega$, and moreover, the eigenvalue problems

$$\begin{cases} -\operatorname{div}\left(\dfrac{|\nabla\psi_1|^{p-2}\nabla\psi_1}{|x|^{\gamma p}}\right) = \nu_1(n)\alpha_n(x)\psi_1^{p-1}, & x \in \Omega, \\ u(x) = 0, & x \in \partial\Omega, \end{cases} \tag{5.10}$$

verify the following:

1. The principal eigenvalue is isolated and simple.
2. We can choose the corresponding eigenfunction ψ_1 positive.
3. The sequence of principal eigenvalues satisfies $\nu_1(n) \searrow \lambda_{N,p,\gamma}$ as $n \to \infty$.

However, the final construction does not work.

REFERENCES

[1] B. Abdellaoui and I. Peral, *Harnack inequality for degenerate parabolic equations related to Caffarelli-Kohn-Nirenberg inequalities*, Nonlinear Anal., 57 (2004), pp. 971–1003.

[2] B. Abdellaoui and I. Peral, *Existence and nonexistence results for quasilinear parabolic equations related to Caffarelli-Kohn-Nirenberg inequalities*, to appear.
[3] B. Abdellaoui, E. Colorado, and I. Peral, *Existence and nonexistence results for a class of linear and semilinear parabolic equations related to some Caffarelli-Kohn-Nirenberg inequalities*, J. Eur. Math. Soc. (JEMS), 6 (2004), pp. 119–148.
[4] Adimurthi, N. Chaudhuri, and M. Ramaswamy, *An improved Hardy-Sobolev inequality and its application*, Proc. Amer. Math. Soc., 130 (2002), pp. 489–505.
[5] J. A. Aguilar Crespo and I. Peral, *Global behavior of the Cauchy problem for some critical nonlinear parabolic equations*, SIAM J. Math. Anal., 31 (2000), pp. 1270–1294.
[6] A. Ambrosetti, J. García Azorero, and I. Peral, *Multiplicity results for some nonlinear elliptic equations*, J. Funct. Anal., 137 (1996), pp. 219–242.
[7] P. Baras and J. Goldstein, *The heat equation with a singular potential*, Trans. Amer. Math. Soc., 294 (1984), pp. 121–139.
[8] P. Bénilan, L. Boccardo, T. Gallouët, R. Gariepy, M. Pierre, and J. L. Vázquez, *An L^1 theory of existence and uniqueness of solutions of nonlinear elliptic equations*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 22 (1995), pp. 240–273.
[9] D. Blanchard and F. Murat, *Renormalised solutions nonlinear parabolic problems with L^1 data: Existence and uniqueness*, Proc. Roy. Soc. Edinburgh Sect. A, 127 (1997), pp. 1137–1152.
[10] L. Boccardo and F. Murat, *Almost everywhere convergence of the gradients of solutions to elliptic and parabolic equations*, Nonlinear Anal., 19 (1992), pp. 581–597.
[11] H. Brezis and J.L. Vazquez, *Blow-up solutions of some nonlinear elliptic equations*, Rev. Mat. Univ. Complut. Madrid, 10 (1997), pp. 443–469.
[12] X. Cabré and Y. Martel, *Existence versus explosion instantanée pour des équations de la chaleur linéaires avec potentiel singulier*, C. R. Acad. Sci. Paris Sér. I Math., 329 (1999), pp. 973–978.
[13] L. Caffarelli, R. Kohn, and L. Nirenberg, *First order interpolation inequalities with weights*, Compositio. Math., 53 (1984), pp. 259–275.
[14] F. Catrina and Z. Q. Wang, *On the Caffarelli-Kohn-Nirenberg inequalities: Sharp constants, existence (and nonexistence), and symmetry of extremal functions*, Comm. Pure Appl. Math., 54 (2001), pp. 229–258.
[15] F. M. Chiarenza and R. P. Serapioni, *A Harnack inequality for degenerate parabolic equations*, Comm. Partial Differential Equations, 9 (1984), pp. 719–749.
[16] E. Colorado and I. Peral, *Eigenvalues and bifurcation for elliptic equations with mixed Dirichlet-Neumann boundary conditions related to Caffarelli-Kohn-Nirenberg inequalities*, Topol. Methods Nonlinear Anal., to appear.
[17] E. DiBenedetto, *Degenerate Parabolic Equations*, Springer-Verlag, Berlin, 1993.
[18] H. Fujita, *On some nonexistence and nonuniqueness theorems for nonlinear parabolic equations*, in Proc. Sympos. Pure Math. 18, AMS, Providence, RI, 1968, pp. 138–161.
[19] J. García Azorero and I. Peral Alonso, *Hardy inequalities and some critical elliptic and parabolic problems*, J. Differential Equations, 144 (1998), pp. 441–476.
[20] J. Heinonen, T. Kilpelainen, and O. Martio, *Nonlinear Potential Theory of Degenerate Elliptic Equations*, Clarendon Press, Oxford, UK, 1993.
[21] J.-L. Lions, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod, Gauthier-Villars, Paris, 1969.
[22] A. Prignet, *Existence and uniqueness of entropy solution of parabolic problems with L^1 data*, Nonlinear Anal., 28 (1997), pp. 1943–1954.
[23] S. Segura de Leon, *Existence and uniqueness for L^1 data of some elliptic equations with natural growth*, Adv. Differential Equations, 8 (2003), pp. 1377–1408.
[24] J. Simon, *Compact sets in the space $L^p(0,T;B)$*, Ann. Mat. Pura Appl. (4), 146 (1987), pp. 65–96.
[25] G. Stampacchia, *Equations elliptiques du second ordre à coefficients discontinus*, Séminaire de Mathématiques Supérieures 16, Les Presses de l'Université de Montréal, Montréal, Quebec, Canada, 1966.
[26] J. L. Vázquez and E. Zuazua, *The Hardy inequality and the asymptotic behavior of the heat equation with an inverse-square potential*, J. Funct. Anal., 173 (2000), pp. 103–153.

SIAM J. MATH. ANAL.
Vol. 36, No. 3, pp. 717–739

INVARIANCE AND NONINVARIANCE OF CENTER MANIFOLDS OF TIME-t MAPS WITH RESPECT TO THE SEMIFLOW*

TIBOR KRISZTIN†

Dedicated to Professors István Győri and László Hatvani for their 60*th birthdays*

Abstract. This paper contains two examples related to the problem of whether the local center manifolds of the time-t maps of a semiflow at an equilibrium point are also invariant under the semiflow itself. An ordinary differential equation in $\mathbb{R}^2$ is given to show that, for almost all choices of the localization functions, the center manifold of the time-1 map at the origin is not locally invariant under the flow. The second example is an abstract functional differential equation. Although a variation-of-constants formula is not known to exist in the phase space, we prove that the classical approach works: The semiflow can be modified outside a neighborhood of the equilibrium point so that the center manifold of the time-t map of the modified semiflow will be locally invariant under the original semiflow.

Key words. invariant manifold, time-t map, smoothness, modification of the nonlinearity, Lyapunov–Perron method, abstract functional differential equation

AMS subject classifications. 34C30, 37L10, 34G20, 34K19, 34K30

DOI. 10.1137/S0036141003419170

1. Introduction.

Center manifolds play a fundamental role in the study of dynamical systems near nonhyperbolic equilibrium points. They were discovered by Pliss [25] and Kelley [18] in the 1960s and later developed by many others (e.g., [4, 16, 27, 31]). There are essentially two methods for proving the existence of center manifolds. The Lyapunov–Perron approach obtains the manifold as a fixed point of a certain integral equation for flows, and of a corresponding equation with sums for maps (e.g., [5, 8, 30, 31]). The Hadamard approach is more geometrical; it uses the graph transform technique (see, e.g., [1, 16]). There is a very extensive literature on applications and various generalizations of the theory of center manifolds; see [1, 4, 6, 7, 8, 9, 13, 19, 20, 32] and the references therein. For some interesting properties of center manifolds we refer to [3, 4, 27].

In this paper we consider the time-t, $t > 0$, map of a smooth semiflow on a Banach space with an equilibrium point at zero and study the problem of whether the local center manifolds (i.e., local center-stable, center-unstable, and center manifolds) of the time-t map are invariant with respect to the semiflow. If the answer for a given semiflow were affirmative, then the local center manifolds of the time-t map would also be local center manifolds of the semiflow. In particular, the problem is very important for semiflows, for which smooth local center manifolds are not known to exist because of certain technical difficulties, while for their time-t maps smooth local center manifolds have been constructed (see, e.g., [13]). We note that at a hyperbolic equilibrium point, under natural conditions it is straightforward to show that the stable and unstable manifolds of the time-t maps are invariant with respect to the semiflow.

*Received by the editors January 31, 2003; accepted for publication (in revised form) February 13, 2004; published electronically September 24, 2004. This work was done while the author was visiting the University of Giessen and was supported by a Humboldt Fellowship. This research was also partially supported by the Hungarian Foundation for Scientific Research, grant T 034336.
http://www.siam.org/journals/sima/36-3/41917.html

†Bolyai Institute, University of Szeged, Aradi vértanúk tere 1, H-6720 Szeged, Hungary (krisztin@math.u-szeged.hu).

In section 3, for an ordinary differential equation in $\mathbb{R}^2$, we construct a center manifold of the time-1 map so that it is not locally invariant with respect to the flow. In the example, the noninvariance holds for an open and dense subset of the localization functions. It is expected that the general case is analogous, although the calculations are specific to the example.

There is a classical approach for solving the above invariance problem (see, e.g., [5]): Assume that the semiflow can be extended from a small neighborhood of the equilibrium point to a global semiflow, which is a small and globally Lipschitzian perturbation of the linearized semiflow with a small Lipschitz constant. Then the center manifold (if it exists) of the time-t map of the modified semiflow will be locally positively invariant with respect to the original semiflow. In section 4 we show that the above extendability property holds for the semiflows generated by a class of abstract functional differential equations, for which the existence of C^k-smooth local center-stable, center, and center-unstable manifolds was a problem for a long time since a variation-of-constants formula in the phase space was not known. Namely, we consider the semilinear functional differential equation

$$\dot{u}(t) = A_T u(t) + Bu_t + F(u_t) \tag{1.1}$$

in a Banach space X, where $r \geq 0$, $C = C([-r, 0]; X)$ is the Banach space of continuous mappings from $[-r, 0]$ into X with the supremum norm, $u_t \in C$ is defined by $u_t(\theta) = u(t + \theta)$ for $\theta \in [-r, 0]$, $B : C \to X$ is a bounded linear operator, $A_T : D(A_T) \subset X \to X$ is the infinitesimal generator of a compact C_0-semigroup of linear operators on X, and F is C^k-smooth from an open neighborhood of 0 in C into X with $F(0) = 0$, $DF(0) = 0$. The mild solutions of (1.1) generate a local semiflow Ψ on C. It is shown by Faria, Huang, and Wu [13] that, for $\tau > r$, the time-τ map $\Psi(\tau, \cdot)$ has C^k-smooth local center-stable, center-unstable, and center manifolds at 0. Therefore, (1.1) is a new example for the general method of [5] to construct smooth local center manifolds for semiflows.

Reaction-diffusion equations with time delay are examples for (1.1). These equations have served as models for many ecology, chemistry, and biology problems [32]. A simple yet important equation is Fisher's equation with a delayed logistic type of reaction term

$$\frac{\partial U(t,x)}{\partial t} = \frac{\partial^2 U(t,x)}{\partial x^2} + lU(t,x)[1 - U(t-r,x)], \quad t > 0,\ x \in (0,\pi), \tag{1.2}$$

with the Dirichlet boundary condition

$$U(t,0) = U(t,\pi) = 0, \tag{1.3}$$

where $l > 0$ and $r > 0$. It was shown for $l \leq 1$ that the origin is a global attractor of all positive solutions of (1.2), (1.3). As l exceeds 1, the origin loses its stability, and a unique positive equilibrium $U_l(x)$ bifurcates from the origin. By [17], the spatially nonconstant $U_l(x)$ is locally stable provided $rl \max\{U_l(x) : x \in [0,\pi]\} < \pi/2$. In addition, for $l > 1$ and $l - 1$ small, as the delay increases and crosses some critical values, Hopf bifurcations occur [2]. The stability of the nontrivial periodic solutions on the center manifold arising from the first Hopf bifurcation point was stated without a proof in [2]. Recently, Faria and Huang [12] completed the proof by applying normal form theory from [11, 13] and C^k-smoothness of center manifolds from [13]. However, the C^k-smooth center manifold of a time-τ map as obtained in [13] is not necessarily invariant with respect to the semiflow (see section 3 in this paper), and thus it is not

necessarily a C^k-smooth center manifold of the semiflow. Our result in this paper provides the C^k-smoothness of the center manifold. This fact was used without proof in the papers [11, 12]. We refer to [12, 32] to see how problem (1.2), (1.3) can be written in the abstract form (1.1) with $X = L^2([0,\pi],\mathbb{R})$ after time-scaling and translating the equilibrium U_l to the origin. Of course, the same approach works for more general reaction-diffusion equations and different bifurcation problems.

We conclude this introduction by listing some notation. $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{R}$, and $\mathbb{C}$ denote the set of nonnegative integers, integers, real, and complex numbers, respectively.

Spectra of linear operators in a Banach space E over $\mathbb{R}$ are defined as spectra of complexifications. If a decomposition $E = E_s \oplus E_c \oplus E_u$ into closed linear subspaces is given, then $E_{sc} = E_s \oplus E_c$, $E_{cu} = E_c \oplus E_u$, $E_{su} = E_s \oplus E_u$, and Pr_{E_s} denotes the associated projection along E_{cu} onto E_s, and similarly for $\mathrm{Pr}_{E_{sc}}$, Pr_{E_c}, Pr_{E_u}, $\mathrm{Pr}_{E_{su}}$, and $\mathrm{Pr}_{E_{cu}}$.

By a (local) semiflow Φ on E we mean a continuous semiflow defined on an open subset of $[0,\infty) \times E$. If the domain of Φ is $[0,\infty) \times E$, then Φ is called a global semiflow on E. The semiflow Φ is said to be C^k-smooth on E if, for each $t \geq 0$, $\Phi(t,\cdot)$ is C^k-smooth from its domain into E. If $I \subset \mathbb{R}$ is an interval and $y : I \to E$ is a curve with $y(t+s) = \Phi(t, y(s))$ for all $s \in I$ and for all $t \geq 0$ with $t + s \in I$, then y is called a trajectory of Φ. If $x \in E$, $I \subset (-\infty, 0]$ is an interval with $0 \in I$, and $y : I \to E$ is a trajectory of Φ with $y(0) = x$, then y is called a backward trajectory of Φ through x.

Let $V \subset U \subset E$, and let Φ be a semiflow on E. We say that V is positively invariant with respect to the semiflow Φ relative to U if for each $x \in V$ and for every $t > 0$,

$$\{\Phi(s,x) : s \in [0,t]\} \subset U \quad \text{implies} \quad \{\Phi(s,x) : s \in [0,t]\} \subset V.$$

We say that V is negatively invariant with respect to Φ relative to U if for each $x \in V$ the following holds: If a backward trajectory through x exists, then there exist a $t_x > 0$ and a backward trajectory $y : (-t_x, 0] \to E$ through x, with t_x maximal, so that for every $t \in (0, t_x)$,

$$\{y(s) : s \in [-t,0]\} \subset U \quad \text{implies} \quad \{y(s) : s \in [-t,0]\} \subset V.$$

V is invariant relative to U if it is both positively and negatively invariant relative to U.

The set $V \subset E$ is called locally positively invariant with respect to the semiflow Φ if for each $x \in V$ there exists $t_x > 0$ so that $\Phi(s,x) \in V$ for all $s \in [0, t_x]$. It is easy to see that if $V \subset U \subset E$, U is open, $\{0\} \times U$ belongs to the domain of Φ, and V is positively invariant with respect to Φ relative to U, then V is also locally positively invariant with respect to Φ.

For $\eta > 0$ and a Banach space E with norm $|\cdot|$, let E_η denote the Banach space of all sequences $\chi = (x_n)_0^\infty \in E^{\mathbb{N}}$ with $\sup_{j\in\mathbb{N}} |x_j|\eta^{-j} < \infty$ and norm $||\chi||_{E_\eta} = \sup_{j\in\mathbb{N}} |x_j|\eta^{-j}$.

2. Invariant manifolds for maps. In this section we recall some steps of the construction of invariant manifolds of maps from [13, 19] with a slight modification. It is shown in section 3 that if the map is the time-1 map of a flow, then the construction of [13, 19] may lead to center manifolds of the map which are not locally invariant with respect to the flow. A modified construction is used in section 4 to get smooth invariant manifolds for a semiflow generated by an abstract functional differential equation. This application motivates some of the particular assumptions on the map below.

Let $f : V \to E$ be a C^k-smooth map, $k \in \mathbb{N} \setminus \{0\}$, on an open subset V of a Banach space E over $\mathbb{R}$, with $0 \in V$ and $f(0) = 0$. Let $L = Df(0)$ and assume that E has the decomposition

$$E = E_s \oplus E_c \oplus E_u$$

such that E_s, E_c, and E_u are closed subspaces of E, E_c and E_u are finite-dimensional, $E_c \neq \{0\}$, $L(E_s) \subset E_s$, $L(E_c) \subset E_c$, $L(E_u) \subset E_u$, and the spectra σ_s, σ_c, and σ_u of the induced maps $L_s : E_s \ni x \mapsto Lx \in E_s$, $L_c : E_c \ni x \mapsto Lx \in E_c$, and $L_u : E_u \ni x \mapsto Lx \in E_u$ are contained in compact subsets of $\{z \in \mathbb{C} : |z| < 1\}$, $\{z \in \mathbb{C} : |z| = 1\}$, and $\{z \in \mathbb{C} : |z| > 1\}$, respectively.

Recent works of Faria, Huang, and Wu [13] and Krisztin, Walther, and Wu [19] show that under the above conditions on f, C^k-smooth local center-stable, center-unstable, and center manifolds of the map f can be obtained by using a discrete version of the Lyapunov–Perron method, and by following the approach of Vanderbauwhede and van Gils [31] for flows, and Diekmann et al. [9] for semiflows. The construction of these manifolds starts with an extension and modification of f. As $E_c \oplus E_u$ is finite-dimensional, by applying a suitable cut-off function, it is possible to modify and extend $f - L$ outside a small neighborhood of 0 in E to get a small Lipschitzian $r_\delta : E \to E$ with a small Lipschitz constant such that r_δ is C^k-smooth on small strips containing the center-unstable space $E_c \oplus E_u$. The global center-stable, center-unstable, and center manifolds of the map $g_\delta := L + r_\delta$ are defined as initial points of forward, backward, and global trajectories, respectively, of g_δ having a small exponential growth bound. These global manifolds are graphs of Lipschitz continuous maps from the spaces $E_s \oplus E_c$, $E_c \oplus E_u$, and E_c into their complementary spaces, respectively. Since the global center-unstable and center manifolds are situated in regions where g_δ is C^k-smooth, their C^k-smoothness can be also verified. This is not surprising since analogous results for flows and semiflows were known. The significance of [13, 19] is that the same modification works also for the infinite-dimensional center-stable manifold, although in a different way: Only that part of the global center-stable manifold is C^k-smooth which is in a small strip containing $E_c \oplus E_u$. This is sufficient because the local center-stable, center-unstable, and center manifolds of the map f are obtained as intersections of the corresponding global manifolds of g_δ with suitable small neighborhoods of 0 in E.

We remark that the finite dimensionality of E_c and E_u is made to guarantee the existence of smooth cut-off functions. This restriction is not necessary if E is a Hilbert space, or if it is a Banach space with the "C^k extension property" (see, e.g., [26]). In the example of section 4, $\dim E_c < \infty$ and $\dim E_u < \infty$ hold.

Set $a = \sup_{\lambda \in \sigma_s} |\lambda|$ and $b = \inf_{\lambda \in \sigma_u} |\lambda|$. In case $E_s = \{0\}$ let $a = 1/2$, and in case $E_u = \{0\}$ let $b = 3/2$. Fix an $\epsilon > 0$ with $a + \epsilon < 1$ and $(1 + \epsilon)^k < b - \epsilon$. There exists a norm $|\cdot|$ on E which is equivalent to the one given originally and satisfies $|x| = |\operatorname{Pr}_{E_s} x| + |\operatorname{Pr}_{E_c} x| + |\operatorname{Pr}_{E_u} x|$ and

$$\begin{aligned} |L \operatorname{Pr}_{E_s} x| &\leq (a + \epsilon) |\operatorname{Pr}_{E_s} x|, \\ |L \operatorname{Pr}_{E_c} x| &\leq (1 + \epsilon) |\operatorname{Pr}_{E_c} x|, \\ |L \operatorname{Pr}_{E_u} x| &\geq (b - \epsilon) |\operatorname{Pr}_{E_u} x| \end{aligned} \tag{2.1}$$

for all $x \in E$.

Define $r^* : V \ni x \mapsto f(x) - Lx \in E$, and extend r^* to a map $r : E \to E$ by $r(x) = 0$ for all $x \in E \setminus V$. Let $g = L + r$.

For $q > 0$ let $\mathcal{R}_q$ denote the set of C^∞-smooth functions $\rho : \mathbb{R} \to \mathbb{R}$ satisfying $\rho(t) = 1$ for $t \le 1$, $0 < \rho(t) < 1$ for $1 < t < 2$, $\rho(t) = 0$ for $t \ge 2$, and $|\rho'(t)| < q$ for all $t \ge 0$. Clearly, for a sufficiently large $q > 0$, $\mathcal{R}_q \neq \emptyset$. We fix such a $q > 0$, and set $\mathcal{R} = \mathcal{R}_q$. Let $\mathcal{R}$ be equipped with the metric d defined by $d(\rho, \zeta) = \sup_{t\in\mathbb{R}} |\rho(t)-\zeta(t)|$, $\rho, \zeta \in \mathcal{R}$.

Fix a norm $|\cdot|_{cu}$ on the finite-dimensional E_{cu}, which is C^∞-smooth on $E_{cu} \setminus \{0\}$. Then $||\cdot|| : E \ni x \mapsto \max\{|\operatorname{Pr}_{E_s} x|, |\operatorname{Pr}_{E_{cu}} x|_{cu}\} \in \mathbb{R}$ is a new norm on E which is equivalent to $|\cdot|$. Let $E(\delta) = \{x \in E : ||x|| < \delta\}$ for $\delta > 0$.

In section 3 we shall consider a set of modifications of the nonlinearity r. This motivates the introduction of a parametrized family of modifications $r_{\delta p}$, given below.

Let a $\delta^* > 0$, a set $\mathcal{P}$, and, for each $\delta \in (0, \delta^*]$ and $p \in \mathcal{P}$, a mapping $r_{\delta p} : E \to E$ be given. Assume that the family of mappings $r_{\delta p}$, $\delta \in (0, \delta^*]$, $p \in \mathcal{P}$, satisfies the following two hypotheses:

(R1) $E(\delta^*) \subset V$; $r_{\delta p}|_{E(\delta)} = r|_{E(\delta)}$ for all $\delta \in (0, \delta^*]$ and $p \in \mathcal{P}$; there exists a nondecreasing function $\lambda : [0, \delta^*] \to [0, 1]$ with $\lim_{\delta\to 0^+} \lambda(\delta) = 0 = \lambda(0)$ such that for every $\delta \in (0, \delta^*]$, for every $p \in \mathcal{P}$, and for all x, y in E,

$$|r_{\delta p}(x)| \le \delta\lambda(\delta),$$
$$|r_{\delta p}(x) - r_{\delta p}(y)| \le \lambda(\delta)|x - y|.$$

(R2) For every $\delta \in (0, \delta^*]$ and $p \in \mathcal{P}$ the restriction $r_{\delta p}|_{\{x\in E:|\operatorname{Pr}_{E_s} x|<\delta\}}$ is C^k-smooth, and all lth derivatives, $l \in \{1, \dots, k\}$, of $r_{\delta p}|_{\{x\in E:|\operatorname{Pr}_{E_s} x|<\delta\}}$ are bounded.

For every $\delta > 0$ and $\rho \in \mathcal{R}$, define $\tilde{r}_{\delta\rho} : E \to E$ by

$$\tilde{r}_{\delta\rho}(x) = r(x)\rho\left(\frac{|\operatorname{Pr}_{E_{cu}} x|_{cu}}{\delta}\right)\rho\left(\frac{|\operatorname{Pr}_{E_s} x|}{\delta}\right),$$

and set $\tilde{g}_{\delta\rho} = L + \tilde{r}_{\delta\rho}$. One can fix a $\delta_0 > 0$ so that $\overline{E(3\delta_0)} \subset V$ and that $r|_{E(3\delta_0)}$ is C^k-smooth, and all lth derivatives, $l \in \{1, \dots, k\}$, of $r|_{E(3\delta_0)}$ are bounded. If $\delta > 0$, $\rho \in \mathcal{R}$, and $x \in E$ with $|\operatorname{Pr}_{E_s} x| < \delta$, then

$$\tilde{r}_{\delta\rho}(x) = r(x)\rho\left(\frac{|\operatorname{Pr}_{E_{cu}} x|_{cu}}{\delta}\right).$$

It follows that, for every $\delta \in (0, \delta_0)$ and $\rho \in \mathcal{R}$, the map $\tilde{r}_{\delta\rho}|_{\{x\in E:|\operatorname{Pr}_{E_s} x|<\delta\}}$ is C^k-smooth, and all lth derivatives, $l \in \{1, \dots, k\}$, of $\tilde{r}_\delta|_{\{x\in E:|\operatorname{Pr}_{E_s} x|<\delta\}}$ are bounded. By an argument completely analogous to [19, Proposition II.2] there exist $\delta_1 \in (0, \delta_0)$ and a nondecreasing function $\tilde{\lambda} : [0, \delta_1] \to [0, 1]$ with $\lim_{\delta\to 0^+} \tilde{\lambda}(\delta) = 0 = \tilde{\lambda}(0)$ so that for each $\delta \in (0, \delta_1]$ and $\rho \in \mathcal{R}$, and for all x, y in E,

$$|\tilde{r}_{\delta\rho}(x)| \le \delta\tilde{\lambda}(\delta), \qquad |\tilde{r}_{\delta\rho}(x) - \tilde{r}_{\delta\rho}(y)| \le \tilde{\lambda}(\delta)|x - y|. \tag{2.2}$$

The construction of $\tilde{\lambda}$ (see [19, Proposition II.2]) uses the facts $r(0) = 0$, $Dr(0) = 0$ and that ρ' is bounded on $[0, \infty)$. Since $\sup_{t\ge 0} |\rho'(t)| < q$ for all $\rho \in \mathcal{R}$, a function $\tilde{\lambda}$ can be constructed so that (2.2) is satisfied for all $\rho \in \mathcal{R}$.

The choice $\delta^* = \delta_1$, $\mathcal{P} = \mathcal{R}$ and $r_{\delta\rho} = \tilde{r}_{\delta\rho}$, $\delta \in (0, \delta_1]$, $\rho \in \mathcal{R}$, clearly satisfies hypotheses (R1) and (R2). This case is used in section 3. In section 4, the case $\mathcal{P} = \{\rho\}$ with a fixed $\rho \in \mathcal{R}$ is applied. However, there the mappings $r_{\delta\rho}$ are necessarily different from $\tilde{r}_{\delta\rho}$ in order to guarantee that the local center manifolds of the time-t maps of the semiflow generated by (1.1) are invariant with respect to the semiflow.

Choose the reals η_0 and η_1 so that $1+\epsilon<\eta_0<\eta_0^k<\eta_1<b-\epsilon$, and define

$$c=\max\left\{\frac{1}{s-1-\epsilon}+\frac{1}{b-\epsilon-s}:s\in[\eta_0,\eta_1]\right\}.$$

Assuming that a family of mappings $r_{\delta p}$, $\delta\in(0,\delta^*]$, $p\in\mathcal{P}$, is given with properties (R1) and (R2), we choose $\delta^{**}\in(0,\delta^*)$ so small that $\lambda(\delta^{**})<1/(2c)$ and $\lambda(\delta^{**})<(1-a-\epsilon)^2/2$.

In the following result we state a slightly modified version of Theorem 5.1 in [13].

THEOREM 2.1. *Let E, f, L, and r be defined as above. Let a $\delta^*>0$, a set $\mathcal{P}$, and, for each $\delta\in(0,\delta^*]$ and $p\in\mathcal{P}$, a mapping $r_{\delta p}:E\to E$ be given such that hypotheses* (R1) *and* (R2) *are satisfied. Set $g_{\delta p}=L+r_{\delta p}$ for $\delta\in(0,\delta^*)$ and $p\in\mathcal{P}$.*

Then for all $\delta\in(0,\delta^{**})$ and $p\in\mathcal{P}$ the following holds:

(i) *There is a Lipschitz continuous map $w:E_{sc}\to E_u$ with Lipschitz constant 2 and $w(0)=0$ so that the set*

$$W^{cs}=\{z+w(z):z\in E_{sc}\}$$

is equal to the set

$$\{x\in E:\text{There is }(x_n)_0^\infty\in E_{\eta_0}\text{ with }x_0=x,\text{ and }x_{n+1}=g_{\delta p}(x_n)\text{ for all }n\in\mathbb{N}\}.$$

(ii) *There exist convex open neighborhoods N_{sc} of 0 in E_{sc}, N_u of 0 in E_u, $N\subset V$ of 0 in E such that $N_{sc}+N_u\subset E(\delta)$, $w(N_{sc})\subset N_u$, $w|_{N_{sc}}$ is C^k-smooth, $Dw(0)=0$, and the set*

$$W^{cs}_{loc}=\{z+w(z):z\in N_{sc}\}$$

satisfies $f(W^{cs}_{loc}\cap N)\subset W^{cs}_{loc}$ and $\cap_{n=0}^\infty f^{-n}(N_{sc}+N_u)\subset W^{cs}_{loc}$.

For a proof of Theorem 2.1 we can refer to the proof of Theorem 5.1 in [13] (see also [19, Theorem II.1]). Our observation is that the particular modification used in [13] can be replaced by a parametrized family of mappings. The reason for this is that all the required smallness conditions on δ are expressed by the function λ, and λ is independent of the choice of $p\in\mathcal{P}$.

Remarks. 1. The set $W^{cs}=\{z+w(z):z\in E_{sc}\}$ is called the global center-stable manifold of the map $g_{\delta p}$, while W^{cs}_{loc} is a local center-stable manifold of f. Since, for the fixed η, w depends on the constant δ and on the parameter p, the sets W^{cs} and W^{cs}_{loc} also depend on δ and p. This dependence on δ and p is related also to the nonuniqueness of local center manifolds (see [3, 4]). As the function λ in hypothesis (R1) is independent of $p\in\mathcal{P}$, for a fixed $\delta>0$, the neighborhood N_{sc} of 0 in E_{sc} can be chosen independently of $p\in\mathcal{P}$.

2. Under the same conditions as those in Theorem 2.1, analogous results can be stated for the center-unstable and center manifolds. For the center-unstable case we refer to [19, Theorem III.1]; for the center case we refer to [13]; and for works containing results for semiflows we refer to [5, 9, 30, 31].

3. The proofs of Theorem 5.1 in [13] and Theorem II.1 in [19] fix a $\rho\in\mathcal{R}$ and use the modification $r_\delta=\tilde{r}_{\delta\rho}$, $\delta>0$, of the nonlinearity r.

In the next section we need the dependence of the manifolds obtained in Theorem 2.1 on the parameters in a particular case.

PROPOSITION 2.2. *Let E, f, L, r, η_0, η_1, and c be the same as in Theorem 2.1. Assume that the family of mappings $\tilde{r}_{\delta\rho} : E \to E$, $\delta \in (0, \delta_1]$, $\rho \in \mathcal{R}$, is given as above. Define $\delta^{**} \in (0, \delta_1)$ so that $\tilde{\lambda}(\delta^{**}) < \min\{1/(2c), (1-a-\epsilon)^2/2\}$. Set $r_0 = \sup_{x \in E(2\delta^{**})} |r(x)|$.*

Then for each $\delta \in (0, \delta^{**})$ and for all $\rho, \zeta \in \mathcal{R}$, the mappings $w_{\delta\rho} : E_{sc} \to E_u$ and $w_{\delta\zeta} : E_{sc} \to E_u$, guaranteed by Theorem 2.1, satisfy

$$\sup_{z \in E_{sc}} |w_{\delta\rho}(z) - w_{\delta\zeta}(z)| \le 4cr_0\, d(\rho, \zeta).$$

The proof is omitted since it goes in a standard way.

3. Noninvariance: An example. In this section we consider an ordinary differential equation in $\mathbb{R}^2$ so that the origin is an equilibrium point with a one-dimensional center and unstable subspaces. Fixing a smooth norm $||\cdot||$ in $\mathbb{R}^2$, we modify the nonlinearity r of the time-1 map as

$$\tilde{r}_{\delta\rho} = r(x)\rho\left(\frac{||x||}{\delta}\right), \quad \delta > 0,\ \rho \in \mathcal{R}.$$

For a sufficiently small fixed $\delta > 0$, Theorem 2.1 guarantees local center-stable manifolds $W^{cs}_{loc}(\rho)$ for all $\rho \in \mathcal{R}$, which are center manifolds since the stable subspace is trivial. We show that, for an open and dense subset of $\mathcal{R}$, these local manifolds are not locally positively invariant with respect to the flow.

The idea of the proof is simple. Let $\rho \in \mathcal{R}$ be fixed. The global center-stable manifold W of the modified time-1 map $G = L + \tilde{r}_{\delta\rho}$ consists of those points $z \in \mathbb{R}^2$ for which the trajectories $(G^n(z))_{n=0}^{\infty}$ do not grow faster than $(\eta^n)_{n=0}^{\infty}$ with some fixed η between 1 and the greatest eigenvalue of the linear part L of the time-1 map. There is a smooth function $h : \mathbb{R} \to \mathbb{R}$ with $h(0) = 0$ and $h'(0) = 0$ so that $W = \{(x, h(x)) : x \in \mathbb{R}\}$. If the local center manifold $W(\gamma) = \{(x, h(x)) : |x| < \gamma\}$, with a sufficiently small $\gamma > 0$, is locally positively invariant with respect to the flow, then $W^+(\gamma) = \{(x, h(x)) : 0 < x < \gamma\}$ is a single trajectory of the flow. For another cut-off function $\rho_\kappa \in \mathcal{R}$ the same holds replacing G, h, W, $W(\gamma)$, $W^+(\gamma)$ by G_κ, h_κ, W_κ, $W_\kappa(\gamma)$, $W^+_\kappa(\gamma)$, respectively. Consequently, if both $W(\gamma)$ and $W_\kappa(\gamma)$ are locally positively invariant with respect to the flow, then either $W^+(\gamma) = W^+_\kappa(\gamma)$, or $W^+(\gamma)$ and $W^+_\kappa(\gamma)$ are disjoint sets. If ρ_κ is chosen such that ρ_κ differs from ρ only in a small interval $(c_1, c_2) \subset (1, 2)$, then $G(z) = G_\kappa(z)$ for all z from the set $\mathbb{R}^2 \setminus \{z \in \mathbb{R}^2 : c_1\delta < ||z|| < c_2\delta\}$. This fact allows us to construct a sequence $(u_n)_{n\in\mathbb{Z}}$ in the open right half plane $\mathbb{R}^2_+$ such that $u_n \in W \cap W_\kappa$, and $u_n \to 0$ as $n \to -\infty$. On the other hand, $\rho \ne \rho_\kappa$ makes it possible to find another sequence $(v_n)_{n\in\mathbb{Z}}$ in $\mathbb{R}^2_+$ with $v_n \in W_\kappa \setminus W$ for all integers $n \le 0$, and $v_n \to 0$ as $n \to -\infty$. Thus, for each $\gamma > 0$, $W^+(\gamma) \ne W^+_\kappa(\gamma)$ and $W^+(\gamma) \cap W^+_\kappa(\gamma) \ne \emptyset$. Therefore, both $W(\gamma)$ and $W_\kappa(\gamma)$ cannot be locally positively invariant with respect to the flow. It also follows that the noninvariance holds for a dense subset of $\mathcal{R}$. By using Proposition 2.2, there is an open subset of $\mathcal{R}$ so that the corresponding time-1 maps are noninvariant with respect to the flow.

It is expected that the general case is analogous, although the calculations below are specific to the example.

Now we give the details of the promised example. Consider the two-dimensional system of ordinary differential equations

$$\begin{cases} \dot{x} = x^3, \\ \dot{y} = y - x^2. \end{cases} \tag{3.1}$$

In $\mathbb{R}^2$ the solutions of (3.1) define the flow

$$\Phi(t,x,y)=\begin{pmatrix} \frac{x}{\sqrt{1-2tx^2}} \\ e^t y - e^t \int_0^t \frac{e^{-s}x^2}{1-2sx^2}ds \end{pmatrix},$$

$-\infty < t < 1/(2x^2)$. Let us fix an integer $k \geq 1$. It is clear that, for each $t \in \mathbb{R}$, $\Phi(t,\cdot)$ is C^k-smooth. Set

$$V = \left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right) \times \mathbb{R}.$$

The time-1 map $f = \Phi(1,\cdot)$ is defined and C^k-smooth on V, and it is given by

$$f(x,y)=\begin{pmatrix} \frac{x}{\sqrt{1-2x^2}} \\ ey - e\int_0^1 \frac{e^{-s}x^2}{1-2sx^2}ds \end{pmatrix} \qquad ((x,y)\in V).$$

We have

$$Df(0,0)=\begin{pmatrix} 1 & 0 \\ 0 & e \end{pmatrix},$$

and $E=\mathbb{R}^2$ has the decomposition $E = E_s \oplus E_c \oplus E_u$ with $E_s = \{(0,0)\}$, $E_c = \mathbb{R}\times\{0\}$, and $E_u = \{0\}\times\mathbb{R}$. In this case $a = 1/2$, $b = e$. Choose $\epsilon \in (0,1/2)$ so that $(1+\epsilon)^k < e - \epsilon$, and let the norm $|\cdot|$ on $\mathbb{R}^2$ be given by

$$|(x,y)| = |(x,0)| + |(0,y)| = |x| + |y|$$

Then the trichotomy (2.1) holds for $L = Df(0,0)$.

Define $r : \mathbb{R}^2 \to \mathbb{R}^2$ by $r(x,y) = (0,0)$ for $(x,y) \in \mathbb{R}^2 \setminus V$, and

$$r(x,y)=\begin{pmatrix} \frac{x}{\sqrt{1-2x^2}} - x \\ -e\int_0^1 \frac{e^{-s}x^2}{1-2sx^2}\,ds \end{pmatrix} \quad \text{for } (x,y)\in V.$$

In order to make certain estimates simpler, we choose a particular smooth norm in $\mathbb{R}^2$.

PROPOSITION 3.1. *There exists a norm $||\cdot||$ on $\mathbb{R}^2$ which is C^∞-smooth on $\mathbb{R}^2\setminus\{(0,0)\}$ and satisfies the following:*

$$\begin{aligned}
&|x| \leq \frac{3}{4} \text{ and } |y| \leq \frac{3}{4} \quad \text{imply} \quad ||(x,y)|| < 1,\\
&|x| = 1 \text{ and } |y| \leq \frac{3}{4} \quad \text{imply} \quad ||(x,y)|| = 1,\\
&|x| \leq \frac{3}{4} \text{ and } |y| = 1 \quad \text{imply} \quad ||(x,y)|| = 1,\\
&|x| \geq 1 \text{ or } |y| \geq 1 \quad \text{implies} \quad ||(x,y)|| \geq 1.
\end{aligned} \tag{3.2}$$

Proof. Suppose that $\nu : \mathbb{R} \to \mathbb{R}$ is a $\pi/2$-periodic positive C^∞-smooth function. Let $\mathcal{C}$ be the simple closed curve in $\mathbb{R}^2$ given by $r = \nu(\phi)$, $\phi \in [0, 2\pi]$, in polar coordinates $[r,\phi]$. Then $\operatorname{int}\mathcal{C}$, the interior of $\mathcal{C}$, is symmetric about zero. Assume that $\operatorname{int}\mathcal{C}$ is convex. For $(x,y) \in \mathbb{R}^2$ set $s(x,y) = \sqrt{x^2+y^2}$, and let $\phi(x,y)$ denote the

polar angle of the point (x, y). Then, by the symmetry and convexity of $\operatorname{int}\mathcal{C}$, it is not difficult to see that the map

$$\mathbb{R}^2 \ni (x, y) \mapsto \frac{s(x, y)}{\nu(\phi(x, y))} \in \mathbb{R}$$

defines a norm $||\cdot||$ on $\mathbb{R}^2$ which is C^∞-smooth on $\mathbb{R}^2 \setminus \{(0, 0)\}$. In fact, the norm is the Minkowski functional associated to the bounded balanced convex absorbing set $\operatorname{int}\mathcal{C}$. Let $|\mathcal{C}|$ denote the trace of $\mathcal{C}$. If

$$\left\{(x, y) : |x| = 1,\ |y| \le \frac{3}{4}\right\} \bigcup \left\{(x, y) : |x| \le \frac{3}{4}, |y| = 1\right\} \subset |\mathcal{C}| \tag{3.3}$$

also holds, then $||\cdot||$ satisfies (3.2). It is an elementary exercise to find a ν with the above properties. □

For $\rho \in \mathcal{R}$ and $\delta > 0$, define $\tilde{r}_{\delta\rho} : \mathbb{R}^2 \to \mathbb{R}^2$ by

$$\tilde{r}_{\delta\rho}(x, y) = r(x, y)\rho\left(\frac{||(x, y)||}{\delta}\right).$$

Set $\tilde{g}_{\delta\rho} = Df(0, 0) + \tilde{r}_{\delta\rho}$. We have seen in section 2 that there exists $\delta_1 > 0$ so that the family of mappings $r_{\delta\rho} = \tilde{r}_{\delta\rho}$, $\delta \in (0, \delta_1]$, $\rho \in \mathcal{R}$, satisfies hypotheses (R1) and (R2) with $\delta^* = \delta_1$ and $\mathcal{P} = \mathcal{R}$. Therefore, we can fix $\eta_0 \in (1 + \epsilon, e - \epsilon)$ and $\delta^{**} \in (0, \delta^*)$ such that statements (i) and (ii) of Theorem 2.1 are satisfied for all $\delta \in (0, \delta^{**})$ and for all $\rho \in \mathcal{R}$.

Let us fix $\rho \in \mathcal{R}$ and $\delta \in (0, \delta^{**})$ so that

$$\delta < \min\left\{\frac{e - 2}{16e}, \frac{1}{\sqrt{32q}}\right\}, \tag{3.4}$$

where $q > 0$ appears in the definition of $\mathcal{R}$. Set $G = L + \tilde{r}_{\delta\rho}$. We have

$$G(x, y) = \begin{pmatrix} x + \left(\frac{x}{\sqrt{1-2x^2}} - x\right)\ \rho\left(\frac{||(x,y)||}{\delta}\right) \\ ey - e\int_0^1 \frac{e^{-s}x^2}{1-2sx^2} ds\, \rho\left(\frac{||(x,y)||}{\delta}\right) \end{pmatrix} \quad ((x, y) \in \mathbb{R}^2).$$

By Theorem 2.1, the global center-stable manifold W of G exists, and there is a map $w : \mathbb{R} \times \{0\} \to \{0\} \times \mathbb{R}$ such that $w((0, 0)) = (0, 0)$ and

$$\begin{aligned} W &= \{z + w(z) : z \in \mathbb{R} \times \{0\}\} \\ &= \{u \in \mathbb{R}^2 : \text{There is a sequence } (u_n)_0^\infty \text{ in } \mathbb{R}^2_{\eta_0} \\ &\qquad \text{with } u_0 = u, \text{ and } u_{n+1} = G(u_n) \text{ for all } n \in \mathbb{N}\}, \end{aligned}$$

$$|w(x_1, 0) - w(x_2, 0)| \le 2|(x_1, 0) - (x_2, 0)| = 2|x_1 - x_2| \quad (x_1, x_2 \in \mathbb{R}).$$

We shall use the notation

$$((x, y))_1 = x, \quad ((x, y))_2 = y$$

for $(x, y) \in \mathbb{R}^2$.

Define $h : \mathbb{R} \to \mathbb{R}$ by

$$h(x) = \big(w((x, 0))\big)_2.$$

Then $h(0) = 0$,

$$|h(x_1) - h(x_2)| \leq 2|x_1 - x_2| \quad (x_1, x_2 \in \mathbb{R}),$$

and

$$W = \{(x, h(x) : x \in \mathbb{R}\}.$$

PROPOSITION 3.2. *$h(x) = 0$ for $|x| \geq 2\delta$, and $0 < h(x) < \delta/2$ for $0 < |x| < 2\delta$.*

Proof. Assume $|x| \geq 2\delta$. Then $||(x, y)|| \geq 2\delta$ for all $y \in \mathbb{R}$, and $G^n(x, h(x)) = (x, e^n h(x))$ for all $n \in \mathbb{N}$. The fact $(x, h(x)) \in W$ implies $((x, e^n h(x)))_0^\infty \in \mathbb{R}^2_{\eta_0}$. Since $1 < \eta_0 < e$, $h(x) = 0$ follows.

Assume that $|x| < 2\delta$ and $h(x) \geq \delta/2$. By the choice of δ, we have $\delta < (e-2)/(16e)$, and thus

$$e \int_0^1 \frac{e^{-s}x^2}{1 - 2sx^2} ds \leq \frac{ex^2}{1 - 2x^2} < \frac{4e\delta^2}{1 - 8\delta^2} < 8e\delta^2 < \frac{8e(e-2)}{16e}\delta = \frac{e-2}{2}\delta.$$

Then

$$(G(x, h(x)))_2 = eh(x) - e \int_0^1 \frac{e^{-s}x^2}{1 - 2sx^2} ds\, \rho\left(\frac{||(x, h(x))||}{\delta}\right) \geq \frac{e\delta}{2} - \frac{e-2}{2}\delta = \delta.$$

By using the facts $G(x, h(x)) \in W$ and $h(u) = 0$ for $|u| \geq 2\delta$, we conclude

$$|(G(x, h(x)))_1| < 2\delta.$$

Hence

$$(G^2(x, h(x)))_2 \geq e\delta - \frac{e-2}{2}\delta = \left(\frac{e}{2} + 1\right)\delta > 2\delta$$

follows. Then $||G^2(x, h(x))|| \geq 2\delta$ holds. Consequently,

$$G^{n+2}(x, h(x)) = \begin{pmatrix} (G^2(x, h(x)))_1 \\ e^n (G^2(x, h(x)))_2 \end{pmatrix} \quad \text{for all } n \in \mathbb{N}.$$

Clearly, $(G^n(x, h(x)))_0^\infty \notin \mathbb{R}^2_{\eta_0}$ because of $\left(G^2(x, h(x))\right)_2 > 2\delta$. On the other hand, the definition of W yields $(G^n(x, h(x)))_0^\infty \in \mathbb{R}^2_{\eta_0}$, a contradiction. Therefore, $h(x) < \delta/2$.

If $y < 0$ and $x \in \mathbb{R}$, then

$$(G^n(x, y))_2 \leq e^n y \quad \text{for all } n \in \mathbb{N}.$$

Obviously, $(G^n(x, y))_0^\infty \notin \mathbb{R}^2_{\eta_0}$, and $(x, y) \notin W$.

If $0 < |x| < 2\delta$ and $y = 0$, then $||(x, y)|| = |x| < 2\delta$. Then

$$\left(G(x, 0)\right)_2 = -e \int_0^1 \frac{e^{-s}x^2}{1 - 2x^2} ds\, \rho\left(\frac{|x|}{\delta}\right) < 0.$$

Hence $(x, 0) \notin W$ follows as above. Consequently, $h(x) > 0$ for $0 < |x| < 2\delta$. ∎

Define

$$U = (0, 2\delta) \times \left(0, \frac{\delta}{2}\right).$$

If $(x,y) \in U$ and $x \le (3/4)\delta$, then $||(x,y)|| < \delta$, and $\rho(||(x,y)||/\delta) = 1 = \rho(x/\delta)$. If $(x,y) \in U$ and $x > (3/4)\delta$, then $|x| > (3/2)|y|$, and thus $||(x,y)|| = |x| = x$. Consequently,

$$\rho\left(\frac{||(x,y)||}{\delta}\right) = \rho\left(\frac{x}{\delta}\right) \quad \text{for all } (x,y) \in U.$$

Therefore, the first component of G in U is given by the function

$$a : [0, 2\delta] \ni x \mapsto x + \left(\frac{x}{\sqrt{1-2x^2}} - x\right)\rho\left(\frac{x}{\delta}\right) \in \mathbb{R}.$$

The function a is C^∞-smooth, $a(0) = 0$, $a(2\delta) = 2\delta$. Moreover, it has nice properties.

PROPOSITION 3.3. *The function a is strictly increasing on $[0, 2\delta]$. For each $x \in (0, 2\delta)$ there is a unique sequence $(x_n)_{n\in\mathbb{Z}}$ in $(0, 2\delta)$ so that $x_{n+1} = a(x_n)$ for all $n \in \mathbb{Z}$. For the sequence $(x_n)_{n\in\mathbb{Z}}$, $\lim_{n\to-\infty} x_n = 0$ and $\lim_{n\to\infty} x_n = 2\delta$ are satisfied.*

Proof. We have

$$a'(x) = 1 + \left(\frac{1}{(1-2x^2)^{3/2}} - 1\right)\rho\left(\frac{x}{\delta}\right) + \left(\frac{x}{\sqrt{1-2x^2}} - x\right)\rho'\left(\frac{x}{\delta}\right)\frac{1}{\delta}.$$

For the last term

$$\left|\left(\frac{x}{\sqrt{1-2x^2}} - x\right)\rho'\left(\frac{x}{\delta}\right)\frac{1}{\delta}\right| = \frac{2x^3|\rho'(x/\delta)|}{\delta\sqrt{1-2x^2}(1+\sqrt{1-2x^2})} \le \frac{32\delta^3 q}{\delta} = 32q\delta^2 < 1$$

holds because of the choice of δ. Then

$$a'(x) \ge \left(\frac{1}{(1-2x^2)^{3/2}} - 1\right)\rho\left(\frac{x}{\delta}\right) > 0 \quad \text{for all } x \in (0, 2\delta).$$

Consequently, a is strictly increasing. Observe that $a(x) > x$ for all $x \in (0, 2\delta)$. Combining these facts with $a(0) = 0$ and $a(2\delta) = 2\delta$, the last two statements follow immediately. ∎

As a has an inverse, the sequence $(x_n)_{n\in\mathbb{Z}}$ guaranteed by Proposition 3.3 will be denoted also by $(a^n(x))_{n\in\mathbb{Z}}$.

PROPOSITION 3.4. *For each $(u,v) \in U$ there is a unique $(x,y) \in U$ with $G(x,y) = (u,v)$.*

Proof. Let $(u,v) \in U$ be given. Define

$$x = a^{-1}(u), \quad y = \frac{1}{e}v + \int_0^1 \frac{e^{-s}x^2}{1-2sx^2}ds\,\rho\left(\frac{x}{\delta}\right).$$

Clearly, $x \in (0, 2\delta)$ and

$$0 < y < \frac{\delta}{2e} + \frac{4\delta^2}{1-8\delta^2} \le \frac{\delta}{2e} + 8\delta^2 < \frac{\delta}{2e} + 8\delta\frac{e-2}{16e} < \frac{\delta}{e} < \frac{\delta}{2}.$$

Thus $(x,y) \in U$, and then $G(x,y) = (u,v)$. Obviously, x is uniquely determined by u. Hence the uniqueness of y also follows. ∎

By Proposition 3.4, the restriction $G|_U$ has an inverse, denoted by G^{-1}.

PROPOSITION 3.5. *For each $(x,y) \in U$,*

$$G^{-n}(x,y) \to (0,0) \quad \text{as } n \to \infty.$$

Proof. Let $(x, y) \in U$ be given. Set $(x_{-n}, y_{-n}) = G^{-n}(x, y)$, $n \in N$. By Proposition 3.3,

$$x_{-n} = a^{-n}(x) \to 0 \quad \text{as } n \to \infty.$$

Let $\tilde{\epsilon} > 0$ be given. There exists $n_0 \in \mathbb{N}$ such that

$$\int_0^1 \frac{e^{-s}x_{-n}^2}{1 - 2sx_{-n}^2} ds\, \rho\left(\frac{x_{-n}}{\delta}\right) < \frac{\tilde{\epsilon}}{2}$$

for all $n \geq n_0$. For all $k \in \mathbb{N}$, we have

$$y_{-n_0-k} = ey_{-n_0-k-1} - e\int_0^1 \frac{e^{-s}x_{-n_0-k-1}^2}{1 - 2sx_{-n_0-k-1}^2} ds\, \rho\left(\frac{x_{-n_0-k-1}}{\delta}\right),$$

from which

$$|y_{-n_0-k-1}| \leq \frac{1}{e}|y_{-n_0-k}| + \frac{\tilde{\epsilon}}{2} \quad \text{for all } k \in \mathbb{N}.$$

Using $|y_{-n_0}| < \delta/2$, an induction argument yields

$$|y_{-n_0-k-1}| \leq \frac{1}{e^{k+1}}\frac{\delta}{2} + \left(\sum_{j=0}^{k} \frac{1}{e^j}\right)\frac{\tilde{\epsilon}}{2} \quad \text{for all } k \in \mathbb{N}.$$

Hence

$$\limsup_{n\to\infty} |y_{-n}| < \tilde{\epsilon}$$

follows. As $\tilde{\epsilon} > 0$ was arbitrary, we conclude $\lim_{n\to\infty} y_{-n} = 0$. ∎

Define

$$c_1 = \frac{3\delta + a(\delta)}{4\delta}, \quad c_2 = \frac{\delta + 3a(\delta)}{4\delta}$$

and

$$x^* = \frac{\delta + a(\delta)}{2}, \quad I = \left(x^* - \frac{a(\delta) - \delta}{4}, x^* + \frac{a(\delta) - \delta}{4}\right) = (c_1\delta, c_2\delta).$$

Let $\beta : \mathbb{R} \to \mathbb{R}$ be a C^∞-smooth function so that $\beta(t) > 0$ for $c_1 < t < c_2$, and $\beta(t) = 0$ for all $t \in \mathbb{R} \setminus (c_1, c_2)$. Fix $\kappa > 0$ so small that

$$\rho + \kappa\beta \in \mathcal{R}.$$

Set $\rho_\kappa = \rho + \kappa\beta$. If ρ is replaced by ρ_κ, then, with the same fixed $\eta_0 > 0$ and $\delta > 0$, we get G_κ, W_κ, h_κ, a_κ, a_κ^{-1}, and G_κ^{-1} instead of G, W, h, a, a^{-1}, and G^{-1}, respectively, and G_κ, W_κ, h_κ, a_κ, a_κ^{-1}, and G_κ^{-1} have the same properties as obtained above for G, W, h, a, a^{-1}, and G^{-1}, respectively.

We have

$$\rho\left(\frac{x}{\delta}\right) = \rho_\kappa\left(\frac{x}{\delta}\right) \quad \text{for all } x \in (0, 2\delta) \setminus I.$$

As $I \subset (\delta, a(\delta))$ and $(a^n(\delta))_{n\in\mathbb{Z}}$ is increasing, we obtain that

$$G(a^n(\delta), h(a^n(\delta))) = G_\kappa(a^n(\delta), h(a^n(\delta))) \quad \text{for all } n \in \mathbb{Z}.$$

The positive invariance of W with respect to G and

$$\big(G(a^n(\delta), h(a^n(\delta)))\big)_1 = a^{n+1}(\delta)$$

combined yield

$$G(a^n(\delta), h(a^n(\delta))) = \big(a^{n+1}(\delta), h(a^{n+1}(\delta))\big)$$

for all $n \in \mathbb{Z}$. Hence

$$G_\kappa(a^n(\delta), h(a^n(\delta))) = \big(a^{n+1}(\delta), h(a^{n+1}(\delta))\big) \quad \text{for all } n \in \mathbb{Z}$$

follows. From $(a^n(\delta), h(a^n(\delta))) \in W$, $n \in \mathbb{Z}$, we get

$$\big(a^{n+k}(\delta), h(a^{n+k}(\delta))\big)_{k=0}^{\infty} \in \mathbb{R}^2_{\eta_0} \quad \text{for all } n \in \mathbb{Z}.$$

These facts imply

$$(a^n(\delta), h(a^n(\delta))) \in W_\kappa \quad \text{for all } n \in \mathbb{Z}.$$

PROPOSITION 3.6. $G_\kappa(x^*, h(x^*)) \notin W$.

Proof. Assume $G_\kappa(x^*, h(x^*)) \in W$. Clearly, $G(x^*, h(x^*)) \in W$ because of the positive invariance of W with respect to G. We have

$$G(x^*, h(x^*)) = \begin{pmatrix} x^* + \left(\frac{x^*}{\sqrt{1-2x^{*2}}} - x^*\right) \rho\left(\frac{x^*}{\delta}\right) \\ eh(x^*) - e\int_0^1 \frac{e^{-s}x^{*2}}{1-2sx^{*2}} ds\, \rho\left(\frac{x^*}{\delta}\right) \end{pmatrix}$$

and

$$G_\kappa(x^*, h(x^*)) = \begin{pmatrix} x^* + \left(\frac{x^*}{\sqrt{1-2x^{*2}}} - x^*\right) \left(\rho\left(\frac{x^*}{\delta}\right) + \kappa\beta\left(\frac{x^*}{\delta}\right)\right) \\ eh(x^*) - e\int_0^1 \frac{e^{-s}x^{*2}}{1-2sx^{*2}} ds \left(\rho\left(\frac{x^*}{\delta}\right) + \kappa\beta\left(\frac{x^*}{\delta}\right)\right) \end{pmatrix}.$$

Using the fact that W is the graph of h, and h is Lipschitz continuous with Lipschitz constant 2, one gets

$$\left|\big(G(x^*, h(x^*))\big)_2 - \big(G_\kappa(x^*, h(x^*))\big)_2\right| \le 2\left|\big(G(x^*, h(x^*))\big)_1 - \big(G_\kappa(x^*, h(x^*))\big)_1\right|,$$

that is

$$e\int_0^1 \frac{e^{-s}x^{*2}}{1-2x^{*2}} ds\, \kappa\beta\left(\frac{x^*}{\delta}\right) \le 2\left(\frac{x^*}{\sqrt{1-2x^{*2}}} - x^*\right) \kappa\beta\left(\frac{x^*}{\delta}\right).$$

As $x^* > 0$, $\kappa > 0$, and $\beta(x^*/\delta) > 0$, the last inequality is equivalent to

$$e\int_0^1 \frac{e^{-s}}{-2sx^{*2}} ds \le \frac{4x^*}{\sqrt{1-2x^{*2}}(1+\sqrt{1-2x^{*2}})}.$$

By $0 < x* < 2\delta$ and the choice of δ,

$$1 < e\int_0^1 \frac{e^{-s}}{1-2sx^{*2}} ds \le \frac{4x^*}{\sqrt{1-2x^{*2}}(1+\sqrt{1-2x^{*2}})} \le 16\delta < \frac{e-2}{e},$$

a contradiction. Therefore, $G_\kappa(x^*, h(x^*)) \notin W$. ∎

Define $y^* \in (0, \delta/2)$ so that

$$(x^*, y^*) = G_\kappa^{-1}(a_\kappa(x^*), h(a_\kappa(x^*))).$$

Clearly, $G_\kappa(x^*, y^*) = (a_\kappa(x^*), h(a_\kappa(x^*))) \in W$. By Proposition 3.6, $y^* \neq h(x^*)$, that is

$$(x^*, y^*) \notin W.$$

Observe that $a_\kappa(x) \geq a(x)$ for all $x \in [0, 2\delta]$. Hence we get

$$a_\kappa(x^*) \geq a(x^*) > a(\delta),$$

and for each $n \in \mathbb{N}$,

$$a_\kappa^n(a_\kappa(x^*)) = a^n(a_\kappa(x^*)) > a(\delta).$$

This fact and the invariance of W combined give

$$\begin{aligned} G_\kappa^{n+1}(x^*, y^*) &= G_\kappa^n(G_\kappa(x^*, y^*)) = G^n\left(a_\kappa(x^*), h(a_\kappa(x^*))\right) \\ &= (a^n(a_\kappa(x^*)), h(a^n(a_\kappa(x^*)))) \in W \end{aligned}$$

for all $n \in \mathbb{N}$. It follows that

$$(G_\kappa^n(x^*, y^*))_0^\infty \in \mathbb{R}^2_{\eta_0}.$$

By the analogue of Proposition 3.4 for G_κ, $G_\kappa^{-n}(x^*, y^*) \in U$ is well defined for all $n \in \mathbb{N}$. Consequently,

$$G_\kappa^n(x^*, y^*) \in W_\kappa \quad \text{for all } n \in \mathbb{Z}.$$

We have

$$G_\kappa^n(x^*, y^*) = (a_\kappa^n(x^*), h_\kappa(a_\kappa^n(x^*))) \quad \text{for all } n \in \mathbb{Z}.$$

If $x \in [\delta, x^*]$, then $a_\kappa(x) \geq a(x) \geq a(\delta)$. Thus $a_\kappa^{-1}(x^*) < \delta$, and

$$a_\kappa^{-n}(x^*) < \delta \quad \text{for all } n \in \mathbb{N} \setminus \{0\}.$$

Since $\rho(x/\delta) = \rho_\kappa(x/\delta) = 1$ for $0 < x < \delta$, we obtain that

$$G_\kappa^{-n}(x^*, y^*) = G^{-n}(x^*, y^*) \quad \text{for all } n \in \mathbb{N}.$$

This equality, the positive invariance of W with respect to G, and $(x^*, y^*) \notin W$ combined yield

$$G_\kappa^{-n}(x^*, y^*) \notin W \quad \text{for all } n \in \mathbb{N}.$$

By the analogue of Proposition 3.5 for G_κ, $G_\kappa^{-n}(x^*, y^*) \to (0, 0)$ as $n \to \infty$ also holds.

For $\gamma \in (0, \delta)$, the sets

$$W(\gamma) = \{(x, h(x)) : |x| < \gamma\}, \quad W_\kappa(\gamma) = \{(x, h_\kappa(x)) : |x| < \gamma\}$$

are local center-stable manifolds of f. Set

$$W^+(\gamma) = \{(x, h(x)) : 0 < x < \gamma\}, \quad W_\kappa^+(\gamma) = \{(x, h_\kappa(x)) : 0 < x < \gamma\}.$$

PROPOSITION 3.7. *If $W(\gamma)$ and $W_\kappa(\gamma)$ are locally positively invariant with respect to Φ, then*

$$W^+(\gamma) = \{\Phi(t, \gamma, h(\gamma)) : t < 0\},$$
$$W_\kappa^+(\gamma) = \{\Phi(t, \gamma, h_\kappa(\gamma)) : t < 0\}.$$

Proof. We know that

$$G^n(\gamma, h(\gamma)) = (a^n(\gamma), h(a^n(\gamma))) \in W \quad \text{for all } n \in \mathbb{Z},$$

where $(a^n(\gamma))_{n\in\mathbb{Z}}$ is a strictly increasing sequence with $\lim_{n\to-\infty} a^n(\gamma) = 0$. It is also true that

$$G^{-n}(\gamma, h(\gamma)) = f^{-n}(\gamma, h(\gamma)) = \Phi(-n, \gamma, h(\gamma)) \in W(\gamma)$$

for all $n \in \mathbb{N} \setminus \{0\}$. Assume that $t_0 < 0$ and $\Phi(t_0, \gamma, h(\gamma)) \notin W^+(\gamma)$. Choose $n_0 \in \mathbb{N} \setminus \{0\}$ with $-n_0 < t_0 < -n_0 + 1$. Define

$$t_1 = \sup\left\{s \in [-n_0, t_0) : \Phi(s, \gamma, h(\gamma)) \in W^+(\gamma)\right\}.$$

Clearly, $-n_0 \le t_1 < t_0$ and

$$\Phi(t_1, \gamma, h(\gamma)) \in W^+(\gamma).$$

The local positive invariance of $W(\gamma)$ gives $t^* > 0$ so that $\Phi(s, \Phi(t_1, \gamma, h(\gamma))) = \Phi(t_1 + s, \gamma, h(\gamma)) \in W^+(\gamma)$ for all $s \in [0, t^*]$. This contradicts the definition of t_1. Consequently,

$$\{\Phi(t, \gamma, h(\gamma)) : t < 0\} \subset W^+(\gamma).$$

The inclusion $\{\Phi(t, \gamma, h(\gamma)) : t < 0\} \supset W^+(\gamma)$ follows from $\{\big(\Phi(t, \gamma, h(\gamma))\big)_1 : t < 0\} = (0, \gamma)$.

The proof for $W_\kappa^+(\gamma)$ is analogous. □

Finally, we can state the following.

PROPOSITION 3.8. *One of the local manifolds $W(\gamma)$ and $W_\kappa(\gamma)$ is not locally positively invariant with respect to Φ.*

Proof. Assume that both $W(\gamma)$ and $W_\kappa(\gamma)$ are locally positively invariant with respect to Φ. By Proposition 3.7 either

$$W^+(\gamma) = W_\kappa^+(\gamma)$$

or

$$W^+(\gamma) \cap W_\kappa^+(\gamma) = \emptyset$$

holds. On the other hand,

$$(a^n(\delta), h(a^n(\delta))) \in W \cap W_\kappa \quad \text{for all } n \in \mathbb{Z},$$
$$G_\kappa^{-n}(x^*, y^*) \in W_\kappa \setminus W \qquad \text{for all } n \in \mathbb{N},$$

and $\lim_{n\to-\infty} a^n(\delta) = 0$, $\lim_{n\to\infty} G_\kappa^{-n}(x^*, y^*) = (0, 0)$ combined imply

$$W^+(\gamma) \neq W^+_\kappa(\gamma)$$

and

$$W^+(\gamma) \cap W^+_\kappa(\gamma) \neq \emptyset,$$

which is a contradiction. □

We remark that in the above example W and W_κ are in fact the global center manifolds of G and G_κ, respectively, because of $E_s = \{(0,0)\}$. Then $W(\gamma)$ and $W_\kappa(\gamma)$, $\gamma \in (0,\delta)$, are local center manifolds of the time-1 map $\Phi(1,\cdot)$ at $(0,0)$. These local center manifolds depend on δ and ρ, ρ_κ, too. Thus, in the next result, where their dependence on the elements of $\mathcal{R}$ is considered, the notation $W_{\delta\rho}(\gamma)$ will be used instead of $W(\gamma)$.

PROPOSITION 3.9. *Let Φ, r, the norm $||\cdot||$ in $\mathbb{R}^2$, the family of mappings $\tilde{r}_{\delta\rho}$, $0 < \delta \leq \delta_1$, $\rho \in \mathcal{R}$, and the constants η_0, δ^{**} be given as above. Let $\delta \in (0, \delta^{**})$ be fixed so that* (3.4) *holds. Fix a $\gamma \in (0,\delta)$.*

Then the set of those elements $\rho \in \mathcal{R}$, for which the local center manifold $W_{\delta\rho}(\gamma)$ of the time-1 map $\Phi(1,\cdot)$ is not locally positively invariant with respect to Φ, is an open and dense subset of $\mathcal{R}$.

Proof. Let $\mathcal{Q}$ be the set of $\rho \in \mathcal{R}$ such that $W_{\delta\rho}(\gamma)$ is not locally positively invariant with respect to Φ.

If $\rho \in \mathcal{R} \setminus \mathcal{Q}$, then the above construction shows that for all sufficiently small $\kappa > 0$, $W_{\delta\rho_\kappa}(\gamma) \in \mathcal{Q}$. Thus, $\mathcal{Q}$ is dense in $\mathcal{R}$.

If $\rho \in \mathcal{Q}$, then there exist $u_0 \in W_{\delta\rho}(\gamma)$ and a $t_0 > 0$ such that $\Phi(t_0,u_0) \notin W_{\delta\rho}(\gamma)$ and $\left|\left(\Phi(t,u_0)\right)_1\right| < \gamma$ for all $t \in [0,t_0]$. There is an $\epsilon_1 > 0$ so that for every $u_1 \in \mathbb{R}^2$ with $|u_1 - u_0| < \epsilon_1$ we have $\operatorname{dist}(\Phi(t_0,u_1), W_{\delta\rho}(\gamma)) > \epsilon_1$, and $\left|\left(\Phi(t,u_1)\right)_1\right| < \gamma$ for all $t \in [0,t_0]$. By Proposition 2.2, there is $\epsilon_2 > 0$ such that for each $\zeta \in \mathcal{R}$ with $d(\rho,\zeta) < \epsilon_2$, $\operatorname{dist}(u_0, W_{\delta\zeta}(\gamma)) < \epsilon_1$ and $\operatorname{dist}(\Phi(t_0,u_1), W_{\delta\zeta}(\gamma)) > \epsilon_1/2$ for all $u_1 \in \mathbb{R}^2$ with $|u_1 - u_0| < \epsilon_1$. Let $\zeta \in \mathcal{R}$ be given with $d(\rho,\zeta) < \epsilon_2$, and choose $u_1 \in W_{\delta\zeta}(\gamma)$ such that $|u_1 - u_0| < \epsilon_1$. Then $\Phi(t_0,u_1) \notin W_{\delta\zeta}(\gamma)$ and $\left|\left(\Phi(t,u_1)\right)_1\right| < \gamma$ for all $t \in [0,t_0]$. Define

$$t_1 = \sup\{t \in [0,t_0] : \Phi(t,u_1) \in W_{\delta\zeta}(\gamma)\}.$$

It is easy to see that $t_1 < t_0$, $\Phi(t_1,u_1) \in W_{\delta\zeta}(\gamma)$, and $\Phi(t,u_1) \notin W_{\delta\zeta}(\gamma)$ for all $t \in (t_1,t_0]$. It follows that $W_{\delta\zeta}(\gamma)$ is not locally positively invariant under Φ. So, $\zeta \in \mathcal{Q}$. Therefore, $\mathcal{Q}$ is an open subset of $\mathcal{R}$. □

4. Invariance: An abstract functional differential equation. Let Φ be a C^k-smooth semiflow on the Banach space E with $\Phi(t,0) = 0$ for all $t \geq 0$. Fix $\tau > 0$. There exists an open neighborhood V of 0 in E so that the time-τ map $f = \Phi(\tau,\cdot)$ is defined on V. Let $L = Df(0)$, and assume that E has the decomposition and the exponential trichotomy with the norm $|\cdot|$ as described in section 2. Let $r : E \to E$ and the equivalent norm $||\cdot||$ on E also be given as in section 2.

The classical solution to the invariance problem consists in assuming that the semiflow can be modified outside a small neighborhood of the equilibrium point so that the modified semiflow is a small and Lipschitzian perturbation of the linearized semiflow, and taking as a modification of the time-τ map the time-τ map of the modified semiflow [5]. We formulate this as a proposition for the semiflow Φ with the conditions on the time-τ map f given in section 2. The proof is omitted since it is essentially the same as the proof in [5].

PROPOSITION 4.1. *Let Φ, L, and r be given as above. Assume that there is a $\delta^*>0$, and that for each $\delta \in (0,\delta^*]$ there exists a semiflow $\Phi_\delta : [0,\infty)\times E \to E$ such that*

(a) $\Phi_\delta|_{[0,\tau]\times E(\delta)} = \Phi|_{[0,\tau]\times E(\delta)}$ *for all* $\delta \in (0,\delta^*]$;

(b) *the family of mappings* $r_\delta := \Phi_\delta(\tau,\cdot) - L$, $0 < \delta \le \delta^*$, *satisfies hypotheses* (R1) *and* (R2);

(c) *for every* $\delta \in (0,\delta^*]$, *there exists* $K > 0$ *so that* $|\Phi_\delta(t,x)| \le K|x|$ *for all* $(t,x) \in [0,\tau]\times E$.

Then the C^k-smooth local center-stable manifold W^{cs}_{loc} of $\Phi(\tau,\cdot)$, given in Theorem 2.1, *is invariant with respect to Φ relative to $N_{sc}+N_u$.*

Analogous results can be stated for the local center-stable and center manifolds.

The paper [5] by Chen, Hale, and Tan contains several examples where a modification Φ_δ of Φ with the required properties in Theorem 4.1 can be obtained. All of these semiflows are generated by evolutionary equations, and a variation-of-constants formula is valid for the evolutionary equation.

Now we demonstrate the applicability of Proposition 4.1 to (1.1), which does not fall into the class of examples considered in [5] since a suitable variation-of-constants formula in the phase space is not known.

Consider the abstract semilinear functional differential equation

$$\dot u(t) = A_T u(t) + Bu_t + F(u_t) \tag{4.1}$$

in a Banach space X over $\mathbb{R}$, where $r \ge 0$, $C = C([-r,0];X)$ is the Banach space of continuous mappings from $[-r,0]$ into X with the supremum norm $||\cdot||_0$, $u_t \in C$ is defined by $u_t(\theta) = u(t+\theta)$ for $\theta \in [-r,0]$, $B : C \to X$ is a bounded linear operator, $A_T : D(A_T) \subset X \to X$ is the infinitesimal generator of a compact C_0-semigroup $(T(t))_{t\ge 0}$ of linear operators on X, $F : V_1 \to X$ is C^k-smooth, $k \in \mathbb{N}\setminus\{0\}$ on an open neighborhood V_1 of 0 in C, and $F(0) = 0$, $DF(0) = 0$.

The following results on the decomposition of C and on the existence and smoothness of solutions of (4.1) can be found in the works [13, 14, 29, 32]. For other results and applications of (4.1) we refer to [10, 11, 20, 21, 22, 28, 32].

For any $\phi \in C$ there exists a unique continuous function $u = u(\phi) : [-r,\infty) \to X$ such that

$$u(t) = T(t)\phi(0) + \int_0^t T(t-s)Bu_s\,ds, \quad t \ge 0,$$

and $u_0 = \phi$ holds. The operators $U(t) : C \to C$, given by $U(t)\phi = u_t(\phi)$, form a C_0-semigroup of bounded linear operators on C; moreover, $U(t)$ is compact for all $t > r$.

Fix $\tau > r$. Since $U(\tau)$ is compact, its spectrum $\sigma(U(\tau))$ consists of three disjoint compact sets σ_s, σ_c, and σ_u, where σ_s is contained in $\{z \in \mathbb{C} : |z| < 1\}$, and σ_c and σ_u are finite sets in $\{z \in \mathbb{C} : |z| = 1\}$ and $\{z \in \mathbb{C} : |z| > 1\}$, respectively. We assume $\sigma_c \ne \emptyset$. Set $a = \max_{\lambda\in\sigma_s}|\lambda|$ and $b = \min_{\lambda\in\sigma_u}|\lambda|$. As in section 2, $a = 1/2$ if $\sigma_s = \emptyset$, $b = 3/2$ if $\sigma_u = \emptyset$. Choose $\epsilon > 0$ with $a+\epsilon < 1$ and $(1+\epsilon)^k < b - \epsilon$. Let $\alpha = (1/\tau)\log(a+\epsilon)$, $\beta = (1/\tau)\log(b-\epsilon)$, $\kappa = (1/\tau)\log(1+\epsilon)$. The space C has the decomposition

$$C = C_s \oplus C_c \oplus C_u$$

into closed subspaces C_s, C_c, C_u of C so that C_c and C_u are finite-dimensional, the spaces C_s, C_c, and C_u are invariant under $U(t)$ for all $t \ge 0$, and $(U(t))_{t\ge 0}$ can be extended to a group on C_c and C_u. There exists a norm $|\cdot|$ on C which is equivalent to $||\cdot||_0$ and satisfies $|\phi| = |\operatorname{Pr}_{C_s}\phi| + |\operatorname{Pr}_{C_c}\phi| + |\operatorname{Pr}_{C_u}\phi|$ and

$$
\begin{aligned}
&|U(t)\,\mathrm{Pr}_{C_s}\,\phi| \le e^{\alpha t}|\,\mathrm{Pr}_{C_s}\,\phi| \quad (t \ge 0),\\
&|U(t)\,\mathrm{Pr}_{C_c}\,\phi| \le e^{\kappa|t|}|\,\mathrm{Pr}_{C_c}\,\phi| \quad (t \in \mathbb{R}),\\
&|U(t)\,\mathrm{Pr}_{C_u}\,\phi| \le e^{\beta t}|\,\mathrm{Pr}_{C_u}\,\phi| \quad (t \le 0)
\end{aligned} \tag{4.2}
$$

for all $\phi \in C$. In addition, $|\,\mathrm{Pr}_{C_s}\,| = |\,\mathrm{Pr}_{C_c}\,| = |\,\mathrm{Pr}_{C_u}\,| = 1$ also holds.

For any $\phi \in V_1$ there exist a maximal $t_\phi > 0$ and a unique continuous function $u^\phi : [-r, t_\phi) \to X$ such that $u_0^\phi = \phi$ and

$$u^\phi(t) = T(t)\phi(0) + \int_0^t T(t-s)\left[Bu_s^\phi + F(u_s^\phi)\right]\,ds$$

for all $t \in [0, t_\phi)$. The function u^ϕ is called the mild solution of (4.1) through $(0, \phi)$.

There is an open neighborhood $V_2 \subset V_1$ of 0 in C so that Ψ, given by $\Psi(t, \phi) = u_t^\phi$, is a C^k-smooth semiflow on V_2 with $\Psi(t, 0) = 0$ and $D_2\Psi(t, 0) = U(t)$ for all $t \ge 0$.

Let $V \subset V_2$ be an open neighborhood of 0 in C such that $[0, \tau] \times V$ is in the domain of Ψ. Define $R : C \to C$ so that $R(\phi) = \Psi(\tau, \phi) - U(\tau)\phi$ for $\phi \in V$ and $R(\phi) = 0$ for $\phi \in C \setminus V$.

Under the above assumptions, there exist C^k-smooth local center-stable, center-unstable, and center manifolds of the semiflow Ψ at 0. More precisely, we have the following.

THEOREM 4.2.

(i) *There exist convex open neighborhoods N_{sc} of 0 in C_{sc}, N_u of 0 in C_u, and a C^k-smooth map $w_{cs} : N_{sc} \to C_u$ so that $w_{cs}(0) = 0$, $Dw_{cs}(0) = 0$, $w_{cs}(N_{sc}) \subset N_u$, the set $W_{loc}^{cs} = \{z + w_{cs}(z) : z \in N_{sc}\}$ is invariant with respect to Ψ relative to $N_{sc} + N_u$, and*

$$\{\Psi(t, \phi) : t \ge 0\} \subset N_{sc} + N_u \quad \textit{implies} \quad \phi \in W_{loc}^{cs}.$$

(ii) *There exist convex open neighborhoods N_{cu} of 0 in C_{cu}, N_s of 0 in C_s, and a C^k-smooth map $w_{cu} : N_{cu} \to C_s$ so that $w_{cu}(0) = 0$, $Dw_{cu}(0) = 0$, $w_{cu}(N_{cu}) \subset N_s$, the set $W_{loc}^{cu} = \{z + w_{cu}(z) : z \in N_{cu}\}$ is invariant with respect to Ψ relative to $N_{cu} + N_s$, and for every backward trajectory $y : (-\infty, 0] \to C$ of Ψ,*

$$\{y(t) : t \le 0\} \subset N_{cu} + N_s \quad \textit{implies} \quad y(0) \in W_{loc}^{cu}.$$

(iii) *There exist convex open neighborhoods N_c of 0 in C_c, N_{su} of 0 in C_{su}, and a C^k-smooth map $w_c : N_c \to C_{su}$ so that $w_c(0) = 0$, $Dw_c(0) = 0$, $w_c(N_c) \subset N_{su}$, the set $W_{loc}^{c} = \{z + w_c(z) : z \in N_c\}$ is invariant with respect to Ψ relative to $N_c + N_{su}$, and for every trajectory $y : \mathbb{R} \to C$ of Ψ,*

$$\{y(t) : t \in \mathbb{R}\} \subset N_c + N_{su} \quad \textit{implies} \quad y(0) \in W_{loc}^{c}.$$

We show only statement (i) in Theorem 4.2 since the proof of (ii) is analogous, and (iii) is a consequence of (i) and (ii) by the proof of Theorem 5.4 in [13].

The proof of (i) *in Theorem* 4.2. Choose a norm $|\cdot|_{cu}$ on the finite-dimensional C_{cu} so that $|\cdot|_{cu}$ is C^∞-smooth on $C_{cu} \setminus \{0\}$. Then $||\phi|| = \max\{|\,\mathrm{Pr}_{C_s}\,\phi|, |\,\mathrm{Pr}_{C_{cu}}\,|_{cu}\}$, $\phi \in C$, defines a new norm on C which is equivalent to the norms $||\cdot||_0$ and $|\cdot|$ on C. There are positive constants c_1, c_2, c_3, c_4 such that $c_1|\phi| \le ||\phi|| \le c_2|\phi|$ and $c_3|\phi| \le ||\phi||_0 \le c_4|\phi|$ for all $\phi \in C$. For $\delta > 0$ let $C(\delta) = \{\phi \in C : ||\phi|| < \delta\}$.

Let $\rho \in \mathcal{R}$. Fix a $\gamma_0 > 0$ so that $C(2\gamma_0) \subset V$ and all lth derivatives, $l \in \{0, \dots, k\}$, of $F|_{C(2\gamma_0)}$ are bounded. Let $F^* : C \to X$ be an extension of F so that $F^*(\phi) = 0$ for $\phi \in C \setminus V_1$. For each $\gamma \in (0, \gamma_0]$ define $F_\gamma : C \to X$ by

$$F_\gamma(\phi) = F^*(\phi)\rho\left(\frac{|\operatorname{Pr}_{C_{cu}} \phi|_{cu}}{\gamma}\right)\rho\left(\frac{|\operatorname{Pr}_{C_s} \phi|}{\gamma}\right).$$

Then, for every $\gamma \in (0, \gamma_0]$, F_γ is C^k-smooth on the open set $\{\phi \in C : |\operatorname{Pr}_{C_s} \phi| < \gamma\}$, and all lth derivatives, $l \in \{0, \dots, k\}$, of $F_\gamma|_{\{\phi \in C : |\operatorname{Pr}_{C_s} \phi| < \gamma\}}$ are bounded.

There exits $\gamma_1 \in (0, \gamma_0]$ and a nondecreasing function $\mu : [0, \gamma_1] \to [0, 1]$ (the proof is the same as that of Proposition II.2 in [19]) with $\lim_{\gamma \to 0^+} \mu(\gamma) = 0 = \mu(0)$ such that for each $\gamma \in (0, \gamma_1]$ and for all ϕ, ψ in C,

$$||F_\gamma(\phi)||_X \le \gamma\mu(\gamma), \quad ||F_\gamma(\phi) - F_\gamma(\psi)||_X \le \mu(\gamma)||\phi - \psi||_0.$$

There exists a constant $M \ge 1$ so that $||B|| \le M$, and $||T(t)|| \le M$, $||U(t)|| \le M$ for all $t \in [0, \tau]$.

For each $\gamma \in (0, \gamma_1]$, consider the modification

$$\dot{u}(t) = A_T u(t) + Bu_t + F_\gamma(u_t) \tag{4.3}$$

of equation (4.1). By [32], for every $\phi \in C$ there exists a unique continuous function $u = u^{\phi,\gamma} : [-r, \infty) \to X$ so that $u_0 = \phi$ and

$$u(t) = T(t)\phi(0) + \int_0^t T(t-s)\left(Bu_s + F_\gamma(u_s)\right) ds$$

for all $t \ge 0$. The map

$$\Psi_\gamma : [0, \infty) \times C \ni (t, \phi) \mapsto u_t^{\phi,\gamma} \in C$$

is a global semiflow on C.

By the Gronwall inequality, from the integral equation of $u^{\phi,\gamma}$, it is easy to see that there exists $K \ge 1$ such that, for each $\gamma \in (0, \gamma_1]$,

$$|\Psi_\gamma(t, \phi)| \le K|\phi| \quad (t \in [0, \tau],\ \phi \in C). \tag{4.4}$$

Set $c = \max\{2, Kc_2/c_1\}$, and choose $\gamma_2 \in (0, \gamma_1]$ such that

$$\frac{c_4 + c}{c_3}\tau M^2 e^{2\tau M(M+1)}\mu(\gamma_2) < 1.$$

We state that for each $\gamma \in (0, \gamma_2]$,

$$|U(t)\phi - \Psi_\gamma(t, \phi)| < \frac{\gamma}{2} \quad (t \in [0, \tau],\ \phi \in C). \tag{4.5}$$

Set $u_s = U(s)\phi$ and $v_s = \Psi_\gamma(s, \phi)$ for $s \ge 0$. From the integral equations for u and v, we get

$$||u_t - v_t||_0 \le \int_0^t M(M||u_s - v_s||_0 + \gamma\mu(\gamma))\, ds \le \tau M\gamma\mu(\gamma) + \int_0^t M^2||u_s - v_s||_0\, ds$$

for $t \in [0, \tau]$. Gronwall's inequality yields

$$||u_t - v_t||_0 \le \tau M\gamma\mu(\gamma)e^{\tau M^2} < \frac{c_3}{c_4 + c}\gamma < \frac{c_3}{2}\gamma$$

for $t \in [0,\tau]$. Now (4.5) follows by the equivalence of the norms $||\cdot||_0$ and $|\cdot|$.

We claim that for each $\gamma \in (0,\gamma_2]$,

$$|\operatorname{Pr}_{C_s}\phi| < \frac{\gamma}{2} \quad \text{implies} \quad |\operatorname{Pr}_{C_s}\Psi_\gamma(t,\phi)| < \gamma$$

for all $t \in [0,\tau]$ and $\phi \in C$.

Applying (4.5) and the exponential trichotomy (4.2) for U, we get for every $t \in [0,\tau]$ and for all $\phi \in C$ with $|\operatorname{Pr}_{C_s}\phi| < \gamma/2$ that

$$\begin{aligned}|\operatorname{Pr}_{C_s}\Psi_\gamma(t,\phi)| &\le |\operatorname{Pr}_{C_s}(\Psi_\gamma(t,\phi) - U(t)\phi)| + |\operatorname{Pr}_{C_s}U(t)\phi| \\ &\le \frac{\gamma}{2} + e^{\alpha t}|\operatorname{Pr}_{C_s}\phi| < \frac{\gamma}{2} + \frac{\gamma}{2} = \gamma.\end{aligned}$$

The above result and the C^k-smoothness of $F_\gamma|_{\{\phi\in C:|\operatorname{Pr}_{C_s}\phi|<\gamma\}}$ makes it possible to prove, for each $t \in [0,\tau]$, that the map

$$\left\{\phi \in C : |\operatorname{Pr}_{C_s}\phi| < \frac{\gamma}{2}\right\} \ni \psi \mapsto \Psi_\gamma(t,\psi) \in C$$

is C^k-smooth. We omit the proof since it goes as that of Lemma 6.1 in [13].

For $\gamma \in (0,\gamma_2]$, define $R_\gamma : C \to C$ by

$$R_\gamma(\phi) = \Psi_\gamma(\tau,\phi) - U(\tau)\phi.$$

By a standard induction procedure it can be shown that all jth derivatives, $j \in \{0,\dots,k\}$, of R_γ are bounded on the open set $\{\phi \in C : |\operatorname{Pr}_{C_s}\phi| < \gamma/2\}$.

Finally we show that Proposition 4.1 can be applied with $E = C$, $\Phi = \Psi$, $L = U(\tau)$, $r = R$.

By the equivalence of the norms $|\cdot|$ and $||\cdot||$ in C, and by (4.4), the estimate

$$||\Psi_\gamma(t,\phi)|| \le c||\phi|| \quad (\phi \in C,\ t \in [0,\tau]) \tag{4.6}$$

holds with $c = \max\{2, Kc_2/c_1\}$.

Define $\delta^* = \gamma_2/c$, and for $\delta \in (0,\delta^*]$ set $\Phi_\delta = \Psi_{c\delta}$. Then $r_\delta = R_{c\delta}$, $\delta \in (0,\delta^*]$, and obviously $E(\delta^*) \subset V$.

Since $c\delta \le c\delta^* = \gamma_2$, $\Phi_\delta : [0,\infty)\times E \to E$ is a global semiflow for all $\delta \in (0,\delta^*]$. Hypothesis (c) of Proposition 4.1 follows from (4.4).

Let $\delta \in (0,\delta^*]$ be fixed. If $\phi \in C(\delta)$, that is $||\phi|| < \delta$, then by (4.6)

$$||\Phi_\delta(t,\phi)|| = ||\Psi_{c\delta}(t,\phi)|| \le c||\phi|| < c\delta \quad (0 \le t \le \tau).$$

Using $F_{c\delta}|_{C(c\delta)} = F|_{C(c\delta)}$, it follows that $F_{c\delta}(\Psi_{c\delta}(t,\phi)) = F(\Psi(t,\phi))$ for all $\phi \in C(\delta)$ and $t \in [0,\tau]$. Thus, for $\phi \in C(\delta)$, both $\Psi_{c\delta}(t,\phi)$ and $\Psi(t,\phi)$ satisfy the same integral equation on the interval $[0,\tau]$. By uniqueness, $\Psi_{c\delta}(t,\phi) = \Psi(t,\phi)$ follows for $0 \le t \le \tau$. As $\phi \in C(\delta)$ and $\delta \in (0,\delta^*]$ were arbitrary, hypothesis (a) of Proposition 4.1 is also satisfied.

We have $r_\delta = R_{c\delta} = \Psi_{c\delta}(\tau,\cdot) - U(\tau) = \Phi_\delta(\tau,\cdot) - L$, and thus, by using (a), we get $r_\delta|_{C(\delta)} = r|_{C(\delta)}$. For each $\delta \in (0,\delta^*]$, the map $R_{c\delta}|_{\{\phi\in C:|\operatorname{Pr}_{C_s}\phi|<c\delta/2\}}$ is C^k-smooth, and its jth derivatives, $j \in \{1,\dots,k\}$, are bounded. Since $\delta \le c\delta/2$ it follows that $r_\delta|_{\{\phi\in C:|\operatorname{Pr}_{C_s}\phi|<\delta\}}$ is C^k-smooth, and all jth derivatives, $j \in \{1,\dots,k\}$, of $r_\delta|_{\{\phi\in C:|\operatorname{Pr}_{C_s}\phi|<\delta\}}$ are bounded.

Define $\lambda : [0, \delta^*] \to \mathbb{R}$ by

$$\lambda(\delta) = \mu(c\delta)\frac{c_4 + c}{c_3}\tau M^2 e^{2\tau M(M+1)}.$$

Then λ is nondecreasing, $\lim_{\delta \to 0^+} \lambda(\delta) = 0 = \lambda(0)$, and by the choice of γ_2, $\lambda([0, \delta^*]) \subset [0, 1]$.

Above we obtained

$$||r_\delta(\phi)||_0 = ||R_{c\delta}(\phi)||_0 = ||\Psi_{c\delta}(\tau, \phi) - U(\tau)\phi||_0 \le \tau M e^{\tau M^2} c\delta\mu(c\delta) \le c_3\delta\lambda(\delta)$$

for all $\phi \in C$. Therefore, $|r_\delta(\phi)| \le \delta\lambda(\delta)$, $\phi \in C$.

Let $\phi, \psi \in C$, and set $u_t = U(t)\phi$, $\tilde{u}_t = U(t)\psi$, $v_t = \Psi_{c\delta}(t, \phi)$, $\tilde{v}_t = \Psi_{c\delta}(t, \psi)$ for $0 \le t \le \tau$. Then $u_0 = v_0 = \phi$, $\tilde{u}_0 = \tilde{v}_0 = \psi$. Using the integral equations for v and $\tilde{v}$, $v_0 = \phi$, $\tilde{v}_0 = \psi$, the Lipschitz continuity of $F_{c\delta}$, and $\mu(c\delta) \le 1$,

$$||v_t - \tilde{v}_t||_0 \le M||\phi - \psi||_0 + \int_0^t M(M+1)||v_s - \tilde{v}_s||_0\, ds$$

follows for $0 \le t \le \tau$. The Gronwall inequality yields

$$||v_t - \tilde{v}_t||_0 \le Me^{\tau M(M+1)}||\phi - \psi||_0 \quad (0 \le t \le \tau). \tag{4.7}$$

The integral equations for $u, \tilde{u}, v, \tilde{v}$, the Lipschitz continuity of $F_{c\delta}$, and inequality (4.7) combined give

$$\begin{aligned} ||r_\delta(\phi) - r_\delta(\psi)||_0 &= ||v_t - u_t - \tilde{v}_t + \tilde{u}_t||_0 \\ &\le \int_0^t M^2||v_s - u_s - \tilde{v}_s + \tilde{u}_s||_0\, ds + \int_0^t M\mu(c\delta)||v_s - \tilde{v}_s||_0\, ds \\ &\le \tau M^2 e^{\tau M(M+1)}\mu(c\delta)||\phi - \psi||_0 + \int_0^t M^2||v_s - u_s - \tilde{v}_s + \tilde{u}_s||_0\, ds \end{aligned}$$

for $0 \le t \le \tau$. The Gronwall inequality implies

$$||r_\delta(\phi) - r_\delta(\psi)||_0 \le \tau M^2 e^{2\tau M(M+1)}\mu(c\delta)||\phi - \psi||_0,$$

that is

$$|r_\delta(\phi) - r_\delta(\psi)| \le \frac{c_4}{c_3}\tau M^2 e^{2\tau M(M+1)}\mu(c\delta)|\phi - \psi| \le \lambda(\delta)|\phi - \psi|.$$

Therefore, the family of mappings r_δ, $0 < \delta \le \delta^*$, satisfies hypotheses (R1) and (R2), and thus condition (b) in Proposition 4.1 holds.

Clearly, (4.2) implies that (2.1) is satisfied with $E = C$ and $L = U(\tau)$. Consequently, we can apply Theorem 2.1 with some fixed η_0 and $\delta \in (0, \delta^{**})$ to get a center-stable manifold W^{cs}_{loc} of the time-τ map $\Psi(\tau, \cdot)$ at 0 as described in statements (i) and (ii) of Theorem 2.1. The invariance of W^{cs}_{loc} with respect to Ψ relative to $N_{sc} + N_u$ is guaranteed by Proposition 4.1.

Assume that $\{\Psi(t, \phi) : t \ge 0\} \subset N_{sc} + N_u$ for some $\phi \in C$. Then $(\Psi(n\tau, \phi))_0^\infty$ is a trajectory of the time-τ map $\Psi(\tau, \cdot)$ in $N_{sc} + N_u$. Theorem 2.1 implies that $\phi \in W^{cs}_{loc}$. This completes the proof. □

Remarks. 1. Some conditions in Theorem 4.2 could be relaxed—for example, the compactness assumption on the semigroup $T(t)$. However, such a generalization is not necessary for handling the motivating example (1.2), (1.3).

2. Theorem 4.2 shows that a variation-of-constants formula in the space X is sufficient to apply the classical approach of [5]; it is not necessary to have a variation-of-constants formula in the phase space $C([-r,0];X)$.

3. Recently Hino et al. [15] proved a "limiting variation-of-constants formula" for (1.1) in the phase space $C([-r,0];X)$, and Murakami and Minh [23] used it to get some Lipschitz continuous invariant manifolds. However, it is not clear whether the variation-of-constants formula of [15] is suitable for a proof of smooth local center-stable, center-unstable, and center manifolds of (1.1) at 0.

4. The very recent paper of Minh and Wu [24] constructs smooth local center-unstable and center manifolds (but not center-stable manifolds) for (1.1) by using the graph transform method.

Acknowledgment. The author thanks the referees for their useful comments and suggestions.

REFERENCES

[1] P. W. Bates and C. K. R. T. Jones, *Invariant manifolds for semilinear partial differential equations*, in Dynamics Reported, Vol. 2, Dynam. Report. Ser. Dynam. Systems Appl. 2, Wiley, Chichester, UK, 1989, pp. 1–38.

[2] S. Busenberg and W. Huang, *Stability and Hopf bifurcation for a population delay model with diffusion*, J. Differential Equations, 124 (1996), pp. 80–107.

[3] P. Brunovský, *Controlling nonuniqueness of local invariant manifolds*, J. Reine Angew. Math., 446 (1994), pp. 115–135.

[4] J. L. Carr, *Applications of Center Manifold Theory*, Springer-Verlag, New York, 1981.

[5] X.-Y. Chen, J. K. Hale, and B. Tan, *Invariant foliations for C^1 semigroups in Banach spaces*, J. Differential Equations, 139 (1997), pp. 283–318.

[6] S.-N. Chow and J. Hale, *Methods of Bifurcation Theory*, Springer-Verlag, New York, 1982.

[7] S.-N. Chow, C. Li, and D. Wang, *Normal Forms and Bifurcation of Planar Vector Fields*, Cambridge University Press, Cambridge, UK, 1994.

[8] S.-N. Chow and K. Lu, *Invariant manifolds for flows in Banach spaces*, J. Differential Equations, 74 (1988), pp. 285–317.

[9] O. Diekmann, S. A. van Gils, S. M. Verduyn Lunel, and H.-O. Walther, *Delay Equations, Functional-, Complex-, and Nonlinear Analysis*, Springer-Verlag, New York, 1995.

[10] T. Faria, *Normal forms and Hopf bifurcations for partial differential equations with delay*, Trans. Amer. Math. Soc., 352 (2000), pp. 2217–2238.

[11] T. Faria, *Normal forms for semilinear functional differential equations in Banach spaces and applications. Part* II, Discrete Contin. Dynam. Systems, 7 (2001), pp. 155–176.

[12] T. Faria and W. Huang, *Stability of periodic solutions arising from Hopf bifurcation for a reaction-diffusion equation with time delay*, in Differential Equations and Dynamical Systems, Fields Inst. Commun. 31, AMS, Providence, RI, 2002, pp. 125–141.

[13] T. Faria, W. Huang, and J. Wu, *Smoothness of center manifolds for maps and formal adjoints for semilinear FDEs in general Banach spaces*, SIAM J. Math. Anal., 34 (2002), pp. 173–203.

[14] W. E. Fitzgibbon, *Semilinear functional differential equations in Banach spaces*, J. Differential Equations, 29 (1978), pp. 1–14.

[15] Y. Hino, S. Murakami, T. Naito, and N. V. Minh, *A variation-of-constants formula for abstract functional differential equations in the phase space*, J. Differential Equations, 179 (2002), pp. 336–355.

[16] M. W. Hirsch, C. Pugh, and M. Shub, *Invariant Manifolds*, Lecture Notes in Math. 583, Springer-Verlag, New York, 1977.

[17] W. Huang, *On asymptotic stability for linear delay equations*, Differential Integral Equations, 4 (1991), pp. 1303–1316.

[18] A. Kelley, *The stable, center-stable, center, center-unstable, unstable manifolds*, J. Differential Equations, 3 (1967), pp. 546–570.

[19] T. Krisztin, H.-O. Walther, and J. Wu, *Shape, Smoothness and Invariant Stratification of an Attracting Set for Delayed Monotone Positive Feedback*, Fields Inst. Monogr. 11, AMS, Providence, RI, 1999.

[20] X. Lin, J. So, and J. Wu, *Center manifolds for partial differential equations with delays*, Proc. Roy. Soc. Edinburgh Sect. A, 122 (1992), pp. 237–254.
[21] M. C. Memory, *Stable and unstable manifolds for partial functional differential equations*, Nonlinear Anal., 16 (1991), pp. 131–142.
[22] M. C. Memory, *Bifurcation and asymptotic behavior of solutions of a delay-differential equation with diffusion*, SIAM J. Math. Anal., 20 (1989), pp. 533–546.
[23] S. Murakami and N. V. Minh, *Some Invariant Manifolds for Abstract Functional Differential Equations and Linearized Stabilities*, preprint.
[24] N. V. Minh and J. Wu, *Invariant Manifolds of Evolutionary Processes and Partial Functional Differential Equations Revisited*, preprint.
[25] V. A. Pliss, *Nonlocal Problems of the Theory of Oscillations*, Academic Press, New York, 1966.
[26] D. Ruelle, *Elements of Differentiable Dynamics and Bifurcations Theory*, Academic Press, Boston, MA, 1989.
[27] J. Sijbrand, *Properties of center manifolds*, Trans. Amer. Math. Soc., 289 (1985), pp. 431–469.
[28] J. So, Y. Yang and J. Wu, *Center manifolds for functional partial differential equations: Smoothness and attractivity*, Math. Japonica, 48 (1998), pp. 67–81.
[29] C. C. Travis and G. F. Webb, *Existence and stability for partial functional differential equations*, Trans. Amer. Math. Soc., 200 (1974), pp. 394–418.
[30] A. Vanderbauwhede and G. Iooss, *Center manifolds in infinite dimensions*, in Dynamics Reported: Expositions in Dynamical Systems, Dynam. Report. Expositions Dynam. Systems (N.S.) 1, Springer-Verlag, Berlin, 1992, pp. 125–163.
[31] A. Vanderbauwhede and S. A. van Gils, *Center manifolds and contractions on a scale of Banach spaces*, J. Functional Anal., 71 (1987), pp. 209–224.
[32] J. Wu, *Theory and Applications of Partial Functional Differential Equations*, Springer-Verlag, New York, 1996.

SIAM J. MATH. ANAL.
Vol. 36, No. 3, pp. 740–772

SUPER-BROWNIAN MOTION WITH EXTRA BIRTH AT ONE POINT*

KLAUS FLEISCHMANN† AND CARL MUELLER‡

Abstract. A super-Brownian motion in two and three dimensions is constructed where "particles" give birth at a higher rate, if they approach the origin. This contradicts the intuition suggested by the fact that in more than one dimension Brownian particles do not hit a given point. Via a log-Laplace approach, the construction is based on the work of Albeverio, Brzeźniak, and Dabrowski [*J. Funct. Anal.*, 130 (1995), pp. 220–254], who calculated the fundamental solutions of the heat equation with one-point potential in dimensions less than four.

Key words. super-Brownian motion, measure-valued process, heat equation, singular potential, one-point potential, $(1+\beta)$-branching, log-Laplace approach

AMS subject classifications. Primary, 60J80; Secondary, 60K35

DOI. 10.1137/S0036141002419473

1. Introduction.

1.1. Motivation and background. Measure-valued branching processes, also called superprocesses, arise naturally as limits of particle branching Markov processes. There is an immense literature on this topic; see any of the expositions [Daw93, Dyn94, LG99, Eth00, Per02], for example. Since these models involve mainly "noninteracting particles," many powerful tools are available, and many detailed properties of these processes are known. Building on this success, many probabilists have turned their attention to more complicated models, many of which are governed by singularities. For example, in two or more dimensions, with probability 1, continuous super-Brownian motion takes values in the space of measures whose closed support has Lebesgue measure 0. Nevertheless, in certain situations, pairs of such processes can kill each other when the corresponding "particles" meet (see, e.g., [EP94]).

Another example of singular behavior is catalytic branching. Here, the branching of the "particles" is controlled by a catalytic measure; the higher the "density" of this measure, the faster the "particles" branch or die. This catalytic measure can be supported on a set of Lebesgue measure 0, as long as it is not a polar set of Brownian motion. In other words, individual "particles" must have a positive probability of "hitting the measure." See, e.g., [DF95, Del96, FK99, Kle00].

A further example of singular behavior is mass creation. One could imagine a "mass creation measure," which would give rise to new "particles" whenever the "particles" of the superprocess hit the support of the measure. For the extreme case of a single point source δ_0 in R, see [EF00, ET02]. In particular, a continuous super-Brownian motion in R with a point source makes sense. In higher dimensions, however,

*Received by the editors December 10, 2002; accepted for publication (in revised form) February 27, 2004; published electronically September 24, 2004.
http://www.siam.org/journals/sima/36-3/41947.html

†Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstr. 39, D–10117 Berlin, Germany (fleischm@wias-berlin.de, http://www.wias-berlin.de/~fleischm). This author was supported in part by the research program "Interacting Systems of High Complexity" of the German Science Foundation.

‡Department of Mathematics, University of Rochester, Rochester, NY 14627 (cmlr@troi.cc.rochester.edu, http://www.math.rochester.edu/u/cmlr/). This author was supported in part by an NSA grant and the graduate program "Probabilistic Analysis and Stochastic Processes" of the German Science Foundation.

at first sight one would expect that a super-Brownian motion with single-point mass creation degenerates to ordinary super-Brownian motion, since the Brownian particles do not hit a given point.

Our goal is to disprove this intuition. But first, we mention a deterministic, "one-particle model," which already gives a different picture. This model was developed by mathematical physicists starting in the 1930s; see Albeverio et al. [AGHKH88] for historical background.

Consider the heat equation in R^d with a one-point potential:

$$\frac{\partial u}{\partial t} = \Delta u + \delta_0^{(\alpha)} u =: \Delta^{(\alpha)} u \tag{1.1}$$

(by $f := g$ or $g =: f$, we mean that f is defined to be equal to g). Heuristically, $\Delta^{(\alpha)} = \Delta + \delta_0^{(\alpha)}$ is the limit as $\varepsilon \downarrow 0$ of the operator

$$\Delta_\varepsilon^{(\alpha)} := \Delta + h(d, \alpha, \varepsilon)\varepsilon^{-d} \mathbf{1}_{B_\varepsilon(0)}, \tag{1.2}$$

where $B_\varepsilon(y)$ denotes the open ball around $y \in \mathsf{R}^d$ with radius ε, and $h(d, \alpha, \varepsilon)$ is some additional rescaling factor, depending on a parameter α, at least.

For instance, in dimension $d = 3$,

$$h(3, \alpha, \varepsilon) := \left(k + \frac{1}{2}\right)^2 \pi^2 \varepsilon - 8\pi^2 \alpha \varepsilon^2 - \zeta \varepsilon^3, \quad \alpha \in \mathsf{R},\ \varepsilon > 0, \tag{1.3}$$

where k is any integer and ζ any real number (we rely on [AGHKH88, (H.74)]). Then, in a sense, $\Delta_\varepsilon^{(\alpha)} \to \Delta^{(\alpha)}$ as $\varepsilon \downarrow 0$, where the limit operator $\Delta^{(\alpha)}$ is *independent* of k and ζ (so for simplification one could set $k = 0 = \zeta$).

Actually, in the physics literature, primary attention is paid to the Schrödinger equation, not the heat equation (and the positive definite operator $-\Delta$ is preferred instead of Δ). But note that in the time-stationary case, the Schrödinger and heat equations coincide. So the operators $\Delta^{(\alpha)}$ are relevant in both cases. Physically, the parameter α is related to the "scattering length" $(4\pi\alpha)^{-1}$ (see, for instance, [AGHKH88, p. 13]). In particular, $\Delta^{(\alpha)} \to \Delta$ as $\alpha \uparrow \infty$, giving the "free" case. We understand $\delta_0^{(\alpha)}$ as $\lambda_\alpha \delta_0$. In dimension $d = 3$ the coupling constant λ_α of the point source δ_0 has to be of the form $\lambda_\alpha = \varepsilon - \alpha\varepsilon^2$ with ε "infinitesimal" in a special way.

Even though the number of particles that hit the origin is infinitesimal, one can imagine that they give raise to a positive mass, provided that the birth rate is high enough. This explains why the linear ε-term in (1.3) is not allowed to be too small; in particular, it cannot be negative. In the latter case, particles will simply die, and nothing else will happen. But since there are only infinitesimally many particles hitting the origin, their possible deaths will pass unnoticed.

At this point we would like to understand certain questions from a probabilistic point of view. For instance, in dimensions $d = 3$, why don't all sufficiently high coefficients of the linear ε-term occur, and why is the limit operator $\Delta^{(\alpha)}$ independent of the integer k? Unfortunately, this is outside the scope of the present paper.

Strictly speaking, $\{\Delta^{(\alpha)} : \alpha \in \mathsf{R}\} \cup \{\Delta\}$ is the family of all self-adjoint extensions on $\mathcal{L}^2(\dot{\mathsf{R}}^d, \mathrm{d}x)$, $d = 2, 3$, of the Laplacian Δ acting on $\mathcal{C}_0^{(\infty)}(\dot{\mathsf{R}}^d)$, where $\dot{\mathsf{R}}^d := \mathsf{R}^d \backslash \{0\}$. See, for instance, [AGHKH88, Chapters I.1 and I.5]. We mention that in dimension one there is a 4-parameter family of extensions instead (see [ABD95]), whereas for $d \geq 4$ there is no other extension than the Laplacian Δ.

The *fundamental solutions* P^α to equation

$$\frac{\partial u}{\partial t} = \Delta^{(\alpha)} u \quad \text{on} \quad (0,\infty) \times \dot{\mathsf{R}}^d, \quad d = 2, 3, \tag{1.4}$$

have been computed in [ABD95]. P^α is different from the heat kernel for each $\alpha \in \mathsf{R}$ (only for $\alpha \uparrow \infty$ one gets back the heat kernel; for the case $d = 3$, see subsection 2.4 below). P^α is a basic object in the present paper.

1.2. Sketch of result. Based on the preceding analytical results from [ABD95], the *purpose* of the present paper is to construct a measure-valued super-Brownian motion $X = \{X_t : t \geq 0\}$ in $\dot{\mathsf{R}}^d = \mathsf{R}^d \backslash \{0\}$, $d = 2, 3$,[1] related to the formal log-Laplace equation

$$\begin{cases} \dfrac{\partial v}{\partial t} = \Delta^{(\alpha)} v - \eta v^{1+\beta} \quad \text{on} \quad (0,\infty) \times \dot{\mathsf{R}}^d, \\ \qquad v(0+, x) = \varphi(x) \geq 0, \quad x \in \dot{\mathsf{R}}^d, \end{cases} \tag{1.5}$$

with constants $0 < \beta \leq 1$, $\eta \geq 0$, and where the φ are appropriate test functions. Of course, X is related to (1.5) via the log-Laplace transition functional

$$-\log \mathbf{P}\{\mathrm{e}^{-\langle X_t, \varphi \rangle} \mid X_0\} = \langle X_0, v(t, \cdot) \rangle, \quad t > 0, \tag{1.6}$$

of the Markov process X.

Roughly speaking, we have "many" independent Brownian "particles" which everywhere undergo critical branching with index $1+\beta$ and rate η, but additionally give birth to new particles if they "approach" the origin 0.

Here is a rough formulation of our main result; a more precise statement will be given in Theorem 4.4 in subsection 4.3 below.

THEOREM 1.1 (existence of X). *If $d = 2$, then let $0 < \beta \leq 1$, and if $d = 3$, let $0 < \beta < 1$. Then, for each $\alpha \in \mathsf{R}$, there is a (unique in law) nondegenerate measure-valued (time-homogeneous) Markov process $X = X^\alpha$ having log-Laplace transition functional* (1.6) *with v solving* (1.5).

We call X a *super-Brownian motion in R^d with extra birth at point $x = 0$*. Note that in the case $\eta = 0$, the process degenerates to the deterministic mass flow related to the kernels P^α, the fundamental solutions to (1.4). At the same time, this mass flow is identical to the expectation of X for any η. In particular, $X = X^\alpha$ is different from ordinary super-Brownian motion (corresponding to $\alpha = \infty$).

Remark 1.2 (open problem). The condition $\beta < 1$ in the three-dimensional case, which excludes finite variance branching as in continuous super-Brownian motion, looks a bit strange. We need this condition for technical reasons, to handle some singularities at the point $x = 0$ where extra birth occurs (see Remark 2.8 below).

It would, of course, be interesting to reveal that this superprocess X has strange new properties. However, we leave this task for a future paper (see [FV04]) and present this construction result separately, since it seems to be interesting enough.

[1] Recall that the one-dimensional super-Brownian motion with extra birth at 0, that is, related to the log-Laplace equation

$$\frac{\partial v}{\partial t} = \frac{1}{2}\Delta v + \delta_0 - v^2 \quad \text{on} \quad (0,\infty) \times \mathsf{R},$$

was introduced in [EF00].

1.3. Outline. In section 2 we give some estimates involving the basic solutions P^α and the semigroup related to the linear equation (1.4). This semigroup is not Markovian in the usual sense, since the integral of the kernel P^α is greater than 1. Then, in section 3, we show that the log-Laplace equation (1.5) is well posed. Here we use Picard iteration, but for the nonnegativity of solutions we go back to a linearized equation. For the construction of X in section 4 we use a Trotter product formula, alternating between purely continuous-state branching (Feller's branching diffusion if $\beta = 1$) and deterministic mass flow with single-point mass creation (related to the kernels P^α).

For background from a mathematical physics point of view concerning the operators $\Delta^{(\alpha)}$ we recommend [AGHKH88], and for basic facts on superprocesses we refer to one of the systematic treatments [Daw93, Dyn94, LG99, Eth00, Per02], which we have already mentioned.

2. The heat equation with birth at a single point. After introducing the set Φ of test functions, on which the heat flow acts continuously (Lemma 2.4), we define the kernel P^α in subsection 2.4 and show the strong continuity of the related flow S^α on Φ (Corollary 2.10).

2.1. Preliminaries: Test functions and measures. The letter C denotes a constant which might change its value from occurrence to occurrence. $C_\#$ and $C_{(\#)}$ refer to specific constants which are defined around Lemma #, say, or formula (#), respectively.

Let ϕ denote the *weight and reference function*

$$\phi(x) := |x|^{-(d-1)/2}, \quad x \in \dot{\mathsf{R}}^d = \mathsf{R}^d \backslash \{0\}. \tag{2.1}$$

For each fixed constant $\varrho \geq 1$, we introduce the Lebesgue space $\mathcal{H} = \mathcal{H}^\varrho = \mathcal{L}^\varrho(\dot{\mathsf{R}}^d, \phi(x)\mathrm{d}x)$ of equivalence classes φ of measurable functions on $\dot{\mathsf{R}}^d$ for which $\|\varphi\|_{\mathcal{H}} < \infty$, where[2]

$$\|\varphi\|_{\mathcal{H}} := \left(\int_{\mathsf{R}^d} \mathrm{d}x \phi(x) |\varphi|^\varrho(x) \right)^{1/\varrho}. \tag{2.2}$$

(As usual, we do not distinguish between an equivalence class and its representatives.)

For fixed $\varrho \geq 1$, let $\Phi = \Phi^\varrho$ denote the set of all *continuous* functions $\varphi : \dot{\mathsf{R}}^d \to \mathsf{R}$ such that $\varphi \in \mathcal{H}$ and

$$0 \leq \varphi \leq C_{(2.3)}\phi \quad \text{for some constant } C_{(2.3)} = C_{(2.3)}(\varphi). \tag{2.3}$$

We endow Φ with the topology inherited from $\mathcal{H}$. Note that the set $\mathcal{C}^+_{\mathrm{com}} = \mathcal{C}^+_{\mathrm{com}}(\dot{\mathsf{R}}^d)$ of all nonnegative continuous functions on $\dot{\mathsf{R}}^d$ with compact support is contained in Φ. Note also that $\varphi \in \Phi$ might have a singularity at $x = 0$ of order $|x|^{-\xi}$ with $0 < \xi < \frac{d-1}{2} \wedge \frac{d+1}{2\varrho}$. (Later, in Hypothesis 3.2, we will restrict ϱ to be less than $\frac{d+1}{d-1}$; then at least the singularity orders $\xi < \frac{d-1}{2}$ are allowed.) The functions in Φ will serve as test functions in log-Laplace representations.

Let $\mathcal{M} = \mathcal{M}(\dot{\mathsf{R}}^d)$ denote the set of all measures μ defined on $\dot{\mathsf{R}}^d$ such that $\langle \mu, \varphi \rangle := \int_{\mathsf{R}^d} \mu(\mathrm{d}x)\varphi(x) < \infty$ for all $\varphi \in \Phi$. We equip $\mathcal{M}$ with the vague topology (recall that

[2]Here and in similar cases we use this simplified integration domain R^d, since in the case of integration with respect to Lebesgue measure, including or not including the point $0 \in \mathsf{R}^d$ of singularity makes no difference.

$\mathcal{C}^+_{\rm com} \subset \Phi$). Of course, each measure $\mu \in \mathcal{M}$ can also be considered as a measure on R^d with zero mass at $0 \in \mathsf{R}^d$. But in our pairing $\langle \mu, \varphi \rangle$, $\varphi \in \Phi$, we cannot extend to work with measures μ on R^d allowing positive mass at 0 by the mentioned possible singularities of the $\varphi \in \Phi$.

If μ is a finite measure, we write $\|\mu\|$ for its total mass. The symbol ℓ denotes the Lebesgue measure, $A^{\rm c}$ the complement of A, and $a \vee b$ the maximum of a and b.

2.2. Heat flow estimates on $\mathcal{H}$. In this subsection we fix a dimension $d \geq 1$. Let $P = P(t; x, y)$ refer to the fundamental solution of the heat equation

$$\frac{\partial u}{\partial t} = \Delta u \quad \text{on} \quad (0, \infty) \times \mathsf{R}^d. \tag{2.4}$$

In other words,

$$P(t; x, y) = (4\pi t)^{-d/2} \mathrm{e}^{-|y-x|^2/4t}, \quad t > 0,\ x, y \in \mathsf{R}^d. \tag{2.5}$$

Let $S = \{S_t : t \geq 0\}$ denote the semigroup corresponding to this heat kernel P.

Here is our first estimate (with ϕ the weight and reference function introduced in (2.1)).

LEMMA 2.1 (a heat flow estimate). *There is a constant $C_{2.1} = C_{2.1}(d)$ such that*

$$S_t \phi \leq C_{2.1} \phi, \quad t \geq 0. \tag{2.6}$$

Proof. Without loss of generality, let $t > 0$ and $x \neq 0$. We have to show that

$$\frac{1}{\phi(x)} S_t \phi(x) = \frac{1}{\phi(x)} \int_{\mathsf{R}^d} \mathrm{d}y \phi(y) \frac{1}{(4\pi t)^{d/2}} \mathrm{e}^{-|y-x|^2/4t} \tag{2.7}$$

is bounded in $t > 0$ and $x \neq 0$. By the change of variables $w := t^{-1/2}(y - x)$, and with the notation $z := -t^{-1/2} x$, we get

$$\frac{1}{\phi(x)} S_t \phi(x) = C \int_{\mathsf{R}^d} \mathrm{d}w \phi((w - z)/z) \mathrm{e}^{-|w|^2/4}. \tag{2.8}$$

We have to show that the right-hand side is bounded in $z \neq 0$. If we restrict the integration to $|w| \leq |z|/2$, then $|w - z| \geq |z|/2$, implying $\phi((w - z)/z) \leq 2^{(d-1)/2}$, and the whole restricted integral is bounded by a constant. On the other hand, if we restrict the integration to $|w| > |z|/2$, the exponential expression can be estimated from above by $\mathrm{e}^{-|z|^2/32} \mathrm{e}^{-|w|^2/8}$, and for the restricted integral we get the upper bound

$$C \int_{\mathsf{R}^d} \mathrm{d}w \phi(w - z) \mathrm{e}^{-|w|^2/8}. \tag{2.9}$$

But this integral is bounded in z. To see this, distinguish between $|w - z| \leq 1$ and $|w - z| > 1$. This completes the proof. □

We finish this subsection with a simple maximization result.

LEMMA 2.2 (maximum in the center). *Fix a constant $\varkappa > 0$. Then*

$$S_t \phi^\varkappa(x) \leq S_t \phi^\varkappa(0), \quad t > 0,\ x \in \mathsf{R}^d. \tag{2.10}$$

Proof. We will use the fact that in the integral

$$\int_{\mathsf{R}^d} \mathrm{d}y P(t; x, y) \phi^\varkappa(y) \tag{2.11}$$

the mapping $y \mapsto \phi^{\varkappa}(y)$ is radially symmetric and decreasing in $|y|$. The same is true for $y \mapsto P(t;x,y)$, except for a shift by x.

Step 1 (*simplification*). Let $a,b,c,d \geq 0$. Then, by expanding,

$$(a+b)(c+d)+ac \geq (a+b)c+a(c+d). \tag{2.12}$$

Step 2 (*functions with n steps*). For $n \geq 2$, let

$$f_i := \sum_{j=1}^{n} a_{i,j} 1_{B_j} \geq 0, \quad i=1,2, \tag{2.13}$$

be two step functions defined on $n \geq 2$ cubes $B_1,\dots,B_n$ in R^d of equal volume, say v. For $i=1,2$, let $\bar{f}_i$ be constructed from f_i by rearranging the $a_{i,j}$ to $\bar{a}_{i,1} \geq \cdots \geq \bar{a}_{i,n}$. Then

$$\int_{\mathsf{R}^d} \mathrm{d}x \bar{f}_1(x)\bar{f}_2(x) \geq \int_{\mathsf{R}^d} \mathrm{d}x f_1(x)f_2(x). \tag{2.14}$$

In fact,

$$\int_{\mathsf{R}^d} \mathrm{d}x f_1(x)f_2(x) = \mathrm{v}\sum_{j=1}^{n} a_{1,j}a_{2,j}. \tag{2.15}$$

Rearranging if necessary, we may assume that $f_1 = \bar{f}_1$, that is, $a_{1,j} = \bar{a}_{1,j}$, $1 \leq j \leq n$. Exploiting Step 1, we may switch from f_2 to $\bar{f}_2$ by a sequence of rearrangements which never decrease the integral in (2.15). This then gives the claim (2.14).

Step 3 (*approximation*). We may assume that the right-hand side of (2.10) is finite. Then the "integrals" in (2.10) (recall (2.11)) can be approximated by using step functions as in (2.13). Then (2.10) follows from (2.14) by passing to the limit. □

2.3. Strong continuity of the heat flow on $\mathcal{H}$. Next we will prove the following statement.

LEMMA 2.3 (estimate of S in case of an additional singularity). *Let $d \geq 1$, $0 \leq \beta \leq 1$, and assume that ϱ in* (2.2) *satisfies*

$$\varrho > \frac{1}{1-\beta(d-1)/2d}. \tag{2.16}$$

Then there is a constant $C_{2.3} = C_{2.3}(d,\beta,\varrho)$ such that for all $\varphi \in \mathcal{H} = \mathcal{H}^{\varrho}$,

$$\|S_t(\varphi\phi^{\beta})\|_{\mathcal{H}}^{\varrho} \leq C_{2.3}t^{-\beta\varrho(d-1)/4}\|\varphi\|_{\mathcal{H}}^{\varrho}, \quad t>0. \tag{2.17}$$

Proof. Fix d, β, ϱ as in the lemma. For $t>0$ and $x \in \mathsf{R}^3$, we introduce the measures

$$\mu_{t,x}(\mathrm{d}y) := t^{\kappa}P(t;x,y)\phi^{\lambda}(y)\mathrm{d}y, \tag{2.18}$$

with

$$\kappa := \frac{\beta\varrho(d-1)}{4(\varrho-1)} \quad \text{and} \quad \lambda := \frac{\beta\varrho}{\varrho-1}. \tag{2.19}$$

By Lemma 2.2,

$$\|\mu_{t,x}\| \leq \|\mu_{t,0}\| = \int_{\mathsf{R}^d} \mathrm{d}y P(1;0,y)\phi^{\lambda}(y) =: C_{(2.20)}, \tag{2.20}$$

where in the last step we used Brownian scaling and the identity $\kappa - \lambda(d-1)/4 = 0$. Note that $C_{(2.20)} = C_{(2.20)}(d, \beta, \varrho)$ is finite by our assumption (2.16). Therefore, the measures $\mu_{t,x}$ are finite with total mass at most $C_{(2.20)}$ independent of t and x. Now, for each finite measure μ on R^d, and for measurable φ, by Hölder's inequality,

$$\left[\int_{\mathsf{R}^d} |\varphi|(y)\mu(\mathrm{d}y)\right]^{\varrho} \le \|\mu\|^{\varrho-1} \int_{\mathsf{R}^d} \mu(\mathrm{d}y)|\varphi|^{\varrho}(y). \tag{2.21}$$

Applied to the measures $\mu_{t,x}$ we get

$$\begin{aligned} |S_t(\varphi\phi^\beta)(x)|^\varrho &= t^{-\kappa\varrho}\left|\int_{\mathsf{R}^d} \mu_{t,x}(\mathrm{d}y)\phi^{\beta-\lambda}(y)\varphi(y)\right|^\varrho \\ &\le t^{-\kappa\varrho}\|\mu_{t,x}\|^{\varrho-1}\int_{\mathsf{R}^d} \mu_{t,x}(\mathrm{d}y)\phi^{(\beta-\lambda)\varrho}(y)|\varphi|^\varrho(y) \\ &\le t^{-\kappa(\varrho-1)}C_{(2.20)}^{\varrho-1} S_t|\varphi|^\varrho(x), \end{aligned} \tag{2.22}$$

since $\lambda + (\beta - \lambda)\varrho = 0$ by (2.19). But by Lemma 2.1,

$$\begin{aligned} \int_{\mathsf{R}^d} \mathrm{d}x\phi(x)S_t|\varphi|^\varrho(x) &= \int_{\mathsf{R}^d} \mathrm{d}y|\varphi|^\varrho(y)S_t\phi(y) \\ &\le \int_{\mathsf{R}^d} \mathrm{d}y|\varphi|^\varrho(y)C_{2.1}\phi(y) = C_{2.1}\|\varphi\|_{\mathcal{H}}^\varrho. \end{aligned} \tag{2.23}$$

Hence,

$$\|S_t(\varphi\phi^\beta)\|_{\mathcal{H}}^\varrho \le t^{-\kappa(\varrho-1)}C_{(2.20)}^{\varrho-1}C_{2.1}\|\varphi\|_{\mathcal{H}}^\varrho, \tag{2.24}$$

and the claim follows since $\kappa(\varrho - 1) = \beta\varrho(d-1)/4$. □

Lemma 2.3 with $\beta = 0$ yields the following result.

LEMMA 2.4 (strong continuity of the heat flow on $\mathcal{H}$). *The semigroup S acting on $\mathcal{H} = \mathcal{H}^\varrho$ is strongly continuous.*

Proof. Fix $\varphi \in \mathcal{H}$. By linearity, we may assume that $\varphi \ge 0$. Consider $t \in (0, 1]$.

Step 1 (*reducing to bounded functions on compact sets*). Fix $\varepsilon \in (0, 1]$. We choose a compact set $K \subset \mathsf{R}^d$ so large that $\|\varphi 1_{K^c}\|_{\mathcal{H}} < \varepsilon$, and then a number $N \ge 1$ such that $\|\varphi 1_K 1_{\{\varphi > N\}}\|_{\mathcal{H}} < \varepsilon$. Then

$$\begin{aligned} \|S_t\varphi - \varphi\|_{\mathcal{H}} &\le \|S_t(\varphi 1_{K^c}) - \varphi 1_{K^c}\|_{\mathcal{H}} + \|S_t(\varphi 1_K 1_{\{\varphi>N\}}) - \varphi 1_K 1_{\{\varphi>N\}}\|_{\mathcal{H}} \\ &\quad + \|S_t(\varphi 1_K 1_{\{\varphi\le N\}}) - \varphi 1_K 1_{\{\varphi\le N\}}\|_{\mathcal{H}} \\ &\le C\varepsilon + \|S_t(\varphi 1_K 1_{\{\varphi\le N\}}) - \varphi 1_K 1_{\{\varphi\le N\}}\|_{\mathcal{H}}, \end{aligned} \tag{2.25}$$

where in the last step we used twice Lemma 2.3 with $\beta = 0$. Thus, for the rest of the proof we may assume that φ is bounded by $N \ge 1$ and vanishes outside a compact set $K \subset \mathsf{R}^d$. That is, from now on in this proof we assume that R^d is replaced by K in the definition of $\mathcal{H}$.

Step 2 (*passing to a continuous function*). Fix $\varepsilon \in (0, 1]$. Choose a continuous nonnegative function $f_\varepsilon \le N$ (on K) such that $\varphi = f_\varepsilon$ on a measurable set $A_\varepsilon \subseteq K$ satisfying $\ell(A_\varepsilon^c) \le \varepsilon$. Then, again by twice applying Lemma 2.3 with $\beta = 0$,

$$\begin{aligned} \|S_t\varphi - \varphi\|_{\mathcal{H}} &\le \|S_t(\varphi 1_{A_\varepsilon^c}) - \varphi 1_{A_\varepsilon^c}\|_{\mathcal{H}} + \|S_t(f_\varepsilon 1_{A_\varepsilon}) - f_\varepsilon 1_{A_\varepsilon}\|_{\mathcal{H}} \\ &\le C\|\varphi 1_{A_\varepsilon^c}\|_{\mathcal{H}} + C\|f_\varepsilon 1_{A_\varepsilon^c}\|_{\mathcal{H}} + \|S_t f_\varepsilon - f_\varepsilon\|_{\mathcal{H}}. \end{aligned} \tag{2.26}$$

For $x \in K$ fixed, $S_t f_\varepsilon(x) \to f_\varepsilon(x)$ as $t \downarrow 0$,

$$\sup_{t \geq 0} \|S_t f_\varepsilon\|_\infty \leq \|f_\varepsilon\|_\infty < \infty \tag{2.27}$$

(with $\|\cdot\|_\infty$ denoting the supremum norm), and ϕ is integrable on K. Hence, by dominated convergence, the third term in (2.26) will vanish as $t \downarrow 0$ for fixed ε. On the other hand, $\|\varphi 1_{A_\varepsilon^c}\|_{\mathcal{H}}$ converges to 0 as $\varepsilon \downarrow 0$. Finally, the same is true for $\|f_\varepsilon 1_{A_\varepsilon^c}\|_{\mathcal{H}}$ since $f_\varepsilon \leq N$. This completes the proof. □

2.4. The fundamental solutions P . Fix $\alpha \in \mathsf{R}$. We now introduce the fundamental solutions $P^\alpha = P^\alpha(t; x, y)$ of the heat equation with one-point potential $\delta_0^{(\alpha)}$, that is, of (1.4).

Step 1 ($d = 3$). Based on [ABD95, formula array (3.4)], for $d = 3$, we can *define*

$$\begin{aligned} P^\alpha(t; x, y) &:= P(t; x, y) + \frac{2t}{|x||y|} P(t; |x| + |y|) \\ &\quad - \frac{8\pi\alpha t}{|x||y|} \int_0^\infty \mathrm{d}u \mathrm{e}^{-4\pi\alpha u} P(t; u + |x| + |y|), \end{aligned} \tag{2.28}$$

$t > 0$, $x, y \neq 0$, where with a slight abuse of notation for heat kernel P,

$$P(t; r) := (4\pi t)^{-d/2} \exp(-r^2/4t), \quad t, r > 0. \tag{2.29}$$

Note that the term in (2.28) involving the integral is always finite and that it disappears for $\alpha = 0$. Otherwise, using the substitution $|\alpha|u \to u$ (for $\alpha \neq 0$) one realizes that $P^\alpha(t; x, y)$ is continuous and decreasing in α, and that $P^\alpha \downarrow P$, the heat kernel, pointwise as $\alpha \uparrow \infty$, whereas $P^\alpha \uparrow \infty$ pointwise as $\alpha \downarrow -\infty$.

Step 2 ($d = 2$). On the other hand, by [ABD95, formula (3.15)], for $d = 2$, we may *define*

$$\begin{aligned} P^\alpha(t; x, y) &:= P(t; x, y) + \frac{\sqrt{4\pi t}}{\sqrt{|x||y|}} P(t; |x| + |y|) \\ &\quad \times \int_0^\infty \mathrm{d}u \frac{t^u \mathrm{e}^{-\alpha u}}{\Gamma(u)} \int_0^\infty \mathrm{d}r \frac{r^{u-1} \mathrm{e}^{-(|x|+|y|)^2/4tr}}{(r+1)^{u+1/2}} \tilde{K}_0 \left(\frac{|x||y|}{2t}(r+1) \right), \end{aligned} \tag{2.30}$$

$t > 0$, $x, y \neq 0$, where Γ is the Gamma function,

$$\Gamma(u) := \int_0^\infty \mathrm{d}s s^{u-1} \mathrm{e}^{-s}, \quad u > 0, \tag{2.31}$$

and

$$\tilde{K}_0(z) := \mathrm{e}^z (2z/\pi)^{1/2} K_0(z), \quad z \geq 0, \tag{2.32}$$

with $K_0 \geq 0$ the Macdonald function of order 0. In other words, K_0 is the modified Bessel function of the third kind, of order 0. See [Leb65, p. 109].

Recall that P^α ($d = 2, 3$) is the family of fundamental solutions to (1.4), computed in [ABD95]. Since $\Delta^{(\alpha)}$ is a self-adjoint extension of Δ on $\dot{\mathsf{R}}^d$, the kernel P^α solves the heat equation on $(0, \infty) \times \dot{\mathsf{R}}^d$, as shown in the following corollary.

Corollary 2.5 (solutions of the heat equation). *Let $d = 2, 3$ and $\alpha \in \mathsf{R}$. Then*

$$\frac{\partial}{\partial t} P^\alpha(t; x, y) = \Delta P^\alpha(t; x, y) \quad \textit{on} \quad (0, \infty) \times \dot{\mathsf{R}}^d, \tag{2.33}$$

where the Laplacian acts on x (or y, respectively). In particular, $(t,x,y) \mapsto P^\alpha(t;x,y)$ is jointly continuous on $(0,\infty) \times \dot{\mathsf{R}}^d$.

Let $S^\alpha = \{S^\alpha_t : t \geq 0\}$ denote the semigroup corresponding to the kernel P^α, $\alpha \in \mathsf{R}$. As we noted in subsection 1.3, since $\int_{\mathsf{R}^d} \mathrm{d}y P^\alpha(t;x,y) > 1$, this semigroup is not Markovian in the usual sense.

2.5. Bounds on P . In this subsection we will derive some bounds for the kernels P^α introduced in (2.28) and (2.30), respectively. To this end, we set

$$\bar{P}(t;x,y) := t^{-1/2}\phi(x)\phi(y)\mathrm{e}^{-|x|^2/4t}\mathrm{e}^{-|y|^2/4t} \tag{2.34}$$

for $t > 0$ and $x, y \neq 0$ (recall the weight and reference function ϕ from (2.1)).

LEMMA 2.6 (P^α bound). *Let $d = 2, 3$. For each $\alpha \in \mathsf{R}$ and $T > 0$, there is a constant $C_{2.6} = C_{2.6}(d,\alpha,T)$ such that*

$$P(t;x,y) \leq P^\alpha(t;x,y) \leq P(t;x,y) + C_{2.6}\bar{P}(t;x,y) \tag{2.35}$$

for all $t \in (0,T]$ and $x, y \neq 0$.

Proof.

Step 1 ($d = 3$). By the arguments after (2.29), for $\alpha \geq 0$,

$$P \leq P^\alpha \leq P^0 \leq P + 2(4\pi)^{-3/2}\bar{P}, \tag{2.36}$$

since

$$(|x| + |y|)^2 \geq |x|^2 + |y|^2. \tag{2.37}$$

So we will restrict our attention to $\alpha < 0$. Abbreviating

$$-4\pi\alpha =: \frac{r}{2} > 0 \quad \text{and} \quad |x| + |y| =: R \geq 0 \tag{2.38}$$

and using the last inequality in (2.36) and (2.37), it suffices to verify that

$$\int_0^\infty \mathrm{d}u \frac{r}{2}\,\mathrm{e}^{\frac{r}{2}u} P(t;u+R) \leq C_{(2.39)} P(t;R) \tag{2.39}$$

(recall notation (2.29)), with a positive constant $C_{(2.39)} = C_{(2.39)}(T,r)$ independent of t and R. Fix any

$$u_0 > rT \quad \text{and put} \quad u_1 := u_0 - rT > 0. \tag{2.40}$$

Consider first the integral in (2.39) restricted to $u \in [0,u_0]$. Here we can use $P(t;u+R) \leq P(t;R)$ and the fact that

$$\int_0^{u_0} \mathrm{d}u \frac{r}{2}\,\mathrm{e}^{\frac{r}{2}u} \leq \mathrm{e}^{\frac{r}{2}u_0}, \tag{2.41}$$

resulting in a positive constant independent of t and R. It remains to deal with

$$\int_{u_0}^\infty \mathrm{d}u \frac{r}{2}\,\mathrm{e}^{\frac{r}{2}u}(4\pi t)^{-3/2}\,\mathrm{e}^{-(u+R)^2/4t} \tag{2.42a}$$

$$= \frac{r}{2}(4\pi t)^{-3/2}\,\mathrm{e}^{-(2rRt - r^2t^2)/4t} \int_{u_0}^\infty \mathrm{d}u \mathrm{e}^{-(u+R-rt)^2/4t} \tag{2.42b}$$

for $0 < t \le T$. The exponential factor in front of the integral in (2.42b) is bounded by

$$\mathrm{e}^{r^2T/4} =: C_{(2.43)}, \tag{2.43}$$

which is a positive constant independent of t and R. Substituting $u + R - rt \to u$ and recalling notation (2.40), the integral in (2.42b) can be bounded by

$$\int_{R+u_1}^{\infty} \mathrm{d}u \mathrm{e}^{-u^2/4t} \le \frac{2T}{u_1} \int_R^{\infty} \mathrm{d}u \frac{u}{2t} \mathrm{e}^{-u^2/4t} = \frac{2T}{u_1} \mathrm{e}^{-R^2/4t}. \tag{2.44}$$

Thus for the integral in (2.42a) we found the bound

$$C_{(2.43)} \frac{r}{2} \frac{2T}{u_1} P(t; R), \tag{2.45}$$

which finishes the proof in the case $d = 3$.

Step 2 ($d = 2$). Recall definition (2.32) of $\tilde{K}_0$. According to [ABD95, after (3.14)],

$$\lim_{z \to \infty} \tilde{K}_0(z) = 1. \tag{2.46}$$

Consulting [Tra69, section 1.15, equation (1.66)], we find that

$$K_0(z) \sim -\gamma - \log(z/2) \sim -\log z \quad \text{as } z \downarrow 0, \tag{2.47}$$

where γ is Euler's constant. Therefore,

$$\lim_{z \downarrow 0} \tilde{K}_0(z) = \lim_{z \downarrow 0} [\mathrm{e}^z (2z/\pi)^{1/2} \log z] = 0. \tag{2.48}$$

Since $\tilde{K}_0$ is continuous, relations (2.46) and (2.48) together give

$$\|\tilde{K}_0\|_\infty < \infty. \tag{2.49}$$

Fix $\alpha \in \mathsf{R}$ and consider $0 < t \le T$. We may assume that $T \ge 1$. We start by estimating the inner integral appearing on the right-hand side of definition (2.30) of P^α. For $u > 0$,

$$\begin{aligned} &\int_0^{\infty} \mathrm{d}r \frac{r^{u-1} \mathrm{e}^{-(|x|+|y|)^2/4tr}}{(r+1)^{u+1/2}} \tilde{K}_0 \left(\frac{|x||y|}{2t} (r+1) \right) \\ &\qquad \le \|\tilde{K}_0\|_\infty \int_0^{\infty} \mathrm{d}r \frac{r^{u-1}}{(r+1)^{u+1/2}}. \end{aligned} \tag{2.50}$$

If $r \ge 1$, drop the 1 in the denominator; otherwise drop the r there. Thus, for the inner integral in (2.30) we find the bound

$$\|\tilde{K}_0\|_\infty [2 + 1/u]. \tag{2.51}$$

Using this bound, we turn to the outer integral of (2.30). For the Gamma function Γ of (2.31), Stirling's formula gives

$$\Gamma(u) \sim \sqrt{2\pi} (u-1)^{u-1/2} \mathrm{e}^{-u+1} \quad \text{as } u \uparrow \infty. \tag{2.52}$$

It follows that, for some constant $C_{(2.53)} = C_{(2.53)}(T, \alpha)$,

$$\int_1^\infty \mathrm{d}u \frac{t^u \mathrm{e}^{-\alpha u}}{\Gamma(u)} \le C_{(2.53)}, \quad 0 \le t \le T. \tag{2.53}$$

Next, using integration by parts, we estimate $u\Gamma(u)$ for $u \in (0, 1]$:

$$u\Gamma(u) = \int_0^\infty \mathrm{d}s u s^{u-1} \mathrm{e}^{-s} = \int_0^\infty \mathrm{d}s s^u \mathrm{e}^{-s} \ge \mathrm{e}^{-1} \int_0^1 \mathrm{d}s s =: C_{(2.54)}. \tag{2.54}$$

Finally, for some constant $C_{(2.55)} = C_{(2.55)}(T, \alpha)$, since $T \ge 1$,

$$\int_0^1 \mathrm{d}u \frac{t^u \mathrm{e}^{-\alpha u}}{u\Gamma(u)} \le \frac{1}{C_{(2.54)}} T \mathrm{e}^{|\alpha|} =: C_{(2.55)}. \tag{2.55}$$

Altogether, we found that the double integral appearing on the right-hand side of definition (2.30) of P^α is bounded by a constant depending only on α, T. This gives estimate (2.35) also in the case $d = 2$, since $\tilde{K}_0 \ge 0$, finishing the proof of Lemma 2.6. □

2.6. Strong continuity of S . We abbreviate

$$\bar{S}_t\varphi(x) := \int_{\mathsf{R}^d} \mathrm{d}y \varphi(y) \bar{P}(t; x, y), \quad t > 0, \ x \ne 0, \tag{2.56}$$

with $\bar{P}$ from (2.34), as long as the right-hand side expression makes sense. The estimates (2.35) and Minkowski's inequality then imply that

$$\|S_t\varphi\|_{\mathcal{H}} \le \|S^\alpha_t\varphi\|_{\mathcal{H}} \le \|S_t\varphi\|_{\mathcal{H}} + C_{2.6}\|\bar{S}_t\varphi\|_{\mathcal{H}}, \quad 0 < t \le T, \tag{2.57}$$

for those φ for which the right-hand side of (2.57) is meaningful and finite.

LEMMA 2.7 (estimate of $\bar{S}$ in case of an additional singularity). *Let $d = 2, 3$ as well as $0 \le \beta \le 1$, and assume*

$$\frac{1}{1 - \beta(d-1)/(d+1)} < \varrho < \frac{d+1}{d-1}. \tag{2.58}$$

Then there is a constant $C_{2.7} = C_{2.7}(d, \beta, \varrho)$ such that for all $\varphi \in \mathcal{H} = \mathcal{H}^\varrho$,

$$\|\bar{S}_t(\varphi\phi^\beta)\|^\varrho_{\mathcal{H}} \le C_{2.7}\varepsilon(t, \varphi) t^{-\beta\varrho(d-1)/4} \|\varphi\|^\varrho_{\mathcal{H}}, \quad t > 0, \tag{2.59}$$

where $0 \le \varepsilon(t, \varphi) \le 1$ and $\varepsilon(t, \varphi) \to 0$ as $t \downarrow 0$.

Remark 2.8 (restriction to infinite variance branching if $d = 3$). Note that in dimension $d = 3$ condition (2.58) can only be satisfied for some ϱ if $\beta < 1$ holds.

Proof of Lemma 2.7. This time we work with the measures

$$\mu_t(\mathrm{d}y) := t^{-\kappa} \mathrm{e}^{-|y|^2/4t} \phi^\lambda(y) \, \mathrm{d}y, \quad t > 0, \tag{2.60}$$

on R^d, where

$$\kappa := \frac{d+1}{4} - \frac{(d-1)\beta\varrho}{4(\varrho - 1)} > 0 \quad \text{and} \quad \lambda := 1 + \beta\varrho/(\varrho - 1). \tag{2.61}$$

Note that the measures μ_t have a t-independent total mass

$$\|\mu_t\| = \int_{\mathsf{R}^d} \mathrm{d}y \mathrm{e}^{-|y|^2/4}\phi^\lambda(y) =: C_{(2.62)} = C_{(2.62)}(d,\beta,\varrho), \tag{2.62}$$

which is finite by the left-hand inequality in assumption (2.58). Then, by our definition (2.34) of $\bar{P}$, for $t>0$ and $x\neq 0$,

$$|\bar{S}_t(\varphi\phi^\beta)(x)|^\varrho = t^{-\varrho/2+\kappa\varrho}\phi^\varrho(x)\mathrm{e}^{-\varrho|x|^2/4t}\left|\int_{\mathsf{R}^d}\mu_t(\mathrm{d}y)\phi^{-\lambda+\beta+1}(y)\varphi(y)\right|^\varrho. \tag{2.63}$$

By (2.21) and the definition of μ_t we may continue with

$$|\bar{S}_t(\varphi\phi^\beta)(x)|^\varrho \le t^{-\varrho/2+\kappa\varrho}\phi^\varrho(x)\mathrm{e}^{-\varrho|x|^2/4t}C_{(2.62)}^{\varrho-1}t^{-\kappa}\int_{\mathsf{R}^d}\mathrm{d}y\mathrm{e}^{-|y|^2/4t}\phi(y)|\varphi|^\varrho(y), \tag{2.64}$$

since $(-\lambda+\beta+1)\varrho+\lambda=1$. We may assume that $\varphi\neq 0$. Define

$$\varepsilon(t,\varphi) := \frac{1}{\|\varphi\|_{\mathcal{H}}^\varrho}\int_{\mathsf{R}^d}\mathrm{d}y\phi(y)\mathrm{e}^{-|y|^2/4t}|\varphi|^\varrho(y). \tag{2.65}$$

Note that $0<\varepsilon(t,\varphi)\le 1$ and that $\varepsilon(t,\varphi)\to 0$ as $t\downarrow 0$, by dominated convergence. Consequently,

$$|\bar{S}_t(\varphi\phi^\beta)(x)|^\varrho \le t^{-\varrho/2+\kappa\varrho}\phi^\varrho(x)\mathrm{e}^{-\varrho|x|^2/4t}C_{(2.62)}^{\varrho-1}t^{-\kappa}\varepsilon(t,\varphi)\|\varphi\|_{\mathcal{H}}^\varrho. \tag{2.66}$$

Therefore,

$$\|\bar{S}_t(\varphi\phi^\beta)\|_{\mathcal{H}}^\varrho \le C_{(2.62)}^{\varrho-1}\varepsilon(t,\varphi)t^{-\varrho/2+\kappa\varrho-\kappa}\|\varphi\|_{\mathcal{H}}^\varrho\int_{\mathsf{R}^d}\mathrm{d}x\phi^{\varrho+1}(x)\mathrm{e}^{-\varrho|x|^2/4t}.$$

But the latter integral is finite since $-(\varrho+1)(d-1)/2+d>0$ by the right-hand inequality in assumption (2.58). Moreover, using a change of variables, the integral gives an additional factor $t^{d/2-(\varrho+1)(d-1)/4}$, so that the whole t-term equals $t^{-\beta\varrho(d-1)/4}$. This finishes the proof. □

Since condition (2.58) is stronger than (2.16), combining Lemmas 2.3 and 2.7 with inequality (2.57) gives the following result.

COROLLARY 2.9 (estimate of S^α in case of an additional singularity). *Let $d=2,3$ as well as $0\le\beta\le 1$. Suppose* (2.58). *To each $T>0$ there is a constant $C_{2.9}=C_{2.9}(d,T,\alpha,\beta,\varrho)$ such that for all $\varphi\in\mathcal{H}=\mathcal{H}^\varrho$,*

$$\|S_t^\alpha(\varphi\phi^\beta)\|_{\mathcal{H}} \le C_{2.9}t^{-\beta(d-1)/4}\|\varphi\|_{\mathcal{H}}, \quad 0<t\le T. \tag{2.67}$$

In particular,

$$\sup_{t\le T}\|S_t^\alpha\varphi\|_{\mathcal{H}} \le C_{2.9}\|\varphi\|_{\mathcal{H}} < \infty, \quad \varphi\in\mathcal{H}. \tag{2.68}$$

Another consequence of Lemma 2.7 is shown in the following corollary.

COROLLARY 2.10 (strong continuity of S^α). *Let $d=2,3$. For each $\alpha\in\mathsf{R}$, the semigroup S^α acting on $\mathcal{H}=\mathcal{H}^\varrho$ with $\varrho\in(1,(d+1)/(d-1))$ is strongly continuous.*

Proof. Fix $\varphi\in\mathcal{H}$. By linearity, we may additionally assume that $\varphi\ge 0$. Consider $0<t\le T$. Decompose

$$S_t^\alpha\varphi = S_t\varphi + (S_t^\alpha - S_t)\varphi, \tag{2.69}$$

where by Lemma 2.6,

$$0 \le (S_t^\alpha - S_t)\varphi \le C_{2.6}\bar{S}_t\varphi, \tag{2.70}$$

implying

$$\|(S_t^\alpha - S_t)\varphi\|_{\mathcal{H}} \le C_{2.6}\|\bar{S}_t\varphi\|_{\mathcal{H}}. \tag{2.71}$$

From (2.69) and (2.71)

$$\begin{aligned}\|S_t^\alpha\varphi - \varphi\|_{\mathcal{H}} &\le \|S_t\varphi - \varphi\|_{\mathcal{H}} + \|(S_t^\alpha - S_t)\varphi\|_{\mathcal{H}} \\ &\le \|S_t\varphi - \varphi\|_{\mathcal{H}} + C_{2.6}\|\bar{S}_t\varphi\|_{\mathcal{H}}.\end{aligned} \tag{2.72}$$

But by Lemma 2.7 with $\beta = 0$, the second term in (2.72) goes to 0 as $t \downarrow 0$, whereas the first term does by Lemma 2.4. By (2.68), this finishes the proof. □

2.7. S as a flow on Φ. Recall our set Φ of continuous nonnegative test functions introduced in subsection 2.1. From the proof of Lemma 2.7 we also get the following result.

COROLLARY 2.11 (S^α bound). *Let $d = 2, 3$, assume $\varrho \in (1, (d+1)/(d-1))$, and assume $\varphi \in \mathcal{H}^\varrho$ satisfies*[3] (2.3). *Then, to each $T > 0$, there is a constant $C_{2.11} = C_{2.11}(d, T, \alpha, \varrho, \varphi)$ such that*

$$0 \le S_t^\alpha\varphi \le C_{2.11}(1 + t^{-1/2+(d+1)(\varrho-1)/4\varrho})\phi, \quad 0 < t \le T. \tag{2.73}$$

In particular, $S_t^\alpha\varphi \in \Phi$ for all $t > 0$.

Proof. From Lemma 2.6,

$$0 \le S_t^\alpha\varphi \le S_t\varphi + C_{2.6}\bar{S}_t\varphi. \tag{2.74}$$

Moreover, by assumption (2.3) on φ and by Lemma 2.1,

$$S_t\varphi \le C_{(2.3)}S_t\phi \le C\phi. \tag{2.75}$$

On the other hand, raising estimate (2.66) (with $\beta = 0$ there, implying $\kappa = (d+1)/4$ and $\lambda = 1$) into the power $1/\varrho$ gives

$$\bar{S}_t\varphi \le Ct^{-1/2+(d+1)(\varrho-1)/4\varrho}\phi\|\varphi\|_{\mathcal{H}}. \tag{2.76}$$

Putting together (2.74)–(2.76) yields (2.73). Finally, $(t, x) \mapsto S_t^\alpha\varphi$ is continuous on $(0, \infty) \times \dot{\mathsf{R}}^d$, since it solves the heat equation; recall Corollary 2.5. This finishes the proof. □

Combining Corollaries 2.10 and 2.11, we get the following result.

COROLLARY 2.12 (S^α acting on Φ). *Let $d = 2, 3$ and $\varrho \in (1, (d+1)/(d-1))$. Then S^α is a strongly continuous linear semigroup acting on $\Phi = \Phi^\varrho$.*

3. Analysis of the log-Laplace equation. The main result of this section is the well-posedness of the log-Laplace equation (Theorem 3.3). Uniqueness follows from a contraction argument (Lemma 3.7). Existence is shown via a Picard iteration (Lemmas 3.8, 3.9, and 3.11), whereas nonnegativity follows using a linearized equation (Lemma 3.10).

[3] Of course, an inequality on an element $\varphi \in \mathcal{H}$ means that the inequality holds for each representative in Lebesgue at almost every point.

3.1. Preliminaries and purpose. Formally, we can rewrite the log-Laplace equation (1.5) as the following *integral equation*:[4]

$$v(t,x) = S_t^\alpha \varphi(x) - \eta \int_0^t \mathrm{d}s S_{t-s}^\alpha (v^{1+\beta}(s))(x), \tag{3.1}$$

$t \geq 0$, $x \neq 0$ (with constants $\alpha \in \mathsf{R}$, $\eta \geq 0$, $0 < \beta \leq 1$, and where $\varphi \geq 0$ has yet to be specified). Here in writing $v^{1+\beta}$ we have in mind that $v \geq 0$. Note also that this nonnegativity implies the following *domination:*

$$0 \leq v(t) \leq S_t^\alpha \varphi, \quad t \geq 0. \tag{3.2}$$

The task of this section is to verify that the log-Laplace equation (3.1) is well posed in Φ.

DEFINITION 3.1 (Φ-valued solution). *Let $\varphi \in \Phi$. A measurable map $t \mapsto v(t) = V_t\varphi$ of R_+ into Φ is called a* solution *of* (3.1) *if* (3.1) *is true for all $x \neq 0$ and $t \geq 0$.*

For convenience, we introduce the following hypothesis.

Hypothesis 3.2 (choice of parameters). Let $\alpha \in \mathsf{R}$, $\eta \geq 0$, and

$$d = 2, 3, \quad 0 < \beta \leq 1, \quad \text{and} \quad \frac{1}{1 - \beta(d-1)/(d+1)} < \varrho < \frac{d+1}{d-1}. \tag{3.3}$$

Recall that for $d = 3$ this requires that $\beta < 1$.

Now we are ready to state the main result of this section.

THEOREM 3.3 (well-posedness of the log-Laplace equation). *If Hypothesis* 3.2 *holds, and if $\varphi \in \Phi$, then* (3.1) *has a unique Φ-valued solution $v = V\varphi = \{V_t\varphi : t \geq 0\}$. Moreover, $V = \{V_t : t \geq 0\}$ is a nonlinear strongly continuous semigroup acting on Φ.*

The rest of this section is devoted to the proof of this theorem.

3.2. First properties of solutions. Now we prepare for the uniqueness proof. Impose Hypothesis 3.2. Fix an integer $T > 0$ for a while, and $\varphi \in \Phi$. We will also fix measurable functions ψ_1, ψ_2 on $(0, T] \times \dot{\mathsf{R}}^d$ such that

$$0 \leq \psi_1(t,x) \leq M(1 + t^{-\kappa})\phi^\beta(x), \tag{3.4a}$$

$$0 \leq \psi_2(t,x) \leq S_t^\alpha \varphi(x), \tag{3.4b}$$

with constants $M = M(T, \psi_1) > 0$ and

$$\kappa := \beta/2 - \beta(d+1)(\varrho - 1)/4\varrho \in (0,1). \tag{3.5}$$

LEMMA 3.4 (properties of the nonlinear term). *There is a constant $C_{3.4} = C_{3.4}(d, M, T, \alpha, \beta, \varrho)$ such that*

$$\left\| \int_0^t \mathrm{d}s S_{t-s}^\alpha (\psi_1(s)\psi_2(s)) \right\|_{\mathcal{H}} \leq C_{3.4} \|\varphi\|_{\mathcal{H}} I(t), \quad 0 < t \leq T, \tag{3.6}$$

where

$$\infty > I(t) := \int_0^t \mathrm{d}s (1 + s^{-\kappa})(t-s)^{-\lambda} \underset{t \downarrow 0}{\searrow} 0, \tag{3.7}$$

[4] We often use notation as $v(s) := v(s, \cdot)$.

with

$$\lambda := \beta(d-1)/4. \tag{3.8}$$

Moreover, if for fixed $t \in (0,T]$,

$$N_t(x) := \int_0^t \mathrm{d}s S^\alpha_{t-s}(\psi_1(s)\psi_2(s))(x), \quad x \in \dot{\mathsf{R}}^d, \tag{3.9}$$

satisfies

$$N_t(x) \le S^\alpha_t \varphi(x), \quad x \in \dot{\mathsf{R}}^d, \tag{3.10}$$

then $N_t \in \Phi$.

Proof. First, by Corollary 2.10, we see that

$$\|S^\alpha_s \varphi\|_{\mathcal{H}} \le C\|\varphi\|_{\mathcal{H}}, \quad 0 \le s \le T, \tag{3.11}$$

where $C = C(T)$. Now, Corollary 2.9 states that

$$\|S^\alpha_t(\phi^\beta \varphi)\|_{\mathcal{H}} \le C_{2.9} t^{-\lambda}\|\varphi\|_{\mathcal{H}}, \quad 0 < t \le T. \tag{3.12}$$

Applying first (3.12) and then (3.11), we obtain

$$\|S^\alpha_{t-s}(\phi^\beta S^\alpha_s \varphi)\|_{\mathcal{H}} \le C_{2.9}(t-s)^{-\lambda}\|\varphi\|_{\mathcal{H}}, \quad 0 \le s < t \le T. \tag{3.13}$$

Exploiting assumption (3.4), we find

$$\|S^\alpha_{t-s}(\psi_1(s)\psi_2(s))\|_{\mathcal{H}} \le C(1+s^{-\kappa})(t-s)^{-\lambda}\|\varphi\|_{\mathcal{H}}. \tag{3.14}$$

However, $I(t)$ from (3.7) can be written as

$$I(t) = \frac{t^{1-\lambda}}{1-\lambda} + t^{1-\lambda-\kappa}\int_0^1 \mathrm{d}s s^{-\kappa}(1-s)^{-\lambda}. \tag{3.15}$$

The positive numbers κ and λ defined in (3.5) and (3.8), respectively, satisfy $\kappa+\lambda < 1$, hence (3.6) and (3.7) follow. Thus, the integrals in (3.9) are finite for almost all x. By assumption (3.10), it remains to show that $N_t(x)$ from (3.9) is continuous in x.

Let $\delta \in (0,t)$. Then

$$\int_0^{t-\delta} \mathrm{d}s S^\alpha_{t-s}(\psi_1(s)\psi_2(s))(x) \;=\; S^\alpha_\delta \int_0^{t-\delta} \mathrm{d}s S^\alpha_{t-\delta-s}(\psi_1(s)\psi_2(s))(x). \tag{3.16}$$

We already showed that the latter integral term belongs to $\mathcal{H}_+$. Then by Corollary 2.11, the left-hand side in (3.16) belongs to Φ, and hence is continuous in x for each δ. To complete the proof, it suffices to show that

$$\int_{t-\delta}^t \mathrm{d}s S^\alpha_{t-s}(\psi_1(s)\psi_2(s))(x) \underset{\delta \downarrow 0}{\longrightarrow} 0 \quad \text{uniformly in } x \in K, \tag{3.17}$$

where K is any compact subset of $\dot{\mathsf{R}}^d$ and is fixed from now on. Next apply assumption (3.4b) and Corollary 2.11 to $S^\alpha_s \varphi$ together with the definition (3.5) of κ to get

$$0 \le \psi_2(s) \le S^\alpha_s \varphi \le C_{2.11}(1+s^{-\kappa/\beta})\phi \le C\phi, \tag{3.18}$$

since s in (3.17) is bounded away from 0. Inserting the assumed upper bound (3.4a) on ψ_1, for the integral in (3.17) we find the estimate

$$C\int_{t-\delta}^{t} \mathrm{d}s S^{\alpha}_{t-s}\phi^{\beta+1}(x) = C\int_{0}^{\delta} \mathrm{d}s S^{\alpha}_{s}\phi^{\beta+1}(x). \tag{3.19}$$

Hence, it suffices to show that

$$s \mapsto \max_{x\in K} S^{\alpha}_{s}\phi^{\beta+1}(x) \quad \text{is integrable on } (0,\delta]. \tag{3.20}$$

By Lemma 2.6,

$$S^{\alpha}_{s}\phi^{\beta+1} \le S_s\phi^{\beta+1} + C_{2.6}\bar{S}_s\phi^{\beta+1}, \quad s>0. \tag{3.21}$$

Now, $(s,x)\mapsto S_s\phi^{\beta+1}(x)$ is finite and satisfies the heat equation on $[0,\delta]\times K$, implying

$$\sup_{(s,x)\in[0,\delta]\times K} S_s\phi^{\beta+1}(x) < \infty. \tag{3.22}$$

Turning to the second term in (3.21), by definition (2.34),

$$\bar{S}_s\phi^{\beta+1}(x) = \; s^{-1/2}\phi(x)\mathrm{e}^{-|x|/4s}\int_{\mathsf{R}^d} \mathrm{d}y\, \mathrm{e}^{-|y|/4s}\phi^{\beta+2}(y), \quad s>0. \tag{3.23}$$

By the substitution $y\mapsto y\sqrt{s}$, the latter integral gives an additional power contribution to $s^{-1/2}$. Moreover,

$$\sup_{x\in K} \phi(x)\mathrm{e}^{-|x|/4s} \le C\mathrm{e}^{-C/s}, \tag{3.24}$$

which together with $s^{-\lambda_0}$ is integrable on $(0,\delta]$ for each $\lambda_0\in\mathsf{R}$. This finishes the proof. □

LEMMA 3.5 (continuity at $t=0$). *Let $\varphi\in\Phi=\Phi^{\varrho}$ and let $v=V\varphi$ be a Φ-valued solution to* (3.1). *Under Hypothesis* 3.2, *for $T\ge 0$ fixed, there is a constant $C_{3.5}=C_{3.5}(d,T,\alpha,\varrho)$ such that*

$$\|V_t\varphi\|_{\mathcal{H}} \le C_{3.5}\|\varphi\|_{\mathcal{H}}, \quad 0\le t\le T. \tag{3.25}$$

Moreover, $V\varphi$ is strongly continuous at $t=0$, where $V_0\varphi=\varphi$.

Proof. By domination (3.2),

$$\|V_t\varphi\|_{\mathcal{H}} \le \|S^{\alpha}_t\varphi\|_{\mathcal{H}}. \tag{3.26}$$

Now (3.25) follows from (2.68). It remains to verify the continuity claim. Clearly, for $t\in(0,T]$,

$$|V_t\varphi-\varphi| \le |V_t\varphi - S^{\alpha}_t\varphi| + |S^{\alpha}_t\varphi-\varphi|. \tag{3.27}$$

By Corollary 2.10, it suffices to deal with the first term at the right-hand side. By (3.1), we have to look at

$$\left|\int_0^t \mathrm{d}s S^{\alpha}_{t-s}(v^{\beta}(s)v(s))\right|. \tag{3.28}$$

But from domination (3.2) and Corollary 2.11,

$$0\le v^{\beta}(s)\le C_{(3.29)}(1+s^{-\kappa})\phi^{\beta}, \quad 0<s\le T, \tag{3.29}$$

with κ from (3.5) and a constant $C_{(3.29)}=C_{(3.29)}(T,\varphi)$ (note that other dependencies are not important in the present proof). Thus, we can apply (3.6) and (3.7) from Lemma 3.4 to finish the proof. □

3.3. Uniqueness of solutions. The following lemma will be useful when we estimate the difference of solutions to (3.1).

LEMMA 3.6 (an elementary observation). *Let $\beta > 0$ and $a, b \in \mathsf{R}$. Then*

$$|a(a \vee 0)^\beta - b(b \vee 0)^\beta| \leq (1+\beta)(|a|+|b|)^\beta |a-b|. \tag{3.30}$$

Proof. First assume that $a, b \geq 0$. By the mean value theorem, there exists a number c between a and b such that

$$|a^{1+\beta} - b^{1+\beta}| = (1+\beta)c^\beta |a-b| \leq (1+\beta)(a+b)^\beta |a-b|. \tag{3.31}$$

This proves (3.30) for $a, b \geq 0$.

Now suppose that $a, b < 0$. In that case the left-hand side in (3.30) disappears, and hence (3.30) holds trivially.

Finally, it remains to consider the case $a < 0 \leq b$. Then

$$|a(a \vee 0)^\beta - b(b \vee 0)^\beta| = b^{1+\beta} \leq (1+\beta) b^\beta b \leq (1+\beta)(|a|+|b|)^\beta |a-b|,$$

and the proof is finished. □

We are ready to prove uniqueness for solutions to (3.1).

LEMMA 3.7 (uniqueness). *Impose Hypothesis* 3.2. *Fix $\varphi \in \Phi$. Suppose that u, v are Φ-valued solutions of* (3.1). *Then $u = v$.*

Proof. Fix $T > 0$. It suffices to prove uniqueness on $[0, T]$. Let

$$D(t,x) := u(t,x) - v(t,x), \quad 0 \leq t \leq T, \ x \neq 0. \tag{3.32}$$

Note that by Lemma 3.5,

$$\|D(t)\|_{\mathcal{H}} \leq 2C_{3.5} \|\varphi\|_{\mathcal{H}}, \quad 0 \leq t \leq T. \tag{3.33}$$

By the elementary inequality (3.30),

$$\begin{aligned} |D(t,x)| &= \eta \left| \int_0^t \mathrm{d}s S^\alpha_{t-s}(u^{1+\beta}(s) - v^{1+\beta}(s))(x) \right| \\ &\leq \eta \int_0^t \mathrm{d}s S^\alpha_{t-s} |u^{1+\beta}(s) - v^{1+\beta}(s)|(x) \\ &\leq 2\eta \int_0^t \mathrm{d}s S^\alpha_{t-s}([u^\beta(s) + v^\beta(s)]|D(s)|)(x). \end{aligned} \tag{3.34}$$

From (3.29), we get

$$|D(t,x)| \leq 4\eta\, C_{(3.29)} \int_0^t \mathrm{d}s (1 + s^{-\kappa}) S^\alpha_{t-s}(|D(s)|\phi^\beta)(x). \tag{3.35}$$

Thus

$$\begin{aligned} \|D(t)\|_{\mathcal{H}} &\leq 4\eta\, C_{(3.29)} \int_0^t \mathrm{d}s (1 + s^{-\kappa}) \|S^\alpha_{t-s}(|D(s)|\phi^\beta)\|_{\mathcal{H}} \\ &\leq 4\eta\, C_{(3.29)} \int_0^t \mathrm{d}s (1 + s^{-\kappa}) C_{2.9}(t-s)^{-\lambda} \|D(s)\|_{\mathcal{H}}, \end{aligned} \tag{3.36}$$

$0 < t \leq T$, where we used Corollary 2.9 and notation (3.8). Setting

$$D_t := \sup_{0 < s \leq t} \|D(s)\|_{\mathcal{H}}, \quad 0 < t \leq T \tag{3.37}$$

(for finiteness, recall (3.33)), since $I(t)$ from (3.7) is increasing in t (recall representation (3.15)), we find

$$D_t \le C_{(3.38)} D_t I(t), \quad 0 < t \le T, \tag{3.38}$$

with some constant $C_{(3.38)} = C_{(3.38)}(T, \varphi)$. Therefore, by (3.7), $D_t = 0$ for all sufficiently small t, say $t < \delta(T, \varphi) < T$. Since the model is time-homogeneous, we can repeat the argument finitely often to extend to the whole interval $[0, T]$. In fact, when iterating the argument, we need to use the bound (3.29) with the constant $C_{(3.29)}$ depending on our original φ, not with φ replaced by the new initial condition at the beginning of the new subinterval, since (3.29) holds on the whole $(0, T]$. Thus, the constants $C_{(3.38)}$ will be the same for each subinterval, and thus each subinterval will have the same length. (For a more complicated time-inhomogeneous situation, see the proof of Lemma 3.8 below.) Because u and v are Φ-valued, we found $u = v$ on $[0, T]$, and the proof is complete. □

3.4. Auxiliary functions w_N, . Recall that we fixed T, φ, and ψ_1 satisfying (3.4a). For fixed integer $N \ge 2$ set

$$\psi_N := \psi_1 \wedge N. \tag{3.39}$$

We inductively define functions $w_{N,n}$. First of all,

$$w_{N,0}(t) := S^\alpha_t \varphi \in \Phi, \quad 0 \le t \le T. \tag{3.40}$$

Assuming that we have defined $w_{N,n}$ for some n, let

$$w_{N,n+1}(t,x) := S^\alpha_t \varphi(x) - \int_0^t \mathrm{d}s S^\alpha_{t-s}(\psi_N(s) w_{N,n}(s))(x), \tag{3.41}$$

$0 \le t \le T$, $x \in \dot{\mathsf{R}}^d$, provided the latter integral makes sense.

LEMMA 3.8 (properties of $w_{N,n}$). *For all $n \ge 0$ and $t \in [0, T]$,*

$$0 \le w_{N,n}(t) \le S^\alpha_t \varphi, \quad \textit{and} \quad x \mapsto w_{N,n}(t,x) \textit{ is continuous.} \tag{3.42}$$

Proof. For $n = 0$, the claim is true by (3.40). Suppose that we have verified (3.42) for some $n \ge 0$. Then the integral in (3.41) is nonnegative, and hence

$$w_{N,n+1}(t) \le S^\alpha_t \varphi, \quad t \in [0, T]. \tag{3.43}$$

Assume for the moment that $w_{N,n+1} \ge 0$ under our induction hypothesis. Then by Lemma 3.4,

$$w_{N,n+1}(t) \in \Phi, \quad t \in [0, T], \tag{3.44}$$

and the proof would be finished.

Next we will verify that $w_{N,n+1}$ is nonnegative on $[0, 1/N]$. Since $\psi_N \le N$, and using the induction assumption, it follows that

$$S^\alpha_{t-s}(\psi_N(s) w_{N,n}(s)) \le S^\alpha_{t-s}(N S^\alpha_s \varphi) \le N S^\alpha_t \varphi. \tag{3.45}$$

Therefore, if $0 \le t \le 1/N$,

$$w_{N,n+1}(t) \ge S^\alpha_t \varphi - N \int_0^{1/N} \mathrm{d}s S^\alpha_t \varphi = 0. \tag{3.46}$$

Now we prove that $w_{N,n+1}$ is nonnegative on $[0,T]$. We use induction on the time intervals $[k/N,(k+1)/N]$, $0 \le k < NT$. To begin with, we have already shown that $w_{N,n+1}$ is nonnegative on $[0,1/N]$. Also, we know already that $w_{N,n+1}(1/N) \in \Phi$. Suppose that we have shown $w_{N,n+1}$ is nonnegative on $[(k-1)/N,k/N]$ for some $0 \le k < NT-1$, and that $w_{N,n+1}(k/N) \in \Phi$. We will shift time and define

$$w^{(k)}_{N,n+1}(t) := w_{N,n+1}(t+k/N), \tag{3.47a}$$

$$\varphi^{(k)}_{N,n+1}(t) := w_{N,n+1}(k/N), \tag{3.47b}$$

$$\psi^{(k)}_{N}(t) := \psi_N(t+k/N), \tag{3.47c}$$

$0 \le t \le 1/N$. We claim that

$$w^{(k)}_{N,n+1}(t) := S^\alpha_t \varphi^{(k)}_{N,n+1} - \int_0^t \mathrm{d}s S^\alpha_{t-s}(\psi^{(k)}_N(s) w^{(k)}_{N,n}(s)), \quad 0 \le t \le \frac{1}{N}. \tag{3.48}$$

Assume for the moment that (3.48) is true. Then the proof that $w_{N,n+1} \ge 0$ on $[k/N,(k+1)/N]$ reduces to showing that $w^{(k)}_{N,n+1} \ge 0$ on $[0,1/N]$. But this follows from the step we have already done. We are left with showing (3.48).

Using definition (3.41), we get

$$S^\alpha_t w_{N,n+1}(r) := S^\alpha_{t+r}\varphi - \int_0^r \mathrm{d}s S^\alpha_{t+r-s}(\psi_N(s) w_{N,n}(s)). \tag{3.49}$$

Let $r = k/N$. Then, for $0 \le t \le 1/N$,

$$S^\alpha_t \varphi^{(k)}_{N,n+1} = S^\alpha_{t+k/N}\varphi - \int_0^{k/N} \mathrm{d}s S^\alpha_{t+k/N-s}(\psi_N(s) w_{N,n}(s)). \tag{3.50}$$

Also, by a change of variables, for $0 \le t \le 1/N$, we obtain

$$\int_0^t \mathrm{d}s S^\alpha_{t-s}(\psi^{(k)}_N(s) w^{(k)}_{N,n}(s)) = \int_{k/N}^{t+k/N} \mathrm{d}s S^\alpha_{t+k/N-s}(\psi_N(s) w_{N,n}(s)). \tag{3.51}$$

Inserting (3.50) and (3.51) into the right-hand side of (3.48), and then using (3.41), we get

$$\begin{aligned} &S^\alpha_t \varphi^{(k)}_{N,n+1} - \int_0^t \mathrm{d}s S^\alpha_{t-s}(\psi^{(k)}_N(s) w^{(k)}_{N,n}(s)) \\ &\quad = S^\alpha_{t+k/N}\varphi - \int_0^{t+k/N} \mathrm{d}s S^\alpha_{t+k/N-s}(\psi_N(s) w_{N,n}(s)) \\ &\quad = w_{N,n+1}(t+k/N) = w^{(k)}_{N,n+1}(t), \end{aligned} \tag{3.52}$$

which proves (3.48). This finishes the proof. □

3.5. Auxiliary functions w . Recall that we fixed T, φ, ψ_1 with (3.4a). For $n \ge 0$, we inductively define functions w_n as follows. Let

$$w_0(t) := S^\alpha_t \varphi, \quad 0 \le t \le T, \tag{3.53}$$

and, given w_n, set

$$w_{n+1}(t,x) := S^\alpha_t \varphi(x) - \int_0^t \mathrm{d}s S^\alpha_{t-s}(\psi_1(s) w_n(s))(x), \tag{3.54}$$

$0 \le t \le T$, $x \in \dot{\mathsf{R}}^d$. Recall the functions $w_{N,n}$ as in Lemma 3.8.

LEMMA 3.9 (properties of w_n). *For each $n \geq 0$ and $t \in [0,T]$,*

$$\lim_{N\uparrow\infty} |w_{N,n}(t) - w_n(t)| = 0 \tag{3.55}$$

pointwise on $\dot{\mathsf{R}}^d$. Moreover,

$$0 \leq w_n(t) \leq S_t^\alpha \varphi, \quad \text{and} \quad x \mapsto w_n(t,x) \text{ is continuous.} \tag{3.56}$$

Proof. Again, we use induction on n. The claims are trivially true for $n = 0$. Suppose they hold for n. By definitions (3.41) and (3.54),

$$\begin{aligned} |w_{n+1}(t) - w_{N,n+1}(t)| &= \left| \int_0^t \mathrm{d}s S_{t-s}^\alpha (\psi_1(s) w_n(s) - \psi_N(s) w_{N,n}(s)) \right| \\ &\leq \left| \int_0^t \mathrm{d}s S_{t-s}^\alpha ([\psi_1(s) - \psi_N(s)] w_n(s)) \right| \\ &\quad + \int_0^t \mathrm{d}s S_{t-s}^\alpha (\psi_1(s) |w_n(s) - w_{N,n}(s)|) =: A_{N,n} + B_{N,n}, \end{aligned} \tag{3.57}$$

with the obvious correspondence. We will show that both $A_{N+1,n}$ and $B_{N+1,n}$ tend to 0 as $N \uparrow \infty$, giving (3.55) for $n+1$. This then yields also the remaining claims in Lemma 3.9 for $n+1$. In fact, by Lemma 3.8 then the inequalities hold in (3.56), and Lemma 3.4 gives the continuity claim.

First note that by the induction hypothesis,

$$\begin{aligned} A_{N,n} &\leq \int_0^t \mathrm{d}s S_{t-s}^\alpha ([\psi_1(s) - \psi_N(s)] S_s^\alpha \varphi) \\ &\leq \int_0^t \mathrm{d}s S_{t-s}^\alpha (\psi_1(s) S_s^\alpha \varphi) < \infty \end{aligned} \tag{3.58}$$

by Lemma 3.4. Thus, by monotone convergence,

$$\lim_{N\uparrow\infty} A_{N,n} = 0. \tag{3.59}$$

By Lemma 3.8 and the induction hypothesis,

$$|w_{N,n}(t) - w_n(t)| \leq 2 S_t^\alpha \varphi. \tag{3.60}$$

Moreover, by the induction assumption, (3.55) holds. Then, by Lemma 3.4, the dominated convergence theorem implies that

$$\lim_{N\uparrow\infty} B_{N,n} = 0, \tag{3.61}$$

finishing the proof. □

3.6. A linearized equation. Next we show that w_n converges as $n \uparrow \infty$ to a solution of a linearized equation.

LEMMA 3.10 (linearized equation). *Fix again $T \geq 1$, $\varphi \in \Phi$, and $\psi_1 : \mathsf{R}_+ \times \mathsf{R}^d \to \mathsf{R}_+$ satisfying* (3.4a). *Then, for $0 \leq t \leq T$, in $\mathcal{H}$ the (nonnegative) limit*

$$w(t) := \lim_{n\uparrow\infty} w_n(t) \tag{3.62}$$

exists and is the unique Φ*-valued solution to*

$$w(t) = S_t^\alpha \varphi - \int_0^t \mathrm{d}s S_{t-s}^\alpha(\psi_1(s)w(s)), \quad 0 \le t \le T, \tag{3.63}$$

implying

$$0 \le w(t) \le S_t^\alpha \varphi, \quad 0 \le t \le T. \tag{3.64}$$

Proof. Extending the uniqueness proof (Lemma 3.7) in the obvious way, one can show that

$$\sup_{0<s\le t} \|w_n(s) - w_m(s)\|_{\mathcal{H}} \xrightarrow[n,m\uparrow\infty]{} 0, \tag{3.65}$$

provided that $t < \delta = \delta(T, \varphi) < T$. Hence, there is a measurable mapping $s \mapsto w(s) \in \mathcal{H}$ on $[0, \delta]$ such that

$$\sup_{0<s\le t} \|w_n(s) - w(s)\|_{\mathcal{H}} \xrightarrow[n\uparrow\infty]{} 0. \tag{3.66}$$

Obviously, for each choice of representatives, w satisfies (3.63). By Lemma 3.4, the $w(s)$ belong to Φ, and the proof is finished. □

3.7. Existence of solutions. Our next goal is to use Lemma 3.10 to prove the existence of a Φ-valued solution for (3.1). Hypothesis 3.2 is still in force.

LEMMA 3.11 (existence). *To each* $\varphi \in \Phi$*, there exists a* Φ*-valued solution* $v = V\varphi$ *to the log-Laplace equation* (3.1).

Proof. We want to construct a sequence of Φ-valued functions v_m satisfying

$$v_m(t) \le S_t^\alpha \varphi. \tag{3.67}$$

In fact, if $m = 0$, set $v_0 := S^\alpha \varphi$. Assume that we have already defined v_m for some $m \ge 0$. Note that by Corollary 2.11,

$$|v_m^\beta(t)| \le M(1 + t^{-\kappa})\phi^\beta, \tag{3.68}$$

with κ from (3.5). Let v_{m+1} be the unique Φ-valued solution to

$$v_{m+1}(t,x) = S_t^\alpha \varphi(x) - \int_0^t \mathrm{d}s S_{t-s}^\alpha(v_m^\beta(s) v_{m+1}(s))(x) \tag{3.69}$$

according to Lemma 3.10, implying (3.67) for $m + 1$. Altogether, by induction we defined Φ-valued functions v_m satisfying (3.69), (3.67), and (3.68).

For $m \ge 0$, let

$$D_m := v_{m+1} - v_m. \tag{3.70}$$

Then, as in the proof of Lemma 3.7 (uniqueness), using Lemma 3.6, for fixed $T > 0$, we find

$$|D_{m+1}(t)| \le C \int_0^t \mathrm{d}s (1 + s^{-\kappa}) S_{t-s}^\alpha(|D_m(s)|\phi^\beta), \tag{3.71}$$

$0 \le t \le T$, with a constant $C = C(T)$, and

$$\begin{aligned} \|D_{m+1}(t)\|_{\mathcal{H}} &\le C \int_0^t \mathrm{d}s(1 + s^{-\kappa})\|S^{\alpha}_{t-s}(|D_m(s)|\phi^{\beta})\|_{\mathcal{H}} \\ &\le C \int_0^t \mathrm{d}s(1 + s^{-\kappa})C_{2.9}(t-s)^{-\lambda}\|D_m(s)\|_{\mathcal{H}}. \end{aligned} \tag{3.72}$$

Setting

$$D_{m,t} := \sup_{0\le s\le t} \|D_m(s)\|_{\mathcal{H}}, \quad 0 \le t \le T, \tag{3.73}$$

we found that

$$D_{m+1,t} \le \varepsilon_t D_{m,t}, \tag{3.74}$$

where the ε_t are independent of m, and $\varepsilon_t \to 0$ as $t \downarrow 0$. Thus, if our $T > 0$ is small enough, then there exists a constant $0 < \gamma < 1$ such that if $0 \le t \le T$, then

$$D_{m+1,t} \le \gamma D_{m,t}, \tag{3.75}$$

and so

$$D_{m,t} \le \gamma^m D_{0,t}. \tag{3.76}$$

Therefore, we can define

$$v(t,x) := \sum_{m=0}^{\infty} D_m(t,x) = \lim_{m\uparrow\infty} v_m(t,x), \quad 0 \le t \le T,\ x \ne 0, \tag{3.77}$$

where the limit is taken in $\mathcal{H}_+$.

From our construction, it follows that

$$0 \le v(t) \le S^{\alpha}_t\varphi \quad \text{and} \quad |v^{\beta}(t)| \le M(1 + t^{-\kappa})\phi^{\beta}. \tag{3.78}$$

Now we want to show that v satisfies (3.1) for $0 \le t \le T$. We start from definition (3.69). First, by (3.77),

$$\lim_{m\to\infty} \|v_{m+1}(t) - v(t)\|_{\mathcal{H}} = 0, \quad 0 \le t \le T. \tag{3.79}$$

As for the integral terms, we first note that for $a, b, c \ge 0$, by (3.30) we have

$$\begin{aligned} |ab^{\beta} - c^{1+\beta}| &\le |a^{1+\beta} - c^{1+\beta}| + |b^{1+\beta} - c^{1+\beta}| \\ &\le (1+\beta)(a+c)^{\beta}|a-c| + (1+\beta)(b+c)^{\beta}|b-c|. \end{aligned} \tag{3.80}$$

Therefore, using the second part of (3.78), we have

$$\begin{aligned} &\left\| \int_0^t \mathrm{d}s S^{\alpha}_{t-s}(v_{m+1}(s)v_m^{\beta}(s)) - \int_0^t \mathrm{d}s S^{\alpha}_{t-s} v^{1+\beta}(s) \right\|_{\mathcal{H}} \\ &\quad \le \int_0^t \mathrm{d}s \|S^{\alpha}_{t-s}|v_{m+1}(s)v_m^{\beta}(s) - v^{1+\beta}(s)|\|_{\mathcal{H}} \\ &\quad \le C \int_0^t \mathrm{d}s(1 + s^{-\kappa})\|S^{\alpha}_{t-s}(\phi^{\beta}|v_{m+1}(s) - v(s)| + \phi^{\beta}|v_m(s) - v(s)|)\|_{\mathcal{H}}. \end{aligned} \tag{3.81}$$

By Corollary 2.9, this chain of inequalities can be continued with

$$\begin{aligned} &\leq C\int_0^t \mathrm{d}s(1+s^{-\kappa})(t-s)^{-\lambda}\|v_{m+1}(s)-v(s)\|_{\mathcal{H}} \\ &\quad + C\int_0^t \mathrm{d}s(1+s^{-\kappa})(t-s)^{-\lambda}\|v_m(s)-v(s)\|_{\mathcal{H}}. \end{aligned} \tag{3.82}$$

By dominated convergence, the latter terms converge to 0 as $m \uparrow \infty$. In fact, the norm expressions tend to 0 by (3.79), the v_m and v are dominated by $S^\alpha\varphi$, which are strongly continuous by Corollary 2.10, and hence bounded in norm on the finite time interval, and recall (3.7). Thus, v satisfies (3.1) in $\mathcal{H}_+$ for $t \leq T$ and for sufficiently small T. By induction on intervals, as in the proof of Lemma 3.8, we extend the solution v from $[0,T]$ to all times. By Lemma 3.4, the constructed solution v is Φ-valued. □

Completion of the proof of Theorem 3.3. With Lemmas 3.7 and 3.11, we already proved the uniqueness and existence claims, respectively. The semigroup property of V follows from the uniqueness of solutions, and the strong continuity of V from Lemma 3.5. □

4. Construction of X. Having available the well-posedness of the log-Laplace equation (Theorem 3.3) under Hypothesis 3.2, we now construct the desired process X via a Trotter product approach to the related log-Laplace semigroup. For this purpose, we introduce an approximating log-Laplace equation related to separating critical continuous-state branching with index $1+\beta$, and mass flow according to S^α on alternate small time intervals (Lemma 4.1). The main work consists in showing that its solutions converge to the solutions to the true log-Laplace equation (Proposition 4.3). Then, in the final subsection, all pieces can easily put together to get our main result, the existence theorem, Theorem 4.4.

4.1. Approximating equation. We continue to impose Hypothesis 3.2. Fix $n \geq 1$ and $\varphi \in \Phi$. We inductively define measurable functions $\bar{v}_n$ on $\mathsf{R}_+ \times \dot{\mathsf{R}}^d$. First of all,

$$\bar{v}_n(0) := S^\alpha_{1/n}\varphi. \tag{4.1}$$

Assume for the moment $\bar{v}_n(\frac{k}{n})$ is defined for some $k \geq 0$. For $\frac{k}{n} \leq t < \frac{k+1}{n}$, set

$$\bar{v}_n(t,x) := \frac{\bar{v}_n(\frac{k}{n},x)}{\left[1+\eta\beta\bar{v}_n^\beta(\frac{k}{n},x)(t-\frac{k}{n})\right]^{1/\beta}}, \quad x \neq 0. \tag{4.2}$$

Note that

$$\frac{\partial}{\partial t}\bar{v}_n(t,x) = -\eta\bar{v}_n^{1+\beta}(t,x) \quad \text{on} \quad \left(\frac{k}{n},\frac{k+1}{n}\right)\times\dot{\mathsf{R}}^d, \tag{4.3}$$

that $\bar{v}_n(\frac{k}{n}+,x) = \bar{v}_n(\frac{k}{n},x)$, and that also the limit $\bar{v}_n(\frac{k+1}{n}-,x)$ exists. Note also that for $x \neq 0$ fixed, (4.2) gives the log-Laplace transition function of a critical continuous-state branching process (on R_+) with index $1+\beta$ (see, for instance, [Lam67]). Put

$$\bar{v}_n\left(\frac{k+1}{n},x\right) := S^\alpha_{1/n}\bar{v}_n\left(\frac{k+1}{n}-,\cdot\right)(x), \quad x \neq 0. \tag{4.4}$$

LEMMA 4.1 (approximating log-Laplace equation). *The function $\bar{v}_n \geq 0$ we have just defined satisfies*

$$\bar{v}_n(t,x) = S^\alpha_{(1+[tn])/n}\varphi(x) - \eta\int_0^t \mathrm{d}s S^\alpha_{([tn]-[sn])/n}(\bar{v}_n^{1+\beta}(s))(x) \tag{4.5}$$

on $\mathsf{R}_+ \times \dot{\mathsf{R}}^d$.

Proof. Differentiating equation (4.5) to $t \neq \frac{k}{n}$, $k \geq 0$, gives the true statement (4.3). On the other hand, for $t = \frac{k}{n}$, $k \geq 0$,

$$\bar{v}_n\left(\frac{k}{n}\right) = S^\alpha_{(1+k)/n}\varphi - \eta\sum_{i=1}^k \left[S^\alpha_{(k-(i-1))/n}\int_{(i-1)/n}^{i/n} \mathrm{d}s\bar{v}_n^{1+\beta}(s)\right]. \tag{4.6}$$

By (4.3) and the fundamental theorem of calculus, the right-hand side of the latter equation equals $\bar{v}_n(\frac{k}{n})$, finishing the proof. □

Set

$$t_n := [tn]/n. \tag{4.7}$$

Since $\bar{v}_n$ is nonnegative, from (4.5) we get the *domination*

$$0 \leq \bar{v}_n(t) \leq S^\alpha_{1/n+t_n}\varphi, \quad t \geq 0,\ n \geq 1, \tag{4.8}$$

implying by Corollary 2.9 with $\beta = 0$,

$$\|\bar{v}_n(t)\|_{\mathcal{H}} \leq C_{(4.9)}\|\varphi\|, \quad 0 \leq t \leq T,\ n \geq 1, \tag{4.9}$$

for each $T > 0$, and where $C_{(4.9)} = C_{(4.9)}(T)$. In particular, $\bar{v}_n$ is $\mathcal{H}_+$-valued. Our aim is to show that the $\bar{v}_n$ converge to the unique solution to (3.1). For this purpose, we will need the following estimate.

LEMMA 4.2 (pointwise bound). *Impose Hypothesis* 3.2. *To each $\varphi \in \mathcal{H}_+$ and $T > 0$, there is a $\varphi_0 = \varphi_0(d,T,\alpha,\varphi,\varrho)$ in $\mathcal{H}_+$, such that*

$$\sup_{T/2\leq t\leq T} S^\alpha_t\varphi \leq \varphi_0. \tag{4.10}$$

Proof. Recall from (2.35) that

$$P^\alpha(t;x,y) \leq P(t;x,y) + C_{2.6}\bar{P}(t;x,y). \tag{4.11}$$

Choose a constant $C_{(4.12)} = C_{(4.12)}(d,T)$ such that, for $T/2 \leq t \leq T$,

$$P(t;x,y) \leq C_{(4.12)}P(T;x,y) \tag{4.12}$$

and

$$\bar{P}(t;x,y) = t^{-1/2}\phi(x)\phi(y)\mathrm{e}^{-|x|^2/4t}\mathrm{e}^{-|y|^2/4t} \leq C_{(4.12)}\bar{P}(T;x,y) \tag{4.13}$$

(recall (2.34)). From Lemma 2.3 (with $\beta = 0$) we conclude that $S_T\varphi$ belongs to $\mathcal{H}_+$, whereas Lemma 2.7 (with $\beta = 0$) gives $\bar{S}_T\varphi \in \mathcal{H}_+$. Therefore, we may set

$$\varphi_0 := C_{(4.12)}(S_T\varphi + C_{2.6}\bar{S}_T\varphi) \tag{4.14}$$

to finish the proof. □

4.2. Convergence to the limit equation. With the functions $\bar v_n$ we may pass to the limit.

PROPOSITION 4.3 (convergence to the limit equation). *Fix $\varphi \in \Phi$. Define $\bar v_n$ as in* (4.1)–(4.4). *Let $v = V\varphi$ be the unique Φ-valued solution to* (3.1) *according to Theorem* 3.3. *Then, for each $t \geq 0$,*

$$\lim_{n\uparrow\infty} \|v(t) - \bar v_n(t)\|_{\mathcal H} = 0. \tag{4.15}$$

Proof. We may restrict our attention to $t \in [0,T]$ for any fixed $T > 1$. For $n \geq 1$ and with t_n defined in (4.7), from (3.1) and (4.5) we have

$$\begin{aligned}
\|v(t) - \bar v_n(t)\|_{\mathcal H} \leq\ & \|S^\alpha_t\varphi - S^\alpha_{1/n+t_n}\varphi\|_{\mathcal H} \\
& + \eta \int_0^{t_n} \mathrm{d}s \|S^\alpha_{t-s} v^{1+\beta}(s) - S^\alpha_{t_n - s_n} v^{1+\beta}(s)\|_{\mathcal H} \\
& + \eta \int_0^{t_n} \mathrm{d}s \|S^\alpha_{t_n - s_n} |v^{1+\beta}(s) - \bar v_n^{1+\beta}(s)|\|_{\mathcal H} \\
& + \eta \left\| \int_{t_n}^t \mathrm{d}s S^\alpha_{t-s} v^{1+\beta}(s) \right\|_{\mathcal H} + \eta \left\| \int_{t_n}^t \mathrm{d}s S^\alpha_{t_n - s_n} \bar v_n^{1+\beta}(s) \right\|_{\mathcal H} \\
=:\ & A_n(t) + B_n(t) + C_n(t) + D_n(t) + E_n(t),
\end{aligned} \tag{4.16}$$

with the obvious correspondence. We will deal with each of these terms separately.

Step 1 ($A_n(t)$). From the semigroup property and boundedness (2.68),

$$A_n(t) = \|S^\alpha_t\varphi - S^\alpha_{1/n+t_n}\varphi\|_{\mathcal H} \leq C_{2.9} \|S^\alpha_{|t-1/n-t_n|}\varphi - \varphi\|_{\mathcal H}. \tag{4.17}$$

But

$$\left| t - \frac{1}{n} - t_n \right| \leq \frac{1}{n}, \quad t \geq 0. \tag{4.18}$$

Hence,

$$\sup_{0\leq t\leq T} A_n(t) \leq C_{2.9} \sup_{0\leq s\leq 2/n} \|S^\alpha_s\varphi - \varphi\|_{\mathcal H} \underset{n\uparrow\infty}{\longrightarrow} 0 \tag{4.19}$$

by strong continuity according to Corollary 2.10.

Step 2 ($D_n(t)$). Clearly, by our estimates (recall (3.29) and (2.67)),

$$D_n(t) \leq C(\varphi)\eta \int_{t_n}^t \mathrm{d}s (1 + s^{-\kappa})(t-s)^{-\lambda} \|\varphi\|_{\mathcal H}. \tag{4.20}$$

By scaling, the integral equals

$$t^{1-\lambda} \int_{t_n/t}^1 \mathrm{d}s (1-s)^{-\lambda} + t^{1-\kappa-\lambda} \int_{t_n/t}^1 \mathrm{d}s s^{-\kappa}(1-s)^{-\lambda} =: I_n(t) + I\!I_n(t),$$

with the obvious correspondence. Take $\varepsilon \in (0,T)$ and let $n > 1/\varepsilon$. Since

$$\frac{t_n}{t} \geq 1 - \frac{1}{tn} \geq 1 - \frac{1}{\varepsilon n}, \quad \varepsilon \leq t \leq T, \tag{4.21}$$

we get

$$\sup_{\varepsilon\le t\le T} I_n(t) \le T^{1-\lambda}\int_{1-1/\varepsilon n}^{1} \mathrm{d}s(1-s)^{-\lambda} \underset{n\uparrow\infty}{\longrightarrow} 0, \tag{4.22}$$

whereas

$$\sup_{0\le t\le \varepsilon} I_n(t) \le \varepsilon^{1-\lambda}\int_{0}^{1} \mathrm{d}s(1-s)^{-\lambda} \le\ C\varepsilon^{1-\lambda}. \tag{4.23}$$

Now ε was arbitrary, and consequently,

$$\sup_{0\le t\le T} I_n(t) \underset{n\uparrow\infty}{\longrightarrow} 0, \tag{4.24}$$

and the same reasoning leads to the analogous statement on $I\!I_n(t)$. Summarizing,

$$\sup_{0\le t\le T} D_n(t) \underset{n\uparrow\infty}{\longrightarrow} 0. \tag{4.25}$$

Step 3 ($E_n(t)$). We may assume that $t > t_n$. By (4.5),

$$E_n(t) = \|\bar{v}_n(t) - \bar{v}_n(t_n)\|_{\mathcal{H}}. \tag{4.26}$$

According to the definition (4.2) of $\bar{v}_n(t)$,

$$0 \le \bar{v}_n(t_n,x) - \bar{v}_n(t,x) = \bar{v}_n(t_n,x)\left(1 - \frac{1}{[1+\eta\beta\bar{v}_n^{\beta}(t_n,x)(t-t_n)]^{1/\beta}}\right). \tag{4.27}$$

Since t is fixed, it follows that for n large enough, $t/2 < t_n$. Using domination and Lemma 4.2, there is a $\varphi_0 = \varphi_0(d,t,\alpha,\varphi,\varrho) \in \mathcal{H}_+$ such that

$$\bar{v}_n(t_n,x) \le S^{\alpha}_{1/n+t_n}\varphi(x) \le \varphi_0. \tag{4.28}$$

But (4.27) is increasing in $\bar{v}_n(t_n,x)$, so we may insert (4.28) to obtain

$$0 \le \bar{v}_n(t_n,x) - \bar{v}_n(t,x) \le \varphi_0(x)\left(1 - \frac{1}{[1+\eta\beta\varphi_0^{\beta}(x)/n]^{1/\beta}}\right) \le \varphi_0(x),$$

since $0 \le t - t_n \le 1/n$. Then, from dominated convergence we get

$$\lim_{n\to\infty} E_n(t) = 0 \tag{4.29}$$

for our fixed t.

Step 4 ($B_n(t)$). First of all, we want to deal with $B_n(t)$ for small t. Clearly, for $0 < s < t_n$,

$$\|S^{\alpha}_{t-s}v^{1+\beta}(s)\|_{\mathcal{H}} \le C(1+s^{-\kappa})(t-s)^{-\lambda}\|\varphi\|_{\mathcal{H}} \tag{4.30}$$

and

$$\|S^{\alpha}_{t_n-s_n}v^{1+\beta}(s)\|_{\mathcal{H}} \le C(1+s^{-\kappa})(t_n-s)^{-\lambda}\|\varphi\|_{\mathcal{H}} \tag{4.31}$$

since

$$t_n - s_n \geq t_n - s > 0. \tag{4.32}$$

Let $0 < \varepsilon < T$. Using notation (3.7), from (4.30) and (4.31), for $0 \leq t \leq \varepsilon$,

$$B_n(t) \leq C[I(t) + I(t_n)]. \tag{4.33}$$

Moreover, since I is increasing (recall (3.15)),

$$\sup_{n\geq 1} \sup_{0\leq t\leq \varepsilon} B_n(t) \leq C \sup_{0\leq t\leq \varepsilon} I(t) \underset{\varepsilon\downarrow 0}{\longrightarrow} 0. \tag{4.34}$$

Now we may restrict our attention to $t \in [\varepsilon, T]$. We want to exploit the strong continuity of the semigroup S^α acting on $\mathcal{H}_+$ (Corollary 2.10). To this end, we truncate $v^{1+\beta}$ to a function in $\mathcal{H}_+$, and consider a small time interval around t_n separately to get rid of the varying upper integration bound. Here are the details.

Take $\delta \in (0, \varepsilon)$ and $N \geq 1$. Set

$$v_{1,N}(t) := (v(t) \wedge N)1_{B_N(0)}, \tag{4.35a}$$

$$v_{2,N}(t) := v(t) - v_{1,N}(t). \tag{4.35b}$$

Then for $\varepsilon \leq t \leq T$,

$$\begin{aligned} B_n(t) \leq & \int_0^{t-\delta} \mathrm{d}s \|S^\alpha_{t-s} v^{1+\beta}_{1,N}(s) - S^\alpha_{t_n-s_n} v^{1+\beta}_{1,N}(s)\|_{\mathcal{H}} \\ & + \int_{t-\delta}^{t} \mathrm{d}s \|S^\alpha_{t-s} v^{1+\beta}(s)\|_{\mathcal{H}} + \int_{t-\delta}^{t_n} \mathrm{d}s \|S^\alpha_{t_n-s_n} v^{1+\beta}(s)\|_{\mathcal{H}} \\ & + \int_0^{t-\delta} \mathrm{d}s \|S^\alpha_{t-s} v^{1+\beta}_{2,N}(s)\|_{\mathcal{H}} + \int_0^{t-\delta} \mathrm{d}s \|S^\alpha_{t_n-s_n} v^{1+\beta}_{2,N}(s)\|_{\mathcal{H}} \\ =: & B_n^{(1)}(t) + \cdots + B_n^{(5)}(t) \end{aligned} \tag{4.36}$$

in the obvious correspondence. Again, we deal with all terms separately.

Step 4.1 ($B_n^{(2)}(t)$). From (4.30) and scaling,

$$B_n^{(2)}(t) \leq C \int_{1-\delta/t}^1 \mathrm{d}s (1-s)^{-\lambda} + C \int_{1-\delta/t}^1 \mathrm{d}s s^{-\kappa}(1-s)^{-\lambda}. \tag{4.37}$$

But $1 - \delta/t \geq 1 - \delta/\varepsilon$, and hence

$$\sup_n \sup_{\varepsilon\leq t\leq T} B_n^{(2)}(t) \underset{\delta\downarrow 0}{\longrightarrow} 0. \tag{4.38}$$

Step 4.2 ($B_n^{(3)}(t)$). Similarly, from (4.31) and (4.32),

$$B_n^{(3)}(t) \leq C \int_{t-\delta}^{t_n} \mathrm{d}s (1 + s^{-\kappa})(t_n - s)^{-\lambda}. \tag{4.39}$$

Now

$$t \geq t_n \geq t - \frac{1}{n} \geq \varepsilon - \frac{1}{n} \geq \delta, \tag{4.40}$$

provided that $n \geq 1/(\varepsilon - \delta)$. Thus, the lower integration bound can be replaced by $t_n - \delta$, and by scaling,

$$B_n^{(3)}(t) \leq C \int_{1-\delta/t_n}^1 \mathrm{d}s(1-s)^{-\lambda} + C \int_{1-\delta/t_n}^1 \mathrm{d}s s^{-\kappa}(1-s)^{-\lambda}. \tag{4.41}$$

By (4.40), the lower integration bounds can be changed to $1 - \delta/(\varepsilon - 1/n)$, implying

$$\limsup_{n\uparrow\infty} \sup_{\varepsilon \leq t \leq T} B_n^{(3)}(t) \underset{\delta\downarrow 0}{\longrightarrow} 0. \tag{4.42}$$

Step 4.3 ($B_n^{(4)}(t)$ *and* $B_n^{(5)}(t)$). By domination and Corollary 2.11,

$$v_{2,N}^{\beta}(s) \leq v^{\beta}(s) \leq (S_s^{\alpha}\varphi)^{\beta} \leq C(1+s)^{-\kappa}\phi^{\beta}. \tag{4.43}$$

Note that for $s \in [0, t-\delta]$,

$$t - s \quad \text{and} \quad t_n - s_n \quad \text{belong to } [\delta/2, t] \text{ if } n > 2/\delta \tag{4.44}$$

(for instance, $t_n - s_n \geq t - \frac{1}{n} - s \geq -\frac{1}{n} + \delta \geq \delta$). Then, for $r \in [\delta/2, t]$ and $n > 2/\delta$,

$$\|S_r^{\alpha} v_{2,N}^{1+\beta}(s)\|_{\mathcal{H}} \leq C(1+s)^{-\kappa} \|S_r^{\alpha}(v_{2,N}(s)\phi^{\beta})\|_{\mathcal{H}} \leq C(1+s)^{-\kappa}\|v_{2,N}(s)\|_{\mathcal{H}},$$

where in the last step we used Lemma 2.7 with φ replaced by $v_{2,N}(s)$. Now by domination and boundedness, for $0 \leq s \leq T$,

$$\|v_{2,N}(s)\|_{\mathcal{H}} \leq \|S_s^{\alpha}\varphi\|_{\mathcal{H}} \leq C\|\varphi\|_{\mathcal{H}} \tag{4.45}$$

and

$$\|v_{2,N}(s)\|_{\mathcal{H}} \underset{N\uparrow\infty}{\searrow} 0. \tag{4.46}$$

Therefore,

$$\begin{aligned} &\limsup_{n\uparrow\infty} \sup_{\varepsilon \leq t \leq T} (B_n^{(4)}(t) + B_n^{(5)}(t)) \\ &\qquad \leq C \int_0^T \mathrm{d}s(1+s)^{-\kappa} \|v_{2,N}(s)\|_{\mathcal{H}} \underset{N\uparrow\infty}{\searrow} 0, \end{aligned} \tag{4.47}$$

by monotone convergence, for the fixed ε and δ.

Step 4.4 ($B_n^{(1)}(t)$). It remains to deal with $B_n^{(1)}(t)$. By the semigroup property, boundedness, and strong continuity,

$$\begin{aligned} &\|S_{t-s}^{\alpha} v_{1,N}^{1+\beta}(s) - S_{t_n - s_n}^{\alpha} v_{1,N}^{1+\beta}(s)\|_{\mathcal{H}} \\ &\qquad \leq C \sup_{0 \leq r \leq 2/n} \|S_r^{\alpha} v_{1,N}^{1+\beta}(s) - v_{1,N}^{1+\beta}(s)\|_{\mathcal{H}} \underset{N\uparrow\infty}{\searrow} 0 \end{aligned} \tag{4.48}$$

for all s and N, since by definition

$$v_{1,N}^{1+\beta}(s) \leq N^{1+\beta} 1_{B_N(0)} \in \mathcal{H}_+. \tag{4.49}$$

Moreover, the supremum in (4.48) is bounded from above by

$$2N^{1+\beta}\|1_{B_N(0)}\|_{\mathcal{H}}. \tag{4.50}$$

Therefore,

$$\sup_{\varepsilon\le t\le T} B_n^{(1)}(t) \le \int_0^T \mathrm{d}s \sup_{0\le r\le 2/n} \|S_r^\alpha v_{1,N}^{1+\beta}(s) - v_{1,N}^{1+\beta}(s)\|_{\mathcal{H}} \underset{n\uparrow\infty}{\searrow} 0, \tag{4.51}$$

by monotone convergence, for all our N, ε, δ.

Step 4.5 (*conclusion*). Putting together (4.34), (4.38), (4.42), (4.47), and (4.51),

$$\sup_{0\le t\le T} B_n(t) \underset{n\uparrow\infty}{\longrightarrow} 0. \tag{4.52}$$

Step 5 ($C_n(t)$). First note that

$$C_n(t) = 0 \quad \text{for } t \le 1/n. \tag{4.53}$$

So we may assume that $t \ge 1/n$. Next we apply Lemma 3.6 to get for the term in abstract value sign in the definition of $C_n(t)$ the bound

$$C[|v|^\beta(s) + |\bar v_n|^\beta(s)]|v(s) - \bar v_n(s)|. \tag{4.54}$$

From domination, the expression in square brackets is bounded by

$$(S_s^\alpha\varphi)^\beta + (S_{1/n+s_n}^\alpha\varphi)^\beta \le C(1+s^{-\kappa})\phi^\beta, \tag{4.55}$$

where we used Corollary 2.11 and $1/n + s_n \ge s$. But by Corollary 2.9,

$$\|S_{t_n-s_n}^\alpha |v(s) - \bar v_n(s)|\phi^\beta\|_{\mathcal{H}} \le C(t_n - s_n)^{-\lambda}\|v(s) - \bar v_n(s)\|_{\mathcal{H}}. \tag{4.56}$$

Setting

$$F_n(t) := \sup_{s\le t} \|v(s) - \bar v_n(s)\|_{\mathcal{H}}, \tag{4.57}$$

we found for $\frac{1}{n} \le t \le T$,

$$\begin{aligned} &\int_0^{t_n} \mathrm{d}s \|S_{t_n-s_n}^\alpha |v^{1+\beta}(s) - \bar v_n^{1+\beta}(s)|\|_{\mathcal{H}} \\ &\qquad \le CF_n(t)\int_0^{t_n} \mathrm{d}s(1+s^{-\kappa})(t_n-s)^{-\lambda} \le CF_n(t)I(t) \end{aligned} \tag{4.58}$$

(recall (3.7)).

Step 6 (*completion of the proof*). By (4.19), (4.25), (4.29), (4.52), and (4.58), from estimate (4.16) we obtain

$$\|v(t) - \bar v_n(t)\|_{\mathcal{H}} \le \varepsilon_n + CF_n(t)I(t) \tag{4.59}$$

for the fixed $t \le T$, where $C = C(T)$ and where $\varepsilon_n = \varepsilon_n(t,T)$ tends to 0 as $n \uparrow \infty$. Imposing additionally on t that $CI(t) < \frac{1}{2}$ (recall (3.7)), we get

$$F_n(t) \le \varepsilon_n + \frac{1}{2}F_n(t), \quad \text{that is, } F_n(t) \le 2\varepsilon_n \underset{n\uparrow\infty}{\longrightarrow} 0. \tag{4.60}$$

Consequently,

$$\|v(t) - \bar v_n(t)\|_{\mathcal{H}} \underset{n\uparrow\infty}{\longrightarrow} 0 \quad \text{for all sufficiently small } t. \tag{4.61}$$

Repeating the argument, we can lift up for $t \in [0,T]$. Since T was arbitrary, the proof of Proposition 4.3 is finished altogether. □

4.3. Construction of the process. Here is now the more precise formulation of our main result, announced in Theorem 1.1.

THEOREM 4.4 (existence of X). *Under Hypothesis* 3.2*, there is a (unique in law) nondegenerate $\mathcal{M}(\dot{\mathsf{R}}^d)$-valued (time-homogeneous) Markov process $X = (X, P_\mu,\ \mu \in \mathcal{M})$ with log-Laplace transition functional* (1.6) *using test functions $\varphi \in \Phi$ and where $v = V\varphi$ is the unique Φ-valued solution to* (3.1)*.*

Remark 4.5 (nondegeneracy). It is easy to see that the following expectation formula holds:

$$P_\mu\langle X_t, \varphi\rangle = \langle \mu, S_t^\alpha \varphi\rangle =: \langle S_t^\alpha \mu, \varphi\rangle, \quad \mu \in \mathcal{M},\ t \geq 0,\ \varphi \in \Phi. \tag{4.62}$$

But

$$V_t\varphi \neq S_t^\alpha \varphi, \quad t > 0,\ \varphi \in \Phi,\ \varphi \neq 0. \tag{4.63}$$

Hence, the log-Laplace formula (1.6) shows that X is different from its expectation; that is, it is nondegenerate.

Proof of Theorem 4.4. For the moment, fix $\mu \in \mathcal{M}$ and $n \geq 1$. Our first purpose is, for $t \geq 0$ fixed, to construct a random measure X_t^n in $\mathcal{M}$ with log-Laplace functional

$$-\log \mathbf{P}\mathrm{e}^{-\langle X_t^n, \varphi\rangle} = \langle \mu, \bar{v}_n(t)\rangle, \quad \varphi \in \Phi, \tag{4.64}$$

where $\bar{v}_n =: V^n\varphi$ is taken from definitions (4.1)–(4.4). Then we later let $n \uparrow \infty$ and obtain (for the fixed t) a random measure X_t in $\mathcal{M}$ with log-Laplace functional

$$-\log \mathbf{P}\mathrm{e}^{-\langle X_t, \varphi\rangle} = \langle \mu, v(t)\rangle, \quad \varphi \in \Phi. \tag{4.65}$$

Actually, we will get a probability kernel Q_t, say, in $\mathcal{M}$, which as a function in t turns out to satisfy Chapman–Kolmogorov. Trivially, Q is then the transition kernel of a time-homogeneous Markov process X, say, in $\mathcal{M}$, regardless of its possible path properties (which are not considered in the theorem) and whether, for fixed n, the family of the X_t^n, $t \geq 0$, is defined on a common probability space (which is necessary to form a random process) or not. Now the details will follow.

As the $\bar{v}_n$ had been constructed by alternating operations of mass flow and continuous-state branching on time intervals of length $\frac{1}{n}$, we will construct also the X_t^n by such an alternation procedure. Note, however, since the two alternating operations do not commute, on the dual level of measures we have to interchange the order of operations.

An essential tool in our construction of the X_t^n will be an $\mathcal{M}$-valued process Y that we will introduce next. As already noted in the beginning of subsection 4.1, the unique solutions g to the ordinary differential equation

$$\frac{\mathrm{d}g}{\mathrm{d}t}(t) = -\eta g^{1+\beta}(t) \quad \text{on} \quad \mathsf{R}_+ \quad \text{with } g(0) = \theta \tag{4.66}$$

yield the log-Laplace transition functions of a critical continuous-state branching process, say $y = \{y_t : t \geq 0\}$, with index $1 + \beta$:

$$-\log \mathbf{P}\{\mathrm{e}^{-y_t\theta} \mid y_0 = a\} = ag(t) \tag{4.67}$$

(see [Lam67]). (Clearly, for $\beta = 1$, we have the famous critical Feller's branching diffusion; otherwise the y_t have infinite variance.) We want to let such processes evolve independently at each spatial point $x \neq 0$. More precisely, consider the $\mathcal{M}$-valued

(time-homogeneous) Markov process $\{Y_t : t \geq 0\}$ with càdlàg paths and log-Laplace transition functional

$$-\log \mathbf{P}\{\mathrm{e}^{-\langle Y_t,\varphi\rangle} \mid Y_0 = \mu\} = \langle \mu, G_t\varphi\rangle, \quad t \geq 0,\ \varphi \in \Phi,\ \mu \in \mathcal{M}, \tag{4.68}$$

where

$$G_t\varphi(x) := g(t,x), \quad t \geq 0,\ x \neq 0, \tag{4.69}$$

and for each $x \neq 0$ fixed, $t \mapsto g(t,x)$ solves (4.66) with θ replaced by $\varphi(x)$. This process Y can also be obtained by starting from a critical super-Brownian motion on R^d with $(1+\beta)$-branching and letting the migration constant of the super-Brownian motion tend to zero; see [DF88, Theorem 5.8]. Although started from any measure $Y_0 = \mu \in \mathcal{M}$, at positive times t the random measures Y_t are atomic, where the (closed) support of Y_t forms a Poisson point field on R^d, whereas the weights of the atoms evolve in time independently as critical continuous-state branching processes y of index $1+\beta$ (see [DF88, discussion after Theorem 3.1]).

For fixed $\mu \in \mathcal{M}$, and $n \geq 1$, we now want inductively to introduce the random measures X_t^n satisfying (4.64). First of all, for $t \in [0, \frac{1}{n})$ set

$$X_t^n \ := \ S_{1/n}^\alpha Y_t, \quad \text{where } Y \text{ starts from } Y_0 := \mu. \tag{4.70}$$

Here we used the notation of smearing out measures according to the flow S^α introduced in (4.62), and Lemma 4.2. Then by (4.68),

$$\mathbf{P}\mathrm{e}^{-\langle X_t^n,\varphi\rangle} = \mathbf{P}\mathrm{e}^{-\langle Y_t, S_{1/n}^\alpha\varphi\rangle} = \mathrm{e}^{-\langle \mu, G_t S_{1/n}^\alpha\varphi\rangle}, \quad \varphi \in \Phi. \tag{4.71}$$

But by uniqueness of the solutions to (4.66) and by (4.3) in the case $k = 0$, we get $G_t S_{1/n}^\alpha \varphi = \bar{v}_n(t)$. Consequently, (4.64) is true for $0 \leq t < \frac{1}{n}$.

Assume now that for some $k \geq 0$ the random measures X_t^n, $t \in [\frac{k}{n}, \frac{k+1}{n})$, are defined (not necessarily on a common probability space) and satisfy (4.64). Recall that this is true for $k = 0$. Then, for fixed $t \in [\frac{k+1}{n}, \frac{k+2}{n})$, conditionally on $X_{t-1/n}^n$, we set

$$X_t^n \ := \ S_{1/n}^\alpha Y_{1/n}, \quad \text{where } Y \text{ starts from } Y_0 := X_{t-1/n}^n. \tag{4.72}$$

Now (4.68) implies

$$\mathbf{P}\mathrm{e}^{-\langle X_t^n,\varphi\rangle} = \mathbf{P}\mathbf{P}\{\mathrm{e}^{-\langle Y_{1/n}, S_{1/n}^\alpha\varphi\rangle} \mid Y_0 = X_{t-1/n}^n\} = \mathbf{P}\mathrm{e}^{-\langle X_{t-1/n}^n, G_{1/n} S_{1/n}^\alpha\varphi\rangle}. \tag{4.73}$$

By the induction hypothesis, the chain of equations (4.73) can be continued with

$$= \mathrm{e}^{-\langle \mu, \bar{v}_n(t-1/n)\rangle}, \qquad \varphi \in \Phi, \tag{4.74}$$

but where $\bar{v}_n(0) = S_{1/n}^\alpha G_{1/n} S_{1/n}^\alpha \varphi$ instead of $S_{1/n}^\alpha\varphi$. However, by the constructions in the beginning of subsection 4.1, the new $\bar{v}_n(t - \frac{1}{n})$ coincides with the original $\bar{v}_n(t)$, yielding (4.64). Consequently, by induction we obtained random measures X_t^n satisfying (4.64) for any $t \geq 0$.

Next we want to let $n \uparrow \infty$. According to Proposition 4.3, for $t \geq 0$ fixed, $\bar{v}_n(t) \to v(t)$ as $n \uparrow \infty$, implying that the right-hand sides of (4.64) converge to $\langle \mu, v(t)\rangle$. Therefore, the log-Laplace transforms at the left-hand side of (4.64) converge to $\langle \mu, v(t)\rangle$, too. Now, by domination as in (3.2), we get $\langle \mu, v(t)\rangle \downarrow 0$ as $\varphi \downarrow 0$.

Therefore, the limit of the log-Laplace transforms in (4.64) is again a log-Laplace transform of a random measure in $\mathcal{M}$, say X_t (see, for instance, [Dyn94, section 3.3.4, pp. 50–51]; in other words, there is no loss of probability mass, that is, the laws of the random measures X_t^n are relatively compact). Consequently, for t fixed, $X_t^n \to X_t$ in law as $n \uparrow \infty$. Since the map $\mu \mapsto \langle \mu, V_t\varphi\rangle$ is measurable, via $\mu \mapsto X_t$ we get a probability kernel Q_t in $\mathcal{M}$ for the fixed t. From the semigroup property of $V\varphi$ it follows that the family $Q := \{Q_t : t > 0\}$ satisfies Chapman–Kolmogorov. Hence, Q is the transition kernel of a time-homogeneous Markov process in $\mathcal{M}$, which is the desired superprocess X. This finishes the proof of Theorem 4.4. □

Remark 4.6 (convergence theorem on càdlàg path space). The previous proofs should be modified and refined as the superprocess X is constructed as a Markov process via convergence of the one-, hence finite-dimensional, distributions of approximating processes X^n with càdlàg paths, say. Now the processes X^n and X should have finite moments of all orders $\theta \in (0, 1 + \beta)$. Using martingale methods, via Aldous's criterion it should be possible to prove tightness of the laws of the processes X^n on Skorohod path space, then sharpening Theorem 4.4 to a convergence theorem on path space with a limiting càdlàg superprocess X.

Acknowledgments. We would like to thank Sergio Albeverio, Zdzisław Brzeźniak, Werner Kirsch, and Karl-Theodor Sturm for discussions on the operators $\Delta^{(\alpha)}$. We are also very grateful to an anonymous referee for his very careful reading and suggestions for improving the exposition, in particular concerning the final construction of the superprocess.

REFERENCES

[ABD95] S. Albeverio, Z. Brzeźniak, and L. Dabrowski, *Fundamental solution of the heat and Schrödinger equations with point interaction*, J. Funct. Anal., 130 (1995), pp. 220–254.

[AGHKH88] S. Albeverio, F. Gesztesy, R. Høegh-Krohn, and H. Holden, *Solvable Models in Quantum Mechanics*, Springer-Verlag, New York, 1988.

[Daw93] D.A. Dawson, *Measure-valued Markov processes*, in École d'été de probabilités de Saint Flour XXI–1991, P.L. Hennequin, ed., Lecture Notes in Math. 1541, Springer-Verlag, Berlin, 1993, pp. 1–260.

[Del96] J.-F. Delmas, *Super-mouvement brownien avec catalyse*, Stochastics Stochastics Rep., 58 (1996), pp. 303–347.

[DF88] D.A. Dawson and K. Fleischmann, *Strong clumping of critical space-time branching models in subcritical dimensions*, Stochastic Process. Appl., 30 (1988), pp. 193–208.

[DF95] D.A. Dawson and K. Fleischmann, *Super-Brownian motions in higher dimensions with absolutely continuous measure states*, J. Theoret. Probab., 8 (1995), pp. 179–206.

[Dyn94] E.B. Dynkin, *An Introduction to Branching Measure-Valued Processes*, AMS, Providence, RI, 1994.

[EF00] J. Engländer and K. Fleischmann, *Extinction properties of super-Brownian motions with additional spatially dependent mass production*, Stochastic Process. Appl., 88 (2000), pp. 37–58.

[EP94] S.N. Evans and E.A. Perkins, *Measure-valued branching diffusions with singular interactions*, Canad. J. Math., 46 (1994), pp. 120–168.

[ET02] J. Engländer and D. Turaev, *A scaling limit theorem for a class of superdiffusions*, Ann. Probab., 30 (2002), pp. 683–722.

[Eth00] A.M. Etheridge, *An Introduction to Superprocesses*, Univ. Lecture Ser. 20, AMS, Providence, RI, 2000.

[FK99] K. Fleischmann and A. Klenke, *Smooth density field of catalytic super-Brownian motion*, Ann. Appl. Probab., 9 (1999), pp. 298–318.

[FV04] K. Fleischmann and P. Vogt, *Long-Term Behaviour of Super-Brownian Motion with a Single Point Source*, WIAS Berlin, 2004, preprint (in preparation).

[Kle00] A. Klenke, *Absolute continuity of catalytic measure-valued branching processes*, Stochastic Process. Appl., 89 (2000), pp. 227–237.

[Lam67] J. Lamperti, *Continuous state branching processes*, Bull. Amer. Math. Soc., 73 (1967), pp. 382–386.

[Leb65] N.N. Lebedev, *Special Functions and Their Applications*, revised English ed., translated and edited by R. A. Silverman, Prentice-Hall, Englewood Cliffs, NJ, 1965.

[LG99] J.-F. Le Gall, *Spatial Branching Processes, Random Snakes and Partial Differential Equations*, Birkhäuser Verlag, Basel, 1999.

[Per02] E.A. Perkins, *Dawson-Watanabe superprocesses and measure-valued diffusions*, in École D'été de Probabilités de Saint Flour XXIX–1999, P. Bernard, ed., Lecture Notes in Math., Springer-Verlag, Berlin, 2002, pp. 125–329.

[Tra69] C. J. Tranter, *Bessel Functions with Some Physical Applications*, Hart Publishing, New York, 1969.

SIAM J. MATH. ANAL.
Vol. 36, No. 3, pp. 773–795

RANDOM SAMPLING OF MULTIVARIATE TRIGONOMETRIC POLYNOMIALS*

RICHARD F. BASS† AND KARLHEINZ GRÖCHENIG‡

Abstract. We investigate when a trigonometric polynomial p of degree M in d variables is uniquely determined by its sampled values $p(x_j)$ on a random set of points x_j in the unit cube (the "sampling problem for trigonometric polynomials") and estimate the probability distribution of the condition number for the associated Vandermonde-type and Toeplitz-like matrices. The results provide a solid theoretical foundation for some efficient numerical algorithms that are already in use.

Key words. sampling, band-limited functions, multivariate trigonometric polynomials, random sampling, block Toeplitz matrix, Vandermonde matrix, condition number, metric entropy

AMS subject classifications. 94A12, 94A20, 15A12, 15A52, 42B99, 42A15, 42A61, 60G50, 60G99

DOI. 10.1137/S0036141003432316

1. Introduction.

The reconstruction, interpolation, or approximation of a function (signal, image) from a given data set is a central task in many problems of data processing. The mathematical problem is to find a function $f(x)$ in a suitable function space V that interpolates or approximates the given data $y_j = f(x_j)$. The set $\mathcal{X} = \{x_j : j = 1, \dots, r\} \subseteq \mathbb{R}^d$ is the sampling set, and the function space V comes from the mathematical modeling of signals or images (e.g., band-limitedness, smoothness). The numerical and theoretical analysis of the sampling problem depends, of course, heavily on the signal model V.

In this paper we focus almost exclusively on multivariate trigonometric polynomials as our model. While this is by no means the only possible model, it is convenient, interesting, and occurs in many applications where standard uniform sampling is not possible. Specifically, the model of trigonometric polynomials has been used in cardiology (one-dimensional) [39]; geophysics (two-dimensional) [30]; image processing (two-dimensional) [37]; as a nonuniform discrete Fourier transform (one- and two-dimensional) [8, 14, 15, 29, 35]; and in computer tomography (two- and three-dimensional) [3, 28, 33]. Furthermore the space of trigonometric polynomials of fixed degree is the appropriate finite-dimensional model for the approximation of band-limited functions from a finite number of samples [20, 21].

Clearly, the sampling operator $f \to \{f(x_j) : j = 1, \dots, r\}$ is linear, and, for a finite-dimensional model space, it can therefore be described by a matrix. For the model of trigonometric polynomials of fixed degree, this matrix possesses an additional structure; namely, it is either a rectangular Vandermonde-like matrix or a square Toeplitz-like matrix. This structure is the basis for efficient and fast numerical algorithms. For dimension $d = 1$ we refer to [8, 16, 17, 31, 38], and for higher dimensions to [28, 30, 33, 37]. These algorithms are fast, stable, and robust, but only in dimension $d = 1$ do the numerical algorithms possess a solid theoretical basis

*Received by the editors July 28, 2003; accepted for publication (in revised form) January 30, 2004; published electronically September 24, 2004.
http://www.siam.org/journals/sima/36-3/43231.html

†Department of Mathematics, The University of Connecticut, Storrs, CT 06269-3009 (bass@math.uconn.edu). This author was partially supported by NSF grant DMS-0244737.

‡Institute of Biomathematics and Biometry, GSF - National Research Center for Environment and Health, Ingolstädter Landstrasse 1, 85764 Neuherberg, Germany (karlheinz.groechenig@gsf.de).

(invertibility, estimates of condition numbers, and rates of convergence for iterative algorithms).

In higher dimensions, there is only numerical evidence that the existing algorithms work; except for some isolated results [19, 23] there has been no theoretical justification for the success of these numerical methods. The main reason for this disparity lies in the nature of zero sets of trigonometric polynomials in one and higher dimensions. In dimension $d = 1$ the zero set of a trigonometric polynomial is finite by the fundamental theorem of algebra, whereas the zero set of a trigonometric polynomial in several variables is an algebraic variety. This difference makes it almost impossible to determine effectively whether the reconstruction problem $\{f(x_j)\} \to f$ is solvable for a fixed multidimensional sampling set $\mathcal{X} \subseteq \mathbb{R}^d$. It seems even more difficult to estimate the condition numbers of the associated matrices. On the other hand, numerical experiments and successful applications make it plausible that for generic sampling sets $\mathcal{X} \subseteq \mathbb{R}^d$ the sampling problem is solvable and well-conditioned.

Our goal is to achieve some understanding for the success of existing numerical methods and to provide more insight into the theoretical issues. To do this we adopt a probabilistic point of view: Instead of seeking analytic statements for a fixed sampling set, we consider the collection of all sampling sets of size r and assume that the *sampling set consists of a finite sequence of independent random variables.* Instead of worst-case estimates, i.e., inequalities within mathematical analysis, we will seek probabilistic estimates (from the realm of probability theory). With this underlying philosophy, we will pursue the following objectives:

(a) We seek to explain and predict the performance of the existing numerical algorithms.

(b) We estimate the distribution of the condition numbers of the associated Vandermonde-like and Toeplitz-like matrices.

(c) We investigate the asymptotic behavior of condition numbers as the number of samples r tends to infinity.

The randomization of the sampling points seems to be a new idea in the investigation of numerical sampling algorithms. So far random sampling has been investigated by Seip and Ulanovskii [32], Chistyakov and Lyubarskii [9], Chistyakov, Lyubarskii, and Pastur [10] for entire functions of exponential type of one complex variable. These results rely on the deep characterization of deterministic sampling sets [26, 27] and, to our knowledge, cannot be extended to higher dimensions. In a different direction, Smale and Zhou [34] have recently used probabilistic methods from learning theory [12] to investigate sampling in reproducing kernel Hilbert spaces.

By contrast, our main contribution is to sampling theory for functions of several variables. In higher dimensions there is currently no satisfactory deterministic theory, and our analysis provides the first clues that existing algorithms and methods do really work. From a more applied point of view, our results suggest that random sampling of images or higher-dimensional objects may be a successful strategy to capture the essential information of multidimensional objects while preserving numerical efficiency and stability.

Description of results. We now describe the main results.

Let $\mathcal{P}_M$ be the space of trigonometric polynomials on $\mathbb{R}^d$ of degree M and period 1, that is, $\mathcal{P}_M$ consists of all functions on $\mathbb{R}^d$ of the form

$$p(x) = \sum_{k \in [-M,M] \cap \mathbb{Z}^d} a_k e^{2\pi i k \cdot x}. \tag{1}$$

Note that the (distributional) Fourier transform of $p \in \mathcal{P}_M$ is $\hat{p} = \sum_{k\in[-M,M]\cap\mathbb{Z}^d} a_k\delta_k$, so supp $\hat{p} \subseteq [-M,M]^d$. The parameter M can be interpreted as the "bandwidth," and indeed trigonometric polynomials have been shown to be the appropriate finite-dimensional model for band-limited functions [20, 21].

Now assume that the samples $p(x_j), j = 1,\dots,r$, of some trigonometric polynomial $p \in \mathcal{P}_M$ are given for some sampling set $\mathcal{X} = \{x_j : j = 1,\dots,r\}$. By our normalization, we may assume that the sampling set $\mathcal{X}$ is contained in the unit cube $[0,1]^d$. Our goal is to reconstruct or to approximate p. Equivalently, we want to determine the coefficients a_k of p from the samples $p(x_j)$. This task can be seen as a nonuniform discrete Fourier transform and is a frequent task in data processing [8, 14, 15, 29, 35].

In its simplest form, the reconstruction of p amounts to solving the r equations

$$\sum_{k\in[-M,M]^d\cap\mathbb{Z}^d} a_k e^{2\pi i k\cdot x_j} = p(x_j) = y_j, \quad j = 1,\dots,r,$$

for the coefficient vector $\mathbf{a} = (a_k)_{k\in\mathbb{Z}^d\cap[-M,M]^d}$. This system of equations can be written in matrix form as

$$\mathcal{U}\mathbf{a} = \mathbf{y}, \tag{2}$$

where $\mathcal{U}$ is the matrix with entries $\mathcal{U}_{jk} = e^{2\pi i k\cdot x_j}, k \in \mathbb{Z}^d \cap [-M,M]^d, j = 1,\dots,r$, and $\mathbf{y}$ is the target vector $\mathbf{y} = (y_j)_{j=1,\dots,r}$. Alternatively, one may try to find $\mathbf{a}$ from the normal equations [18]

$$\mathcal{U}^*\mathcal{U}\mathbf{a} = \mathcal{U}^*\mathbf{y}. \tag{3}$$

In this case the matrix $\mathcal{T} = \mathcal{U}^*\mathcal{U}$ has entries

$$\mathcal{T}_{kl} = \sum_{j=1}^{r} e^{-2\pi i(k-l)\cdot x_j}, \quad k,l \in [-M,M]^d \cap \mathbb{Z}^d.$$

The matrices of these linear systems are highly structured, $\mathcal{U}$ is a *Vandermonde-like matrix*, and $\mathcal{T}$ is a positive semidefinite $D \times D$ matrix with a *block Toeplitz structure*. Both structures have been successfully exploited for fast numerical algorithms [17, 23, 29, 37].

However, before the numerical analysis of the sampling problem can be undertaken, we need to settle a fundamental theoretical issue: Is either of the equations (2) or (3) solvable? Note that both matrices $\mathcal{U}$ and $\mathcal{T}$ depend on the sampling points x_j as parameters. Therefore we ask more precisely, *for which sampling set $\mathcal{X}$ does $\mathcal{U}$ have full rank, or, equivalently, when is $\mathcal{T}$ invertible?*

In dimension $d = 1$, $\mathcal{T}$ is invertible if and only if $r \geq 2M + 1$ (the number of sampling points is greater than the dimension of the space). In higher dimensions no criterion for the invertibility of $\mathcal{T}$ is known, and useful results are sparse. See [23] for a discussion.

In the spirit of probability theory we model the sampling set as a sequence of independent, identically distributed (i.i.d.) random variables in $[0,1]^d$. This means that we treat the sampling points as a sequence of functions $x_j = x_j(\omega)$ on some probability space $(\Omega, \mathbb{P})$. Thus the matrices $\mathcal{U}$ and $\mathcal{T}$ are now random matrices, and their determinants, eigenvalues, and singular values are random variables on $(\Omega, \mathbb{P})$ that depend on the sampling set in a rather complicated way.

The first theorem guarantees the generic invertibility of $\mathcal{T}$.

THEOREM 1.1. *Assume that the finite sequence of random variables $x_1, \dots, x_r$ satisfies the following properties:*

(a) $r \geq (2M+1)^d$.

(b) *The x_j's are independent.*

(c) *The distribution μ_j of each x_j is absolutely continuous with respect to Lebesgue measure on $[0,1]^d$.*

Then with probability one the Toeplitz-like matrix $\mathcal{T}$ is invertible.

Estimates for the condition number. For a stable numerical solution of either of the systems (2) and (3) we need effective invertibility of $\mathcal{T}$. This is usually measured by the condition number $\kappa(\mathcal{T})$ of $\mathcal{T}$. (The condition number $\kappa(M)$ of a rectangular matrix is the ratio of largest to smallest singular value [18]; for a positive-definite square matrix, this is simply the ratio of the largest to the smallest eigenvalue.) To estimate the condition numbers of $\mathcal{U}$ and $\mathcal{T}$ we observe that

$$\sum_{j=1}^{r} |p(x_j)|^2 = \langle \mathbf{y}, \mathbf{y} \rangle = \langle \mathcal{U}\mathbf{a}, \mathcal{U}\mathbf{a} \rangle = \langle \mathcal{U}^*\mathcal{U}\mathbf{a}, \mathbf{a} \rangle = \langle \mathcal{T}\mathbf{a}, \mathbf{a} \rangle. \tag{4}$$

Consequently, if we can prove an inequality of the form

$$A\|p\|_2^2 \leq \sum_{j=1}^{r} |p(x_j)|^2 \leq B\|p\|_2^2 \quad \forall p \in \mathcal{P}_M, \tag{5}$$

then the largest (smallest) eigenvalue of $\mathcal{T}$ is at most B (at least A), since $\|p\|_2 = \|\mathbf{a}\|_2$. Consequently, (5) implies the estimates

$$\kappa(\mathcal{T}) \leq \frac{B}{A} \quad \text{and} \quad \kappa(\mathcal{U}) \leq \Big(\frac{B}{A}\Big)^{1/2}. \tag{6}$$

Our main theorem is the following asymptotic estimate for the condition numbers of $\mathcal{T}$ or $\mathcal{U}$ as $r \to \infty$.

THEOREM 1.2. *Assume that $\mathcal{X} = \{x_j : j \in \mathbb{N}\}$ is a sequence of i.i.d. random variables uniformly distributed over $[0,1]^d$. There exist constants $A, B > 0$ depending only on the bandwidth M and the dimension d such that for any $\mu \in (0,1)$, the sampling inequality*

$$(1-\mu)r\|p\|_2^2 \leq \sum_{j=1}^{r} |p(x_j)|^2 \leq (1+\mu)r\|p\|_2^2 \quad \forall p \in \mathcal{P}_M \tag{7}$$

holds with probability at least

$$1 - Ae^{-Br\frac{\mu^2}{1+\mu}}.$$

Consequently, with the same probability estimate, the Toeplitz-type matrix $\mathcal{T}$ has condition number $\kappa(\mathcal{T}) \leq \frac{1+\mu}{1-\mu}$ and the Vandermonde-like matrix $\mathcal{U}$ has condition number $\kappa(\mathcal{U}) \leq \sqrt{1+\mu}/\sqrt{1-\mu}$.

For a fixed threshold $\theta > 1$, the probability that $\kappa(\mathcal{T}) \leq \theta$ converges to 1 exponentially fast as the number of samples increases. With some poetic license, we may therefore say that *oversampling improves the condition number.*

We will give two proofs of this result. The first proof is by reduction to a deterministic result. We estimate the probability that the conditions of an existing deterministic result from [23] are satisfied. With this approach we obtain explicit estimates for the

constants. The second proof uses a version of the powerful metric entropy method; see [4, 5, 13] for just a few of its applications to probability theory. This approach is genuinely asymptotic and does not yield effective estimates of the constants. The main advantage of this method is its flexibility and generality. To demonstrate the power of this approach we will formulate versions of Theorem 1.2 for ordinary polynomials in several variables, for almost periodic functions, and for spherical harmonics on the sphere (section 6).

As a consequence of Theorem 1.2 we obtain the following law of the iterated logarithm.

COROLLARY 1.3. *If $\{x_j : j \in \mathbb{N}\}$ is a sequence of i.i.d. random variables that are uniformly distributed over $[0,1]^d$, then*

$$\limsup_{r\to\infty} \frac{\sup_{p\in\mathcal{P}} |\sum_{j=1}^{r} [|p(x_j)|^2 - \|p\|_2^2]|}{\sqrt{r \log\log r}\,\|p\|_2^2} = c \quad a.s. \tag{8}$$

for some positive constant c of order $D = (2M+1)^d$.

With less precision, but more intuitively, the corollary says that with probability one, the condition number of the sampling problem is

$$\kappa(\mathcal{T}) \le (r + c\sqrt{r\log\log r})/(r - c\sqrt{r\log\log r}) \approx 1 + 2c\left(\frac{\log\log r}{r}\right)^{1/2},$$

whenever r is large enough.

Our main theorems validate existing numerical algorithms for nonuniform sampling sets in higher dimensions. Furthermore, they make precise in which sense random sampling of multidimensional objects is better than deterministic sampling.

The paper is organized as follows. In section 2 we collect some facts about multivariate trigonometric polynomials and explain the idea of the simplest numerical algorithms. In section 3 we prove Theorem 1.1 about the almost certain solvability of the sampling problem. In section 4 we provide the first proof of Theorem 1.2 and show a probabilistic covering result that may be of independent interest. In section 5 we develop the metric entropy approach and give a second proof of Theorem 1.2 for the asymptotic estimate of the condition number. Furthermore, we develop some consequences of our main theorem. In section 6 we discuss extensions of the metric entropy method to other sampling problems.

2. Sampling of trigonometric polynomials. We first collect the background information on sampling of trigonometric polynomials and some of the numerical aspects that motivated our investigation.

By $\mathcal{X} = \{x_j : j = 1, \dots, r\}$ we denote a sampling set of r (distinct) points in $[0,1]^d$.

The space of trigonometric polynomials on $\mathbb{R}^d$ of degree M and period 1 in each variable is

$$\mathcal{P}_M = \left\{ p : p(x) = \sum_{k\in[-M,M]^d\cap\mathbb{Z}^d} a_k e^{2\pi i k\cdot x} \right\}. \tag{9}$$

Remarks. (1) The vector space $\mathcal{P}_M$ has dimension $D = (2M+1)^d$. This implies that we need at least $(2M+1)^d$ data points in order to recover a polynomial $p \in \mathcal{P}_M$.

(2) The parameter M can be interpreted as the "bandwidth" and measures the permissible amount of oscillation (smoothness). We will assume that M is given, but

note that the determination of the optimal bandwidth is an important step in the practical application of sampling algorithms [38].

(3) On $\mathcal{P}_M$ the following estimates between equivalent norms hold:

$$\|p\|_2^2 = \int_{[0,1]^d} |p(x)|^2 \, dx = \|\mathbf{a}\|_2,$$

$$\|p\|_\infty \le D^{1/2} \, \|\mathbf{a}\|_2 = D^{1/2} \, \|p\|_2, \tag{10}$$

$$\|p\|_4^4 \le \|p\|_\infty^2 \|p\|_2^2 \le D\|p\|_2^4.$$

The reconstruction of $p \in \mathcal{P}_M$ from given samples $\{p(x_j) : j = 1, \dots, r\}$ amounts to solving the following system of r equations:

$$\sum_{k \in [-M,M]^d \cap \mathbb{Z}^d} a_k e^{2\pi i k \cdot x_j} = f(x_j) = y_j, \quad j = 1, \dots, r. \tag{11}$$

Introducing the matrices $\mathcal{U}$ and $\mathcal{T}$ with entries

$$U_{jk} = e^{2\pi i k \cdot x_j}, \quad j = 1, \dots, r, k \in [-M, M]^d \cap \mathbb{Z}^d, \tag{12}$$

$$\mathcal{T}_{kl} = (\mathcal{U}^* \mathcal{U})_{kl} = \sum_{j=1}^{r} e^{-2\pi i (k-l) \cdot x_j}, \quad k, l \in [-M, M]^d \cap \mathbb{Z}^d, \tag{13}$$

we can then formulate the sampling problem for $\mathcal{P}_M$ in several distinct ways.

LEMMA 2.1. *The following are equivalent:*

(i) *The equations* (11) *possess a unique solution in* $\mathcal{P}_M$.

(ii) *The Vandermonde-type matrix has full rank and* $r \ge D$.

(iii) *There exist* $A, B > 0$ *such that*

$$A\|\mathbf{a}\|_2 \le \|\mathcal{U}\mathbf{a}\|_2 \le B\|\mathbf{a}\|_2 \quad \forall \mathbf{a} \in \mathbb{C}^D.$$

(iv) *The* $D \times D$ *Toeplitz-like matrix* $\mathcal{T}$ *is invertible.*

(v) *There exist* $A, B > 0$ *such that*

$$A\|p\|_2 \le \sum_{j=1}^{r} |p(x_j)|^2 \le B\|p\|_2 \quad \forall p \in \mathcal{P}_M. \tag{14}$$

If any of (i)–(v) hold, we say that $\mathcal{X}$ is a *set of stable sampling* for $\mathcal{P}_M$ [25].

Despite its lack of mathematical substance, this lemma is useful because each of the criteria may be used as a starting point for the theoretical or numerical investigation of the sampling problem. For the mathematical analysis the *sampling inequality* (14) is most appropriate, because it invites the use of analytic methods. For the numerical solution of the sampling problem, the linear algebra criteria (ii), (iii), and (iv) are most useful, because the theory of structured matrices offers fast solution techniques.

A numerical algorithm for the solution of (11) could then be based on the following steps.

ALGORITHM.

Input. Given a sampling set $\mathcal{X} = \{x_j : j = 1, \dots, r\} \subseteq [0,1]^d$ and a data vector $\mathbf{y} = \{y_j : j = 1, \dots, r\}$, assume that $\mathcal{T}$ defined in (13) is invertible.

Step 1. Compute $\mathbf{b} = \mathcal{U}^*\mathbf{y}$, i.e.,

$$b_k = \sum_{j=1}^{r} e^{-2\pi i k\cdot x_j}\, y_j \quad \text{for } k \in [-M, M]^d \cap \mathbb{Z}^d. \tag{15}$$

Step 2. Solve the system of equations

$$\mathbf{a} = \mathcal{T}^{-1}\mathbf{b}. \tag{16}$$

Step 3. Compute $p \in \mathcal{P}_M$ by

$$p(x) = \sum_{k\in[-M,M]^d\cap\mathbb{Z}^d} a_k e^{2\pi i k\cdot x}. \tag{17}$$

Then p is the (unique) least squares approximation of the given data vector $\mathbf{y}$ in the sense that

$$\sum_{j=1}^{r} |y_j - p(x_j)|^2 = \min_{q\in\mathcal{P}_M} \sum_{j=1}^{r} |y_j - q(x_j)|^2. \tag{18}$$

If $\mathbf{y}$ arises as the sampled vector of a polynomial $p \in \mathcal{P}_M$, i.e., $y_j = p(x_j)$, then this algorithm provides the exact reconstruction of p.

Remarks. (1) The numerical implementation of this idea is often referred to as the ACT-algorithm. The decisive step is the solution of matrix equation $\mathcal{T}\mathbf{a} = \mathbf{b}$ in Step 2. Since $\mathcal{T}$ is a positive-definite Toeplitz-like matrix, the exploitation of this structure in conjunction with block Toeplitz solvers and conjugate gradient algorithms has led to fast and efficient reconstruction algorithms in higher dimensions [30, 37]. For numerical issues and real applications we refer to [23].

(2) Since the condition numbers of $\mathcal{U}$ and $\mathcal{T}$ are related by $\kappa(\mathcal{T}) = \kappa(\mathcal{U})^2$, it may be better to solve the Vandermonde-type system $\mathcal{U}\mathbf{a} = \mathbf{y}$ directly; see the work of Potts and Steidl [28].

3. Invertibility almost surely. We first establish that the reconstruction algorithm discussed in section 2 works almost surely. In dimension $d = 1$, $\mathcal{T}$ is invertible if and only if $r \geq 2M + 1$. In higher dimensions, a complete and effective characterization of the invertibility seems out of reach. For this reason we use a probabilistic approach.

First we provide a lemma in which λ will denote Lebesgue measure.

LEMMA 3.1. *Let $p \in \mathcal{P}_M$ be a trigonometric polynomial in d variables. Then its zero set $\mathcal{Z}(p) = \{x \in [0,1]^d : p(x) = 0\}$ has Lebesgue measure* 0.

Proof. This fact is well known; we provide its easy proof for the sake of completeness.

Fix $x_1, \ldots, x_{d-1} \in [0,1]^d$; then $P(x_1, \ldots, x_{d-1}, x_d)$ is a trigonometric polynomial in one variable x_d of degree M and thus has at most $2M + 1$ zeros. The set $\{x \in [0,1] : (x_1, \ldots, x_{d-1}, x) \in \mathcal{Z}(p)\}$ has Lebesgue measure 0. This is true for every choice of $x_1, \ldots, x_{d-1}$, so by Fubini's theorem, we obtain that

$$\lambda(\mathcal{Z}(p)) = \int_{[0,1]^{d-1}} \Big(\int_{[0,1]} \chi_{\mathcal{Z}(p)}(x_1, \ldots, x_{d-1}, x)\, dx \Big)\, dx_1 \ldots dx_{d-1} = 0\,,$$

as desired. ∎

The following result is a first indication why in practice no serious problems have occurred in the application of multidimensional sampling algorithms.

THEOREM 3.2. *Assume that the random variables $\{x_1, \dots, x_r\}$ are independent and that the distribution μ_j of each x_j is absolutely continuous with respect to Lebesgue measure on $[0,1]^d$.*

Then the Vandermonde-like matrix $\mathcal{U}$ is of full rank almost surely. If, in addition, $r \geq D = (2M+1)^d$, then the Toeplitz-like matrix $\mathcal{T} = \mathcal{U}^*\mathcal{U}$ is invertible almost surely.

Proof. Let $m_1, \dots, m_D$ be an enumeration of the index set $[-M, M] \cap \mathbb{Z}^d$ over which we are summing, and let C_N be the $N \times N$ matrix with entries

$$C_{\ell j} = e^{im_\ell \cdot x_j}, \quad 1 \leq \ell,\ j \leq N.$$

Then C_N depends on the sampling points $x_1, \dots x_N$, and we may define the "bad" set

$$\mathcal{B}_N = \{(x_1, \dots, x_N) \in ([0,1]^d)^N : \det C_N = 0\}.$$

We claim that $\lambda(\mathcal{B}_N) = 0$ for all $N \leq \min(r, D)$ and prove this by induction over N. This is certainly true for $N = 1$. So assume that $N < \min(r, D)$ and that $(x_1, \dots, x_N) \notin \mathcal{B}_N$.

Let $a_\ell = (C_{\ell,1}, \dots, C_{\ell,N})$, $\ell \leq N$, be the ℓth row of C_N and let $a_{N+1} = (C_{N+1,1}, \dots, C_{N+1,N})$. Since C_N is invertible, there exist coefficients $b_\ell = b_\ell(x_1, \dots, x_N) \in \mathbb{C}$, not all 0, such that

$$a_{N+1} = b_1 a_1 + \cdots + b_N a_N\,.$$

By looking at the $(N+1)$st column of C_{N+1}, we find that C_{N+1} is invertible if and only if $C_{N+1,N+1} \neq b_1 C_{1,N+1} + \cdots + b_N C_{N,N+1}$, or if and only if

$$e^{im_{N+1} \cdot x_{N+1}} \neq b_1 e^{im_1 \cdot x_{N+1}} + \cdots + b_N e^{im_N \cdot x_{N+1}}.$$

In other words, C_{N+1} is invertible if x_{N+1} is *not* in the set

$$D_N = D_N(x_1, \dots, x_N) = \{x \in [0,1]^d : e^{im_{N+1} \cdot x} = b_1 e^{im_1 \cdot x} + \cdots + b_N e^{im_N \cdot x}\}\,.$$

For fixed $(x_1, \dots, x_N) \in ([0,1]^d)^N$, D_N is the zero set of some trigonometric polynomial, and by Lemma 3.1 D_N has Lebesgue measure 0 in $[0,1]^d$.

Since the bad set $\mathcal{B}_{N+1}$ is contained in $\{(x_1, \dots, x_N, x_{N+1}) \in ([0,1]^d)^{N+1} : x_{N+1} \in D_N(x_1, \dots, x_N)\}$, we see by Fubini's theorem that

$$\begin{aligned}\lambda(\mathcal{B}_{N+1}) &= \int_{([0,1]^d)^N} \left(\int_{[0,1]^d} \chi_{B_{N+1}}(x_1, \dots, x_N, x_{N+1})\, dx_{N+1} \right) dx_1, \dots, dx_N \\ &\leq \int_{([0,1]^d)^N} \int_{[0,1]^d} \lambda(D_N(x_1, \dots, x_N)) dx_1, \dots, dx_N = 0.\end{aligned}$$

The induction step is proved.

If $r \leq D$, then C_r is invertible for almost every choice of $x_1, \dots, x_D$, where "almost every" is with respect to Lebesgue measure λ. Consequently, the $r \times D$ matrix $\mathcal{U}$ has full rank. If $r \geq D$, this also implies that the $D \times D$ square matrix $\mathcal{T} = \mathcal{U}^*\mathcal{U}$ is invertible for almost every choice of $x_1, \dots, x_D$.

Since the distribution μ_j of x_j is absolutely continuous with respect to λ, the bad set $\mathcal{B}_D$ also has measure 0 with respect to $\mu_1 \times \cdots \times \mu_D$. ∎

COROLLARY 3.3. *The Toeplitz-like matrix $\mathcal{T}$ is invertible under each of the following hypotheses on the sampling set:*

(a) *The $x_j, j = 1, \dots, r$, are i.i.d. random variables, each of which is uniformly distributed over $[0,1]^d$.*

(b) *The sampling set is a random perturbation of a uniform sampling set, i.e., it is some enumeration of $\{\frac{1}{N}k + \delta_k : k \in \mathbb{Z}^d \cap [0, N-1]^d\}$, where $N \geq 2M+1$ and the δ_k are i.i.d. random variables uniformly distributed over a neighborhood of 0.*

4. A covering results and reduction to deterministic estimates. Theorem 1.1 guarantees that an implementation of the algorithm in section 2 will work in principle. However, numerical invertibility requires a reasonable bound on the condition number of $\mathcal{T}$ or of $\mathcal{U}$.

This is already a serious problem in dimension $d = 1$. It is easy to construct sampling sets in $[0,1]$ for which the corresponding Toeplitz matrix has condition number of the order 10^{15} [17]. While such a matrix is invertible in theory, for practical purposes it may be considered to be noninvertible.

As a next step we therefore turn to estimates for the condition number of the block Toeplitz matrix $\mathcal{T}$. For this we combine a deterministic result with a probabilistic statement on coverings.

We work with the metric $d(x,y) = \min_{k \in \mathbb{Z}^d} \|x - y + k\|_\infty$ on the torus $\mathbb{T}^d \sim [0,1)^d$ and the associated cubes of side-length 2ρ,

$$B(x,\rho) = \{y \in [0,1]^d : d(y,x) \leq \rho\} = x + [-\rho, \rho]^d.$$

To every sequence of sampling points $\{x_j : j \in \mathbb{N}\} \subseteq [0,1]^d$, let $\{V_j\}$ we assign the "distance function"

$$\delta(r) = \inf \left\{ s : \bigcup_{j=1}^{r} B(x_i, s) \supset [0,1]^d \right\}. \tag{19}$$

The quantity $2\delta(r)$ can be interpreted as the maximum distance of any of the first r sampling points x_j to its next neighbor. Let $V_j, j = 1, \dots, r$, be Voronoi regions

$$V_j = \{y \in [0,1]^d : d(y, x_j) \leq d(y, x_k), k \neq j, 1 \leq j, k \leq r\}$$

and $w_j = \lambda(V_j)$ and consider the weighted Toeplitz-like matrix $\mathcal{T}^w$ with entries

$$\mathcal{T}^w_{kl} = (\mathcal{U}^* \mathcal{U})_{kl} = \sum_{j=1}^{r} w_j e^{-2\pi i (k-l)\cdot x_j}, \quad k, l \in [-M, M]^d \cap \mathbb{Z}^d.$$

Then it is possible to show the following deterministic theorem [19, 23].

THEOREM 4.1. *If*

$$\delta(r) < \frac{\log 2}{2\pi M d}, \tag{20}$$

then, for all $p \in \mathcal{P}_M$,

$$(2 - e^{2\pi d M \delta})^2 \|p\|_2^2 \leq \sum_{j=1}^{r} |p(x_j)|^2 w_j = \langle a, \mathcal{T}^w a \rangle \leq 4 \|p\|_2^2. \tag{21}$$

Consequently, the condition number of $\mathcal{T}^w$ can be estimated by

$$\kappa(\mathcal{T}^w) \leq \frac{4}{(2 - e^{2\pi dM\delta})^2}, \tag{22}$$

and both $\mathcal{T}$ and $\mathcal{T}^w$ are invertible.

Remarks. (1) The specific choice of weights w_j is crucial for the explicit estimate (22). In the numerical implementation of the algorithm of section 2, they serve as a simple and cheap preconditioner.

(2) In higher dimensions, (22) is far from being optimal, since it depends on the dimension d. It is an open problem to obtain improvements to this estimate. For a related result for band-limited functions, see [7].

We next suppose that the sampling points form an infinite sequence of i.i.d. independent random variables $x_j, j \in \mathbb{N}$. We first investigate how the distribution of the associated sequence of random variables $\delta(r)$ depends on the number of sampling points r.

THEOREM 4.2. *If $X = \{x_j : j \in \mathbb{N}\}$ is a sequence of i.i.d. random variables uniformly distributed over $[0,1]^d$, then for every $r, N \in \mathbb{N}$*

$$\mathbb{P}(\delta(r) > 1/N) \leq N^d(1 - N^{-d})^r \leq N^d e^{-r/N^d}. \tag{23}$$

Consequently, $\kappa(\mathcal{T}^w) \leq 4(2 - e^{2\pi Md/N})^{-2}$ and both $\mathcal{T}^w$ and $\mathcal{T}$ are invertible with probability at least

$$1 - N^d(1 - N^{-d})^r \geq 1 - N^d e^{-r/N^d}.$$

Proof. Divide $[0,1]^d$ into N^d disjoint subcubes of side-length $1/N$, i.e., $[0,1]^d = \bigcup_{j=1}^r B(c_j, \frac{1}{2N})$, where the c_j are the centers of these subcubes. Note that if a subcube contains a point x_j, then that subcube is contained in $B(x_j, 1/N)$. So if each of these subcubes contains at least one of the x_j, we conclude $\delta(r) \leq 1/N$.

Since the $x_j, j = 1, \ldots, r$, are chosen independently and uniformly, the number of x_j's in any cube is a binomial random variable. Thus the probability that a particular subcube is empty is

$$(1 - N^{-d})^r$$

(since N^{-d} is the probability that any particular x_j is in this subcube and there are r points). Since there are N^d subcubes altogether, the probability that at least one of the subcubes is empty is bounded by

$$N^d(1 - N^{-d})^r. \tag{24}$$

If $\delta(r) > 1/N$, then at least one of the subcubes must be empty, which proves the left-hand inequality of (23). The right-hand side follows from the obvious inequality $(1 - N^{-d})^r = e^{r\log(1-N^{-d})} \leq e^{-r/N^d}$.

The estimate for the condition number of $\mathcal{T}^w$ and the invertibility of $\mathcal{T}$ now follow from Theorem 4.1. □

Remark. For (20) we need that $\frac{1}{N} < \frac{\log 2}{2\pi Md}$; this means that we need at least $r = N^d \geq \left(\frac{2\pi Md}{\log 2}\right)^d \approx \left(\frac{\pi d}{\log 2}\right)^d D$ sampling points before Theorem 4.2 becomes effective.

Next we derive an asymptotic result for $\delta(r)$, which may be of independent interest.

THEOREM 4.3. *Assume that* $\{x_j : j \in \mathbb{N}\}$ *is a sequence of i.i.d. points uniformly distributed in* $[0,1]^d$. *Then*

$$\limsup_{r\to\infty} \frac{\delta(r)}{(\log r/r)^{1/d}} = c \quad \text{a.s.} \tag{25}$$

for some constant $c \in [\frac{1}{4}, 2^{1+1/d}]$.

Thus for r sampling points the maximum distance to the nearest neighbor is roughly $(\log r/r)^{1/d}$. For comparison, for the $r = N^d$ equispaced points $\{\frac{k}{N} : k \in [0,N]\cap\mathbb{Z}^d\}$, we have $\delta(r) = \frac{1}{2N} = \frac{1}{2}r^{-1/d}$. For r randomly distributed points we need an additional logarithmic term.

Proof of Theorem 4.3.

Step 1. We first show that

$$\limsup_{r\to\infty} \frac{\delta(r)}{(\log r/r)^{1/d}} \le 2^{1+1/d} \quad \text{a.s.} \tag{26}$$

Choose $r_k = 2^k$ as the number of points, and let N_k be the greatest integer less than $(\frac{r_k}{2\log r_k})^{1/d}$. We divide $[0,1]^d$ into N_k^d disjoint subcubes of side-length N_k^{-1}. Let A_k be the event that at least one of the subcubes contains none of the $x_j, j = 1,\dots,r_k$. By (24) we have

$$\mathbb{P}(A_k) \le N_k^d e^{-r_k/N_k^d} \le \frac{r_k}{2\log r_k} e^{-2\log r_k} = \frac{1}{2r_k\log r_k} = \frac{1}{2^{k+1}k\log 2}. \tag{27}$$

Therefore $\sum_{k=1}^{\infty}\mathbb{P}(A_k) < \infty$, and so the Borel–Cantelli lemma [11] implies that the probability of A_k infinitely often is 0. This means for almost every $\omega\in\Omega$ there is a k_0 depending on ω such that for $k \ge k_0$, each of the subcubes of side-length N_k^{-1} will contain at least one of the points of $x_1,\dots,x_{r_k}$.

Now for r arbitrary and sufficiently large (depending on ω), choose k such that $r_k \le r < r_{k+1}$. Then each of the subcubes of side-length N_k^{-1} will contain at least one of the points $x_1,\dots,x_{r_k}$, and hence at least one of the points $x_1,\dots,x_r$. Consequently

$$\delta(r) \le \frac{1}{N_k},$$

and thus

$$\left(\frac{r}{\log r}\right)^{1/d}\delta(r) \le \left(\frac{r_{k+1}}{\log r_{k+1}}\right)^{1/d}\delta(r) \le 2^{1/d}(2N_k+1)\delta(r) \le 2^{1/d}\left(2+\frac{1}{N_k}\right).$$

Taking $r\to\infty$ proves (26).

We prove the converse inequality

$$\limsup_{r\to\infty} \frac{\delta(r)}{(\log r/r)^{1/d}} \ge \frac{1}{4} \quad \text{a.s.} \tag{28}$$

in several steps.

Step 2. Assume for the moment that we have already chosen a sequence r_k (number of sampling points) and N_k. Then we divide $[0,1]^d$ into N_k^d subcubes of side-length N_k^{-1}, and we enumerate the cubes as $C_1, C_2, \dots, C_{N_k^d}$. Let D_j be the event that the cube C_j does not contain any of the points $x_{r_{k-1}+1},\dots,x_{r_k}$. As in (24) the probability of D_j is given by

$$\mathbb{P}(D_j) = (1 - N_k^{-d})^{r_k - r_{k-1}}. \tag{29}$$

For $j \neq k$, $D_j \cap D_k$ is the event that the region $C_j \cup C_k$ does not contain any of the points $x_{r_{k-1}+1}, \dots, x_{r_k}$. Therefore as in (24) we obtain that

$$\begin{aligned} \mathbb{P}(D_j \cap D_k) &= (1 - 2N_k^{-d})^{r_k - r_{k-1}} \\ &\leq (1 - N_k^{-d})^{2(r_k - r_{k-1})} = \mathbb{P}(D_j)\mathbb{P}(D_k), \end{aligned} \tag{30}$$

since $1 - 2x \leq (1-x)^2$ for $x \in [0,1]$.

Step 3. Now let B_k be the event that at least one of the first N_k (out of a total of N_k^d) cubes $C_1, \dots, C_{N_k}$ does not contain any of the points $x_{r_{k-1}}, \dots, x_{r_k}$. (In dimension $d = 1$ we take the first N_k^α of N_k cubes for some $\alpha, 1/2 < \alpha < 1 - 1/e$, and modify the following argument slightly.) If we define the random variable Y_k by

$$Y_k = \sum_{j=1}^{N_k} 1_{D_j},$$

then $B_k = \{Y_k > 0\}$. To find a lower estimate for the probability of B_k, we use an argument due to Kochen and Stone [24]. Using Cauchy–Schwarz we find that

$$\begin{aligned} \mathbb{E}\, Y_k &= \sum_{l=1}^{N_k} l\, \mathbb{P}(Y_k = l) \\ &\leq \left(\sum_{l=1}^{N_k} l^2 \mathbb{P}(Y_k = l) \right)^{1/2} \left(\sum_{l=1}^{N_k} \mathbb{P}(Y_k = l) \right)^{1/2} \\ &= (\mathbb{E}\, Y_k^2)^{1/2} (\mathbb{P}(Y_k > 0))^{1/2}, \end{aligned}$$

whence

$$\mathbb{P}(B_k) = \mathbb{P}(Y_k > 0) \geq \frac{(\mathbb{E}\, Y_k)^2}{\mathbb{E}\, Y_k^2}. \tag{31}$$

On the other hand,

$$\mathbb{E}\, Y_k = \sum_{j=1}^{N_k} \mathbb{P}(D_j) = N_k \mathbb{P}(D_j)$$

and by (30)

$$\begin{aligned} \mathbb{E}\, Y_k^2 &= \sum_{j=1}^{N_k} \mathbb{P}(D_j) + \sum_{k \neq j} \mathbb{P}(D_j \cap D_k) \\ &\leq \mathbb{E}\, Y_k + \sum_{k \neq j} \mathbb{P}(D_j)\mathbb{P}(D_k) \\ &\leq \mathbb{E}\, Y_k + (\mathbb{E}\, Y_k)^2. \end{aligned}$$

Substituting into (31), we obtain

$$\mathbb{P}(B_k) \geq \frac{\mathbb{E}\, Y_k}{1 + \mathbb{E}\, Y_k}. \tag{32}$$

Step 4. Finally we choose $r_k = e^{e^k}$ and N_k the least integer $\geq (2^d r_k/\log r_k)^{1/d}$. Then

$$\mathbb{P}(D_j) = (1 - N_k^{-d})^{r_k - r_{k-1}} \geq \left(1 - \frac{\log r_k}{2^d r_k}\right)^{r_k}.$$

Since $\lim_{x\to\infty} x^{1/2^d}\big(1 - \frac{\log x}{2^d x}\big)^x = 1$, we have $\big(1 - \frac{\log x}{2^d x}\big)^x \geq \frac{1}{2}x^{-1/2^d}$ for x sufficiently large, and consequently

$$\mathbb{E}\, Y_k = N_k \mathbb{P}(D_j) \geq \left(\frac{2^d r_k}{\log r_k}\right)^{1/d} \frac{1}{2} r_k^{-1/2^d} = r_k^{\frac{1}{d} - \frac{1}{2^d}} / (\log r_k)^{1/d} \geq 1 \tag{33}$$

for sufficiently large k ($k \geq 3$). Now (32) implies that $\mathbb{P}(B_k) \geq 1/2$ and so $\sum_{k=1}^{\infty} \mathbb{P}(B_k) = \infty$. Finally we observe that the events B_k are independent, because they depend on disjoint segments of the sequence $x_j, j \in \mathbb{N}$. Therefore the second part of the Borel–Cantelli lemma [11] implies that the probability of B_k infinitely often is 1. This means that for almost every ω there is an infinite subsequence of k's (depending on ω) such that $\omega \in B_k$.

Step 5. It remains to consider the event E_k that one of the points $x_1, \ldots, x_{r_{k-1}}$ is in $\bigcup_{j=1}^{N_k} C_j$. Since the volume of $\bigcup_{j=1}^{N_k} C_j$ is $N_k \cdot N_k^{-d}$, the probability that a particular x_j is in this set is N_k^{1-d}. There are r_{k-1} points to consider, so as in (24)

$$\mathbb{P}(E_k) \leq r_{k-1} N_k^{1-d}.$$

By our choices of r_k and N_k, we have $\sum_{k=1}^{\infty} \mathbb{P}(E_k) < \infty$, and so by the Borel–Cantelli lemma once again, the probability of E_k infinitely often is 0.

Combining Steps 4 and 5 we conclude that with probability 1, infinitely often at least one of the C_ℓ with $\ell \leq N_k$ will contain none of the points $x_1, \ldots, x_{r_k}$. Since C_ℓ contains none of these x_j, the center of C_ℓ is not contained in $\bigcup_{j=1}^{r_k} B(x_j, 1/(2N_k))$. Consequently $\delta(r_k) > 1/(2N_k)$ for infinitely many k almost surely. So

$$\delta(r_k) \left(\frac{r_k}{\log r_k}\right)^{1/d} \geq \delta(r_k) N_k/2 \geq 1/4$$

and (28) is proved.

Step 6. It is clear that if we omit the first M points $x_1, \ldots, x_M$ for any fixed integer M, then this will not affect the value of $\limsup \delta(r)/(\log r/r)^{1/d}$. Therefore this random variable is measurable with respect to the tail σ-field of the sequence $x_1, x_2, \ldots$. By the Kolmogorov's 0-1 law, the value of this random variable must be constant almost surely [11, p. 254]. This completes the proof. □

5. Asymptotic estimates of the condition number. In the previous section we have combined a deterministic argument with a covering argument. Essentially we have calculated the probability that a random sampling set satisfies the sufficient condition already known from deterministic sampling theory.

In this section we develop an alternative approach that is based on a metric entropy argument such as the ones used in [13]. This approach does not rely on deterministic sampling results and can therefore be adapted to other sampling models. On the other hand, it is difficult to keep track of the constants involved, and thus the results are only efficient for large sampling sets.

Once again we start with an infinite sequence of i.i.d. random variables $\{x_j : j \in \mathbb{N}\}$, each of which is uniformly distributed over $[0,1]^d$. Our goal is to estimate the

quantity $\sum_{j=1}^r |p(x_j)|^2 - r\|p\|_2^2$ and its distribution as a function of the number of sampling points r.

For every $p \in \mathcal{P}_M$ we introduce the random variable $Y_j(p) = |p(x_j)|^2 - \|p\|_2^2$. To obtain a sampling inequality of the form $A\|p\|_2^2 \le \sum_{j=1}^r |p(x_j)|^2 \le B\|p\|_2^2$, we have to estimate the probability distribution of the random variable

$$\sup_{p\in\mathcal{P}_M, \|p\|_2=1} \sum_{j=1}^r Y_j(p).$$

This is accomplished in the following theorem.

THEOREM 5.1. *If $\{x_j : j \in \mathbb{N}\}$ is a sequence of i.i.d. random variables that are uniformly distributed over $[0,1]^d$, then there exist constants $A, B > 0$ depending on d and M, such that*

$$\mathbb{P}\left(\sup_{p\in\mathcal{P}_M, \|p\|_2=1} \sup_{s\le r} \left| \sum_{j=1}^s Y_j(p) \right| \ge \lambda \right) \le A \exp\left(- B \frac{\lambda^2}{r+\lambda} \right) \tag{34}$$

for $r \in \mathbb{N}$ and $\lambda \ge 0$.

For the distribution of a sum of random variables we use Bernstein's inequality [6] in the following form.

PROPOSITION 5.2. *Let $Y_j, j = 1, \dots, r$, be a sequence of bounded, independent random variables with $\mathbb{E}\, Y_j = 0$, $\operatorname{Var} Y_j = \sigma^2$, and $\|Y_j\|_\infty \le M$ for $j = 1, \dots, r$. Then*

$$\mathbb{P}\left(\left| \sum_{j=1}^r Y_j \right| \ge \lambda \right) \le 2 \exp\left(- \frac{\lambda^2}{2r\sigma^2 + \frac{2}{3}M\lambda} \right). \tag{35}$$

To apply (35) to the $Y_j(p)$, we need several simple estimates. It suffices to work with the unit ball of $\mathcal{P}_M$, which we denote by $\mathcal{P}^0 = \{p \in \mathcal{P}_M : \|p\|_2 \le 1\}$.

LEMMA 5.3. *Let $p, q \in \mathcal{P}^0$ and $j \in \mathbb{N}$. Then the following identities and inequalities hold:*

$$\mathbb{E}\, Y_j(p) = 0, \tag{36}$$

$$\operatorname{Var} Y_j(p) = \|p\|_4^4 - \|p\|_2^4 \le D - 1, \tag{37}$$

$$\operatorname{Var}\left(Y_j(p) - Y_j(q)\right) \le 8\|p-q\|_\infty^2, \tag{38}$$

$$\|Y_j(p)\|_\infty \le \|p\|_\infty^2 - \|p\|_2^2 \le (D-1), \tag{39}$$

$$\|Y_j(p) - Y_j(q)\|_\infty \le 2(D^{1/2}+1)\|p-q\|_\infty. \tag{40}$$

Proof. Since each x_j is uniformly distributed over $[0,1]^d$, we have

$$\mathbb{E}\left(Y_j(p)\right) = \int_{[0,1]^d} \left(|p(x)|^2 - \|p\|_2^2 \right) dx = 0$$

and consequently (also using (10))

$$\begin{aligned} \operatorname{Var} Y_j(p) = \mathbb{E}\left[Y_j(p)^2\right] &= \int_{[0,1]^d} \left(|p(x)|^2 - \|p\|_2^2 \right)^2 dx \\ &= \|p\|_4^4 - \|p\|_2^4 \le D - 1, \end{aligned}$$

since $\|p\|_2 = 1$. Similarly, we obtain

$$\|Y_j(p)\|_\infty = \sup_{\omega\in\Omega} \left| |p(x_j(\omega))|^2 - \|p\|_2^2 \right| \le \left| \|p\|_\infty^2 - \|p\|_2^2 \right| \le D - 1.$$

Next, since $\mathbb{E}\, Y_j(p) = 0$, we obtain

$$\begin{aligned}\operatorname{Var}(Y_j(p) - Y_j(q)) &= \mathbb{E}\left((Y_j(p) - Y_j(q))^2\right)\\ &= \int_{[0,1]^d} \left(|p(x)|^2 - |q(x)|^2\right)^2 dx - (\|p\|_2^2 - \|q\|_2^2)^2\\ &\le \|p-q\|_\infty^2 \,\|\, |p| + |q| \,\|_2^2 + \|p-q\|_2^2\big(\|p\|_2^2 + \|q\|_2^2\big)\\ &\le 8\|p-q\|_\infty^2.\end{aligned}$$

The last estimate follows similarly from

$$\begin{aligned}\|Y_j(p) - Y_j(q)\|_\infty &\le \sup_{\omega\in\Omega}\Big|\, |p(x_j(\omega))|^2 - |q(x_j(\omega))|^2\Big| + \Big|\|q\|_2^2 - \|p\|_2^2\Big|\\ &\le \|p-q\|_\infty\big(\|p\|_\infty + \|q\|_\infty\big) \;+\; \|q-p\|_2\big(\|p\|_2 + \|q\|_2\big)\\ &\le \|p-q\|_\infty\, D^{1/2}\big(\|p\|_2 + \|q\|_2\big) \;+\; \|q-p\|_\infty\big(\|p\|_2 + \|q\|_2\big)\\ &= 2(D^{1/2}+1)\,\|p-q\|_\infty. \qquad \square\end{aligned}$$

Proof of Theorem 5.1.

Step 1: *A metric entropy argument.* For a given $\delta > 0$, we construct a δ-net for $\mathcal{P}^0$ with respect to the L^∞-norm as follows. Given $p \in \mathcal{P}^0$ with coefficients $\mathbf{a} = (a_k)_{k\in\mathbb{Z}^d\cap[-M,M]^d}$ and $\|\mathbf{a}\|_2 \le 1$, we approximate the real and imaginary parts of each a_k by a number $\frac{\delta}{\sqrt{2}D}\ell$, $\ell \in \mathbb{Z}$; in other words, we choose a vector $\mathbf{b}$ of the form $\mathbf{b} = \frac{\delta}{\sqrt{2}D}(\ell + im)$, $\ell, m \in \mathbb{Z}^d$, to approximate $\mathbf{a}$. Then for each coordinate $a_k, k \in [-M,M]^d \cap \mathbb{Z}^d$, we have

$$|a_k - b_k| \le \frac{\delta}{D},$$

and so

$$\|\mathbf{a}-\mathbf{b}\|_2 \le \big(D \max_k |a_k - b_k|^2\big)^{1/2} = \frac{\delta}{\sqrt{D}}\,.$$

Setting $q(x) = \sum_{k\in I_M} b_k e^{2\pi i k\cdot x}$, we obtain

$$\|p-q\|_\infty \le D^{1/2}\|p-q\|_2 \le D^{1/2}\|\mathbf{a}-\mathbf{b}\|_2 = \delta.$$

We denote the δ-net of all $q \in \mathcal{P}^0$ with coefficients of the form $\mathbf{b} = \frac{\delta}{\sqrt{2}D}(\ell + im)$, $\ell, m \in \mathbb{Z}^d, \|\mathbf{b}\|_2 \le 1$, by $\mathcal{A}(\delta)$. The cardinality of $\mathcal{A}(\delta)$ is estimated as follows:

$$\begin{aligned}\operatorname{card}\mathcal{A}(\delta) &= \operatorname{card}\left\{\mathbf{b} = \frac{\delta}{\sqrt{2}D}(\ell + im),\ \ell, m \in \mathbb{Z}^D, \|\mathbf{b}\|_2 \le 1\right\}\\ &= \operatorname{card}\left\{k \in \mathbb{Z}^{2D} : \|k\|_2 \le \frac{\sqrt{2}D}{\delta}\right\}\\ &\le c_1\delta^{-2D},\end{aligned}$$

where the constant $c_1 \approx \frac{(2\pi)^D}{D}D^{2D}$ is roughly the number of integer lattice points in a ball of radius $\sqrt{2}D$ in $\mathbb{R}^{2D}$.

Given $p \in \mathcal{P}^0$, let p_j be the polynomial in $\mathcal{A}(2^{-j})$ that is closest to p in L^∞-norm, with some convention for breaking ties. Since $\|p - p_j\|_2 \to 0$, we can write

$$Y_j(p) = Y_j(p_0) + (Y_j(p_1) - Y_j(p_0)) + (Y_j(p_2) - Y_j(p_1)) + \cdots.$$

If $\sup_{p\in\mathcal{P}^0}\sup_{s\le r}|\sum_{j=1}^s Y_j(p)| \ge \lambda$, then either

(a) $\sup_{s\le r}|\sum_{j=1}^s Y_j(p)|\ge\lambda/2$ for some $p\in\mathcal{A}(1)$; or
(b) for some $\ell\ge 1$, some $p\in\mathcal{A}(2^{-\ell})$, and some $q\in\mathcal{A}(2^{-\ell+1})$ with $\|p-q\|_\infty\le 3\cdot 2^{-\ell}$ we have $\sup_{s\le r}|\sum_{j=1}^s\big(Y_j(p)-Y_j(q)\big)|\ge\lambda/2(\ell+1)^2$.

(Possibly both (a) and (b) hold.)

If this were not the case, then

$$\sup_{s\le r}\left|\sum_{j=1}^s Y_j(p)\right|\le\sup_{s\le r}\left|\sum_{j=1}^s Y_j(p_0)\right|+\sup_{s\le r}\sum_{\ell=1}^\infty\left|\sum_{j=1}^s(Y_j(p_\ell)-Y_j(p_{\ell-1}))\right|$$
$$\le\sum_{\ell=1}^\infty\frac{\lambda}{2\ell^2}=\frac{\pi^2}{12}\lambda<\lambda.$$

So far the construction is purely deterministic. Now we estimate the probability of each of the events in (a) and (b).

Step 2. For fixed $p\in\mathcal{A}(1)$, the probability of the event in (a) is bounded, using Bernstein's inequality (35) and Lemma 5.3, by

$$2\exp\Big(-\frac{\lambda^2}{2r\operatorname{Var}Y_j(p)+\frac{2}{3}\lambda\|Y_j(p)\|_\infty}\Big)$$
$$\le 2\exp\Big(-\frac{\lambda^2}{2r(D-1)+\frac{2}{3}(D-1)\lambda}\Big).$$

There are at most c_1 polynomials in $\mathcal{A}(1)$, so the probability of (a) is bounded by

$$2c_1\exp\Big(-\frac{\lambda^2}{(D-1)(2r+\frac{2}{3}\lambda)}\Big). \tag{41}$$

Step 3. We estimate (b) in a similar fashion using Lemma 5.3, (38), and (40). If $p\in\mathcal{A}(2^{-\ell})$ and $q\in\mathcal{A}(2^{-\ell+1})$ with $\|p-q\|_\infty\le 3\cdot 2^{-\ell}$, we have

$$\mathbb{P}\Big(\sup_{s\le r}\Big|\sum_{j=1}^s(Y_j(p)-Y_j(q))\Big|>\frac{\lambda}{2(\ell+1)^2}\Big)$$
$$\le 2\exp\Big(-\frac{\lambda^2/4(\ell+1)^4}{144r2^{-2\ell}+4\cdot 2^{-\ell}D^{1/2}\lambda/(\ell+1)^2)}\Big)$$
$$\le 2\exp\Big(-2^\ell\frac{\lambda^2}{c_3(r(\ell+1)^4 2^{-\ell}+D^{1/2}\lambda(\ell+1)^2)}\Big).$$

There are $c_1 2^{(2\ell-2)D}$ trigonometric polynomials in $\mathcal{A}(2^{-\ell+1})$, and for each q the number of trigonometric polynomials $p\in\mathcal{A}(2^{-\ell})$ satisfying $\|p-q\|_\infty\le 3\cdot 2^{-\ell}$ is bounded by a constant c_2 independent of q and j. (Similar to the count in Step 1, $c_2\approx\frac{(6\pi)^D}{D}D^{2D}$ is roughly the number of integer lattice points in a ball of radius $3\sqrt{2}D$ in $\mathbb{R}^{2D}$.) Finally, this can happen for any ℓ. So the probability in (b) is bounded by

$$\sum_{\ell=1}^\infty 2c_1c_2 2^{(2\ell-2)D}\exp\Big(-2^\ell\frac{\lambda^2}{c_3(r(\ell+1)^4 2^{-\ell}+D^{1/2}\lambda(\ell+1)^2)}\Big). \tag{42}$$

Step 4. Estimate of the sum (42).

Since $(\ell+1)^4 2^{-\ell}$ is bounded above and $2^{\ell/2}/(\ell+1)^2$ is bounded below, the above sum is bounded by

$$\sum_{\ell=1}^{\infty} c_4 \exp\left(-2^{\ell/2}\frac{\lambda^2}{c_5(r+\lambda)} + (2\ell-2)D\log 2\right) = (\star). \tag{43}$$

We distinguish two cases. In the first,

$$\frac{\lambda^2}{c_5(r+\lambda)} \geq 64D. \tag{44}$$

Then

$$2^{\ell/2}\frac{\lambda^2}{c_5(r+\lambda)} \geq 2(2\ell-2)D\log 2 \quad \forall\, \ell \geq 1,$$

and so

$$(\star) \leq \sum_{\ell=1}^{\infty} c_4 \exp\left(-2^{\ell/2}\frac{\lambda^2}{2c_5(r+\lambda)}\right).$$

Now we use the fact that $\sum_{\ell=1}^{\infty} e^{-a^\ell x} \leq c_6 e^{-x}$ for any $a > 1$ and $x \geq 1$ (with c_6 depending only on a). Consequently the sum in (43) is bounded by

$$(\star) \leq c_7 \exp\left(-\frac{\lambda^2}{c_8(r+\lambda)}\right).$$

In the second case, (44) does not hold. But then the probability of the event in (b) is at most 1, which is certainly less than or equal to

$$e^{64D}\exp\left(-\frac{\lambda^2}{c_8(r+\lambda)}\right).$$

In either case, we have that the probability of the event in (b) is bounded by

$$c_9 \exp\left(-\frac{\lambda^2}{c_8(r+\lambda)}\right).$$

Step 5. The statement now follows by combining the bounds for (a) and (b), and so we have

$$\mathbb{P}\left(\sup_{p\in\mathcal{P}^0}\sup_{s\leq r}\left|\sum_{j=1}^{s} Y_j(p)\right| \geq \lambda\right) \leq A\exp\left(-B\frac{\lambda^2}{r+\lambda}\right). \qquad \square \tag{45}$$

COROLLARY 5.4. *If $\{x_j : j \in \mathbb{N}\}$ is a sequence of i.i.d. random variables that are uniformly distributed over $[0,1]^d$ and $0 < \mu < 1$, then the sampling inequality*

$$(1-\mu)r\|p\|_2^2 \leq \sum_{j=1}^{r} |p(x_j)|^2 \leq (1+\mu)r\|p\|_2^2 \quad \forall p \in \mathcal{P}_M \tag{46}$$

holds with probability at least

$$1 - Ae^{-Br\frac{\mu^2}{1+\mu}}.$$

Consequently with the same probability estimate the Toeplitz-type matrix $\mathcal{T}$ has condition number $\kappa(\mathcal{T}) \le \frac{1+\mu}{1-\mu}$ and also $\kappa(\mathcal{U}) \le (\frac{1+\mu}{1-\mu})^{1/2}$

Proof. Choose $\lambda = r\mu$ in Theorem 5.1 and observe that the inequality

$$\left| \sum_{j=1}^{r} |p(x_j)|^2 - r \right| \le r\mu$$

for all $p \in \mathcal{P}^0$ is equivalent to the sampling inequality (46) for all $p \in \mathcal{P}_M$. □

From Theorem 5.1 it is straightforward to obtain a law of the iterated logarithm.

COROLLARY 5.5. *If $\{x_j : j \in \mathbb{N}\}$ is a sequence of i.i.d. random variables that are uniformly distributed over $[0,1]^d$, then*

$$\limsup_{r\to\infty} \frac{\sup_{p\in\mathcal{P}} |\sum_{j=1}^{r} [|p(x_j)|^2 - \|p\|_2^2]|}{\sqrt{r \log\log r}\, \|p\|_2^2} = c \quad a.s. \tag{47}$$

for some constant $c \in [\left(\frac{2}{\pi}\right)^d D - 1, \infty)$.

Proof. Let $r_k = 2^k$ and $\lambda_k = \frac{2}{\sqrt{B}} \sqrt{r_k \log\log r_k}$, where B is the constant from (34). Let

$$C_k = \left\{ \sup_{p\in\mathcal{P}^0} \sup_{s\le r_k} \left| \sum_{j=1}^{s} Y_j(p) \right| > \lambda_k \right\}.$$

Then for k large enough, we have $r_k > \lambda_k$. So the probability of C_k is bounded by

$$\begin{aligned}
\mathbb{P}(C_k) &\le A \exp\left(- B \frac{\lambda_k^2}{r_k + \lambda_k} \right) \\
&\le A \exp\left(- B \frac{\lambda_k^2}{2r_k} \right) \\
&\le A \exp\left(- B \frac{4}{B} \frac{r_k \log\log r_k}{2r_k} \right) \\
&\le A \exp(-2 \log k) = \frac{A}{k^2}.
\end{aligned}$$

So $\sum_{k=1}^{\infty} \mathbb{P}(C_k) < \infty$, and by the Borel–Cantelli lemma, the probability of C_k happening infinitely often is 0.

If $|\sum_{j=1}^{r} Y_j(p)| > \frac{2}{\sqrt{B}} \sqrt{r \log\log r}$ for some r, we choose k so that $r_{k-1} \le r < r_k$ and observe that C_k holds. (This is the only place where we need the estimate for $\sup_{s\le r} |\sum_{j=1}^{s} Y_j(p)|$ instead of just $|\sum_{j=1}^{r} Y_j(p)|$.) So this inequality cannot happen for infinitely many r and we therefore have

$$\limsup_{r\to\infty} \frac{\sup_{p\in\mathcal{P}^0} |\sum_{j=1}^{r} [|p(x_j)|^2 - r|}{\sqrt{r \log\log r}} \le c' \quad a.s.$$

for some constant $c' > 0$.

For fixed $p \in \mathcal{P}^0$ the classical law of the iterated logarithm [11, p. 232] says that

$$\limsup_{r\to\infty} \frac{\left|\sum_{j=1}^{r} Y_j(p)\right|}{\sqrt{2r \log\log r}} = \sqrt{\operatorname{Var} Y_j(p)} = \|p\|_4^4 - 1 \quad \text{a.s.}$$

Choosing $p(x) = D^{-1/2} \sum_{k\in[-M,M]^d\cap\mathbb{Z}^d} e^{2\pi i k\cdot x}$, we have $\|p\|_2 = 1$ and the elementary estimate $\|p\|_4 \geq \frac{2}{\pi} D^{1/4}$. So

$$\limsup_{r\to\infty} \frac{\sup_{p\in\mathcal{P}^0} |\sum_{j=1}^r [|p(x_j)|^2 - r]|}{\sqrt{r\log\log r}} \geq \left(\frac{2}{\pi}\right)^4 D - 1\,.$$

The conclusion follows as in the proof of Theorem 4.3 by applying Kolmogorov's 0-1 law. □

This result can be summarized by saying that for large enough r (r depending on ω) we always have the sampling inequality

$$(48)\qquad (r - c\sqrt{r\log\log r})\|p\|_2^2 \leq \sum_{j=1}^r |p(x_j)|^2 \leq (r + c\sqrt{r\log\log r})\|p\|_2^2 \quad \forall p\in\mathcal{P}_M.$$

The condition number of the random matrix $\mathcal{T}$ is therefore

$$\kappa \leq (r + c\sqrt{r\log\log r})/(r - c\sqrt{r\log\log r}) \approx 1 + 2c\left(\frac{\log\log r}{\sqrt{r}}\right)^{1/2}$$

almost surely for some constant c of order D.

6. A universal sampling theorem and examples. The main statements (Theorems 4.2 and 5.1 and Corollary 5.4) reach similar conclusions. At first glance, Theorem 4.2 seems preferable because of its elementary proof and the explicit constants. In this section we focus on the merits of the metric entropy method. This method is extremely flexible and works for many other sampling problems. We formulate a general framework for finite-dimensional sampling theorems and derive a universal sampling theorem in the style of Corollary 5.4. We then will discuss several examples of practical interest.

To begin, we note that the proofs of Theorem 5.1 and Corollary 5.4 do not use any specific properties of trigonometric polynomials. In fact, we have used only the following (interrelated) properties of $\mathcal{P}_M$.

(a) The space $\mathcal{P}_M$ is finite-dimensional and possesses a basis of continuous functions.

(b) All norms on $\mathcal{P}_M$ are equivalent; in the proofs we have used the norms $\|p\|_2, \|p\|_4, \|p\|_\infty$, and $\|\mathbf{a}\|_2$ and the associated equivalence constants. As a consequence the random variables related to the samples $|p(x_j)|^2$ satisfy the uniform estimates of Lemma 5.3.

(c) The unit ball of $\mathcal{P}_M$ is compact. This fact enables the construction of the δ-nets $\mathcal{A}(\delta)$ and suitable estimates for their cardinality.

It is evident that Theorem 5.1 and Corollary 5.4 can be obtained under much more general conditions.

A general framework. We make the following assumptions.

1. Let $S \subseteq \mathbb{R}^d$ be a compact set and let ν be a probability measure on S with $\operatorname{supp}\nu = S$.

2. Let $\mathcal{B}$ be a finite-dimensional subspace of $L^2(S,\nu)$ with a basis $\{e_k : k = 1,\dots,D\}$ of continuous functions. Often this basis is chosen as a finite subset of a Riesz basis for $L^2(S,\nu)$ and in this sense $\mathcal{B}$ may be interpreted as a space of band-limited functions in $L^2(S,\nu)$. Since $p = \sum_{k=1}^D a_k e_k$ for every $p\in\mathcal{B}$, all functions in $\mathcal{B}$ are continuous.

The sampling problem in $\mathcal{B}$. The task is now to interpolate or to approximate a given data set $\{(x_j, p(x_j)) : j = 1, \dots, r\}$ by a function in $\mathcal{B}$. As in section 2 this amounts to solving the system of linear equations

$$\sum_{k=1}^{D} a_k e_k(x_j) = p(x_j) = y_j, \quad j = 1, \dots, r.$$

Let $\mathcal{U}_{jk} = e_k(x_j)$ and

$$\mathcal{T}_{kl} = (\mathcal{U}^*\mathcal{U})_{kl} = \sum_{j=1}^{r} \overline{e_k(x_j)} e_l(x_j); \tag{49}$$

then we need to solve either the $r \times D$ system

$$\mathcal{U}\mathbf{a} = \mathbf{y}$$

or the $D \times D$ normal equations

$$\mathcal{T}\mathbf{a} = \mathcal{U}^*\mathcal{U}\mathbf{a} = \mathcal{U}^*\mathbf{y}.$$

Assume that we can prove the sampling inequality

$$A\|p\|_{2,\nu}^2 \le \sum_{j=1}^{r} |p(x_j)|^2 = \langle \mathcal{T}\mathbf{a}, \mathbf{a}\rangle \le B\|p\|_{2,\nu}^2 \quad \forall p \in \mathcal{B}. \tag{50}$$

Inserting the norm equivalence $\alpha\|\mathbf{a}\|_2 \le \|p\|_{2,\nu} \le \beta\|\mathbf{a}\|_2$, (50) then implies the estimates

$$\kappa(\mathcal{T}) \le \frac{\beta^2 B}{\alpha^2 A} \quad \text{and} \quad \kappa(\mathcal{U}) \le \Big(\frac{\beta^2 B}{\alpha^2 A}\Big)^{1/2} \tag{51}$$

for the condition numbers of these matrices. Furthermore, $p \in \mathcal{B}$ is uniquely determined by its samples, if and only if $\mathcal{T}$ is invertible, or if and only if $r \ge D$ and $\mathcal{U}$ has full rank.

We can now formulate our main theorem for random sampling in finite-dimensional spaces of band-limited functions.

THEOREM 6.1. *If $\{x_j : j \in \mathbb{N}\}$ is a sequence of i.i.d. random variables and if each x_j is ν-distributed over S, then there exist constants $A, B > 0$ depending on S, ν, and D, such that for all $\mu \in (0, 1)$, the sampling inequality*

$$(1-\mu) r\|p\|_{2,\nu}^2 \le \sum_{j=1}^{r} |p(x_j)|^2 \le (1+\mu) r\|p\|_{2,\nu}^2 \quad \forall p \in \mathcal{B} \tag{52}$$

holds with probability at least

$$1 - Ae^{-Br\frac{\mu^2}{1+\mu}}.$$

With the same probability estimate the matrix $\mathcal{T}$ has condition number $\kappa(\mathcal{T}) \le \frac{\beta^2(1+\mu)}{\alpha^2(1-\mu)}$ and also $\kappa(\mathcal{U}) \le (\frac{\beta^2(1+\mu)}{\alpha^2(1-\mu)})^{1/2}$

Proof. We have already done all the work when we proved Theorem 5.1 and Corollary 5.4. The only minor modifications occur in the constants in Lemma 5.3 and in Step 1 of the proof. We now use the random variables $Y_j(p) = |p(x_j)|^2 - \|p\|_{2,\nu}^2 = |p(x_j)|^2 - \mathbb{E}\,[|p(x_j)|^2]$. □

We present the following examples where the general hypotheses are satisfied, and so Theorem 6.1 is applicable. Each example yields a new result on random sampling. In some of these examples it seems to be extremely difficult to derive quantitative deterministic results in the style of Theorem 4.1.

Example 1. Trigonometric polynomials revisited. Choose a closed set $S \subseteq [0,1]^d$ of positive Lebesgue measure and a probability measure ν with $\operatorname{supp} \nu = S$ and equivalent to λ on S. If $p \in \mathcal{P}_M$ vanishes on S, then by Lemma 3.1 $p \equiv 0$ and, consequently, $\|p\chi_S\|_{2,\nu} = \left(\int_S |p(x)|^2 \, d\nu(x)\right)^{1/2}$ is equivalent to the L^2-norm on $\mathcal{P}_M$; i.e., there exist constants $\alpha, \beta > 0$ such that

$$\alpha \|p\|_2 \leq \|p\chi_S\|_{2,\nu} \leq \beta \|p\|_2 \quad \forall p \in \mathcal{P}_M.$$

We state the conclusion of Theorem 6.1 explicitly.

THEOREM 6.2. *Suppose that $\{x_j : j \in \mathbb{N}\} \subseteq S$ is a sequence of i.i.d. random variables that are ν-distributed over S. Then there exist constants $A, B > 0$ depending on S, ν, and D, such that for all $\mu \in (0,1)$ the sampling inequality*

$$\alpha^2(1-\mu) r \|p\|_2^2 \leq \sum_{j=1}^{r} |p(x_j)|^2 \leq \beta^2 (1+\mu) r \|p\|_2^2 \quad \forall p \in \mathcal{P}_M \tag{53}$$

holds with probability at least

$$1 - Ae^{-Br\frac{\mu^2}{1+\mu}}.$$

With the same probability estimate we have $\kappa(\mathcal{T}) \leq \frac{\beta^2(1+\mu)}{\alpha^2(1-\mu)}$.

Comparing with Theorem 5.1 we have been able to change the distribution of the random variables x_j and the target set S in which the samples are taken.

Example 2. Almost periodic functions and trigonometric polynomials with arbitrary frequencies. Assume that $S \subseteq \mathbb{R}^d$ is compact and has positive Lebesgue measure and that ν is equivalent to λ on S. Choose exponentials $e^{i\lambda_k \cdot x}$ with arbitrary frequencies $\lambda_k \in \mathbb{R}^d$ ($\lambda_k \in \mathbb{Z}^d$ is the case of trigonometric polynomials) and consider the subspace of almost periodic functions (trigonometric polynomials) on S,

$$\mathcal{B} = \{p \in L^2(S) : p(x) = \sum_{k=1}^{D} a_k e^{i\lambda_k \cdot x} \ \chi_S(x).$$

Then Theorem 6.1 applies.

Example 3. Algebraic polynomials. Again assume that $S \subseteq \mathbb{R}^d$ has positive Lebesgue measure and that ν is equivalent to λ on S. Choose a finite set $F \subseteq (\mathbb{N} \cap \{0\})^d$ and consider the space of algebraic polynomials on a compact set $S \subseteq \mathbb{R}^d$ defined as

$$\mathcal{P}_F = \{p \in L^2(S) : p(x) = \sum_{k \in F} a_k x^\alpha \ \chi_S(x).$$

Thus Theorem 6.1 applies also to algebraic polynomials of several variables.

Example 4. Local shift-invariant spaces. Let ϕ be a continuous function on $\mathbb{R}^d$ with $\operatorname{supp} \phi \subseteq [-\sigma, \sigma]^d \subseteq S$. The local shift-invariant space $V(\phi, S)$ is defined by

$$V(\phi, S) = \left\{ f \in L^2(S) : f(x) = \sum_{k \in (S + [-\sigma,\sigma]^d) \cap \mathbb{Z}^d} a_k \phi(x-k) \right\}.$$

If we assume that $0 < a \le \sum_{k\in\mathbb{Z}^d} |\hat{\phi}(\omega - k)|^2 \le b$ for all $\omega \in \mathbb{R}^d$, then the translates $\phi(x-k), k \in \mathbb{Z}^d$, form a Riesz basis for the generated subspace, and so any finite subset is linearly independent. Thus Theorem 6.1 applies. In dimension $d = 1$ and for certain "generators" ϕ this model is well understood both numerically [22] and theoretically [1]. In dimension $d > 1$, however, there are no quantitative deterministic estimates. Theorem 6.1 gives the first hint that the numerical methods of [22] also work in higher dimensions. See [2] for a survey of sampling in shift-invariant spaces.

Example 5. Sampling on the sphere and spherical harmonics. Let $S_d = \{x \in \mathbb{R}^{d+1} : |x| = 1\}$ be the unit sphere in $\mathbb{R}^{d+1}$ with surface measure ν_d. We choose the sequence J_ℓ of suitably normalized spherical harmonics [36] as an orthonormal basis for $L^2(S_d, \nu_d)$ and consider the space of band-limited functions on the sphere, namely,

$$\mathcal{B} = \left\{ p \in L^2(S_d, \nu_d) : p = \sum_{\ell=1}^{D} a_\ell J_\ell \right\}.$$

Then the conclusions of Theorem 6.1 hold for every sequence of i.i.d. random variables x_j on S_d with x_j being ν_d-distributed.

Remark. Whereas the asymptotic results for the distribution number hold universally in finite-dimensional vector spaces, the generalization of Theorem 3.2 is more subtle and depends on the support properties of the basis functions. The same proof as in section 3 shows that the system matrix $\mathcal{T}$ defined in (49) is invertible with probability 1 in Examples 1, 2, and 3 whenever $r \ge D$. On the other hand, for Example 4 it can be shown that $\mathcal{T}$ is always singular with positive probability. As this probability depends on the number of samples r, this observation does not contradict Theorem 6.1.

REFERENCES

[1] A. Aldroubi and K. Gröchenig, *Beurling-Landau-type theorems for non-uniform sampling in shift invariant spline spaces*, J. Fourier Anal. Appl., 6 (2000), pp. 93–103.

[2] A. Aldroubi and K. Gröchenig, *Nonuniform sampling and reconstruction in shift-invariant spaces*, SIAM Rev., 43 (2001), pp. 585–620.

[3] A. Averbuch, R. Coifman, D. Donoho, M. Israeli, and J. Walden, *Fast slant stack: A notion of Radon transform for data in a Cartesian grid which is rapidly computable, algebraically exact, geometrically faithful and invertible*, SIAM J. Sci. Comput., to appear.

[4] R. F. Bass, *Law of the iterated logarithm for set-indexed partial sum processes with finite variance*, Z. Wahrsch. Verw. Gebiete, 70 (1985), pp. 591–608.

[5] R. F. Bass, *Probabilistic Techniques in Analysis*, Probab. Appl., Springer-Verlag, New York, 1995.

[6] G. Bennett, *Probability inequalities for the sum of independent random variables*, J. Amer. Statist. Assoc., 57 (1962), pp. 33–45.

[7] A. Beurling, *Local harmonic analysis with some applications to differential operators*, in Some Recent Advances in the Basic Sciences, Vol. 1, Belfer Graduate School of Science, Yeshiva University, New York, 1966, pp. 109–125.

[8] G. Beylkin, *On the fast Fourier transform of functions with singularities*, Appl. Comput. Harmon. Anal., 2 (1995), pp. 363–381.

[9] G. Chistyakov and Y. Lyubarskii, *Random perturbations of exponential Riesz bases in $L^2(-\pi, \pi)$*, Ann. Inst. Fourier (Grenoble), 47 (1997), pp. 201–255.

[10] G. Chistyakov, Y. Lyubarskii, and L. Pastur, *On completeness of random exponentials in the Bargmann-Fock space*, J. Math. Phys., 42 (2001), pp. 3754–3768.

[11] K. L. Chung, *A course in probability theory*, 2nd ed., Probab. Math. Statist. 21, Academic Press, New York, 1974.

[12] F. Cucker and S. Smale, *On the mathematical foundations of learning*, Bull. Amer. Math. Soc. (N.S.), 39 (2002), pp. 1–49.

[13] R. M. Dudley, *Sample functions of the Gaussian process*, Ann. Probab., 1 (1973), pp. 66–103.
[14] A. Dutt and V. Rokhlin, *Fast Fourier transforms for nonequispaced data*, SIAM J. Sci. Comput., 14 (1993), pp. 1368–1393.
[15] A. Dutt and V. Rokhlin, *Fast Fourier transforms for nonequispaced data*, II, Appl. Comput. Harmon. Anal., 2 (1995), pp. 85–100.
[16] H. Fassbender, *On numerical methods for discrete least-squares approximation by trigonometric polynomials*, Math. Comp., 66 (1997), pp. 719–741.
[17] H. G. Feichtinger, K. Gröchenig, and T. Strohmer, *Efficient numerical methods in non-uniform sampling theory*, Numer. Math., 69 (1995), pp. 423–440.
[18] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.
[19] K. Gröchenig, *Reconstruction algorithms in irregular sampling*, Math. Comp., 59 (1992), pp. 181–194.
[20] K. Gröchenig, *Irregular sampling, Toeplitz matrices, and the approximation of entire functions of exponential type*, Math. Comp., 68 (1999), pp. 749–765.
[21] K. Gröchenig, *Non-uniform sampling in higher dimensions: From trigonometric polynomials to bandlimited functions*, in Modern Sampling Theory, Appl. Numer. Harmon. Anal., Birkhäuser Boston, Boston, MA, 2001, pp. 155–171.
[22] K. Gröchenig and H. Schwab, *Fast local reconstruction methods for nonuniform sampling in shift-invariant spaces*, SIAM J. Matrix Anal. Appl., 24 (2003), pp. 899–913.
[23] K. Gröchenig and T. Strohmer, *Numerical and theoretical aspects of non-uniform sampling of band-limited images*, Chapter 6 in Nonuniform Sampling: Theory and Applications, F. Marvasti, ed., Kluwer Academic, Dordrecht, The Netherlands, pp. 283–324.
[24] S. Kochen and C. Stone, *A note on the Borel-Cantelli lemma*, Illinois J. Math., 8 (1964), pp. 248–251.
[25] H. J. Landau, *Necessary density conditions for sampling and interpolation of certain entire functions*, Acta Math., 117 (1967), pp. 37–52.
[26] Y. I. Lyubarskii and K. Seip, *Complete interpolating sequences for Paley-Wiener spaces and Muckenhoupt's* (A_p) *condition*, Rev. Mat. Iberoamericana, 13 (1997), pp. 361–376.
[27] B. S. Pavlov, *The basis property of a system of exponentials and the condition of Muckenhoupt*, Dokl. Akad. Nauk SSSR, 247 (1979), pp. 37–40.
[28] D. Potts and G. Steidl, *Fourier reconstruction of functions from their nonstandard sampled Radon transform*, J. Fourier Anal. Appl., 8 (2002), pp. 513–533.
[29] D. Potts, G. Steidl, and M. Tasche, *Fast Fourier transforms for nonequispaced data: A tutorial*, in Modern Sampling Theory, Appl. Numer. Harmon. Anal., Birkhäuser Boston, Boston, MA, 2001, pp. 247–270.
[30] M. Rauth and T. Strohmer, *Smooth approximation of potential fields from noisy scattered data*, Geophys., 63 (1998), pp. 85–94.
[31] L. Reichel, G. S. Ammar, and W. B. Gragg, *Discrete least squares approximation by trigonometric polynomials*, Math. Comp., 57 (1991), pp. 273–289.
[32] K. Seip and A. M. Ulanovskii, *Random exponential frames*, J. London Math. Soc. (2), 53 (1996), pp. 560–568.
[33] Y. Shkolnisky, A. Averbuch, M. Israeli, R. Coifman, and D. Donoho, 2*d Fourier Based Discrete Radon Transform*, preprint.
[34] S. Smale and D.-X. Zhou, *Shannon sampling and function reconstruction from point values*, Bull. Amer. Math. Soc. (N.S.), 41 (2004), pp. 279–305.
[35] G. Steidl, *A note on fast Fourier transforms for nonequispaced grids*, Adv. Comput. Math., 9 (1998), pp. 337–352.
[36] E. M. Stein and G. Weiss, *Introduction to Fourier Analysis on Euclidean Spaces*, Princeton Math. Ser. 32, Princeton University Press, Princeton, NJ, 1971.
[37] T. Strohmer, *Computationally attractive reconstruction of band-limited images from irregular samples*, IEEE Trans. Image Process., 6 (1997), pp. 540–548.
[38] T. Strohmer, *Numerical analysis of the non-uniform sampling problem*, J. Comput. Appl. Math., 122 (2000), pp. 297–316.
[39] T. Strohmer, T. Binder, and M. Süssner, *How to recover smooth object boundaries in noisy medical images*, in Proceedings of ICIP'96 (International Conference on Image Processing), 1996, pp. 331–334.

SIAM J. MATH. ANAL.
Vol. 36, No. 3, pp. 796–814

ON THE TWO-DIMENSIONAL HYDROSTATIC NAVIER–STOKES EQUATIONS*

DIDIER BRESCH†, ALEXANDRE KAZHIKHOV‡, AND JÉRÔME LEMOINE§

Abstract. This paper concerns some mathematical results on the two-dimensional hydrostatic equations, also called the primitive equations. The uniqueness of weak solutions of the two-dimensional Navier–Stokes equations is well known. Such a result is not known on the two-dimensional hydrostatic equations with Dirichlet boundary condition on the bottom. These equations are derived from the Navier–Stokes equations replacing the vertical component of the momentum equations by the hydrostatic equation on the pressure. We give here some partial answers on the uniqueness, the global strong existence of solutions, and the exponential decay in time of the energy. We assume a basin with a strictly positive depth. The degenerate case, in which the depth vanishes on the shore, remains open.

Key words. thin domains, geophysics, hydrostatic equation, global existence and uniqueness, overdetermined and underdetermined equations

AMS subject classifications. 35Q30, 35B40, 76D05

DOI. 10.1137/S0036141003422242

1. Introduction.

The hydrostatic Navier–Stokes equations correspond to the primitive equations used in oceanography that assume the density to be constant; see [10], [11]. They are obtained from the Navier–Stokes equations with anisotropic viscosity by an asymptotic analysis as the aspect ratio of the domain $\delta = \text{depth/width}$ tends to 0. The reader interested in such asymptotic analysis is referred, for example, to [1], [9]. We also mentioned the different works [12], [13], where they study the hydrostatic Navier–Stokes equations in thin rectangular domains and spherical domains. See also [2], [5] for the hydrostatic Euler equations in the two-dimensional case.

The uniqueness of weak solutions of two-dimensional Navier–Stokes equations is well known. See, for instance, [4] for a survey of uniqueness results and related fields on Navier–Stokes equations. This is not the case for the two-dimensional hydrostatic Navier–Stokes equations. Some partial results have been obtained in several recent papers, such as [6], [3].

The first paper [6] concerns the global strong solutions (u, p) of the hydrostatic equations and a weak strong uniqueness result in two and three space dimensions with Dirichlet condition on the bottom. In the two-dimensional case, they prove that if one has two weak solutions $u_1 = (v_1, w_1)$ and $u_2 = (v_2, w_2)$ of the two-dimensional hydrostatic equations such that v_1 satisfies $\partial_z v_1 \in L^4(0, T; L^2(\Omega))$, these solutions coincide. Moreover they prove a global strong existence result, assuming the data are small enough.

The second paper [3] concerns the existence and uniqueness of weak solutions, assuming that the vertical derivative of the initial data is square integrable and using

*Received by the editors February 3, 2003; accepted for publication (in revised form) February 6, 2004; published electronically October 14, 2004.
http://www.siam.org/journals/sima/36-3/42224.html
†Laboratoire de Modélisation et Calcul, IMAG-CNRS, U.M.R. 5523, Université Joseph Fourier, 38041 Grenoble, France (Didier.Bresch@imag.fr).
‡Lavrentyev Institute of Hydrodynamics, Siberian Branch of Russian Academy of Sciences, Russia (kazhikhov@hydro.nsc.ru).
§Laboratoire de Mathématiques Appliquées, U.M.R. 6620, Unversité Blaise Pascal, Avenue des Landais, 63177 Aubière, France (Jerome.Lemoine@math.univ-bpclermont.fr).

a Chezye condition on the bottom and on the surface, which means a condition of the type $\partial_z v = \alpha v$ with $u \cdot n = 0$. The domain is assumed to be with a depth vanishing on the shore. Regularity on the pressure has to be derived and Hardy's inequality must be used. Such a Chezye condition is used, for instance, in limnology; see, for instance, [8]. This kind of boundary condition is important in the proof. Let us remark that the problem of global existence and uniqueness with the same assumption as for the Navier–Stokes equations is an open problem.

The reader interested in some regularity results concerning the stationary hydrostatic Stokes equations is referred to [14]. We give here a regularity result on the nonstationary hydrostatic Navier–Stokes equations.

We consider the case with a Dirichlet condition on the bottom. At first, we establish a global existence result on v such that v and $\partial_z v$ have a weak regularity. The key of the proof is to find an appropriate boundary condition for $\partial_z v$ on the bottom using the hydrostatic condition and the fact that the integration of v with respect to the vertical coordinate is equal to 0, and we deduce existence of a global strong solution assuming initial data regular enough but without assuming that the data are small enough, as was done in [6]. In conclusion, we give some results related to the uniqueness of weak solutions. Note that our results remain valid if variable density or temperature is introduced.

2. The model. Let s (the horizontal section) denote an open interval and let $h : [0,1] \to \mathbb{R}_+$ denote a nonnegative continuous function on $[0,1]$ with $h \geq c > 0$. Let us consider the two-dimensional domain Ω defined by

$$\Omega = \{(x,z) : x \in (0,1), -h(x) < z < 0\}.$$

The boundary of the domain is $\partial\Omega = b \cup s \cup \bar{l}$, where the bottom b is defined by

$$b = \{(x, -h(x)) \in \mathbb{R}^2 : x \in (0,1)\}.$$

The surface s and lateral side l are given by

$$s = \{(x,0) : x \in (0,1)\}, \quad l = \{(0,z) : z \in (-h(0),0)\} \cup \{(1,z) : z \in (-h(1),0)\}.$$

Let us consider that the velocity (v,w) and the pressure p satisfy the following problem in Ω:

$$\begin{cases} \partial_t v - \nu\Delta v + v\partial_x v + w\partial_z v + \partial_x p = 0, \\ \partial_z p = 0, \\ \partial_x v + \partial_z w = 0. \end{cases} \tag{1}$$

System (1) is supplemented with the boundary condition on the surface,

$$\partial_z v = 0, \qquad w = 0 \text{ on } s,$$

the Dirichlet condition on the bottom,

$$(v,w) = 0 \text{ on } b,$$

and the following condition on the lateral wall side:

$$v = 0 \text{ on } l.$$

In addition, we consider the initial data $u|_{t=0} = (v_0, w_0)$ with $w_0(x,z) = \int_z^0 \partial_x v_0(x,\xi)\, d\xi$ and $\int_{-h}^0 v_0\, dz = 0$. In fact w is given from v by the relation

$$w = -\int_0^z \partial_x v.$$

This will give a system on (v,p) that we will write, in the next section, in another form (2).

Let us remark that the boundary condition on the lateral-side walls corresponds to a condition on the normal velocity and gives a condition on $\partial_z v$. Indeed $v = 0$ on l implies $\partial_z v = 0$ on the vertical lateral sides l.

Another way to write the system. Throughout the paper, we choose the viscosity ν equal to 1. Then we can write the system (1) as follows. The horizontal velocity v has to satisfy the following system:

$$\text{(2)} \qquad \begin{cases} \partial_t v - \Delta v + v\partial_x v + \left(\displaystyle\int_z^0 \partial_x v\right) \partial_z v + \partial_x p = 0 \text{ in } \Omega, \\ \displaystyle\int_{-h}^0 v\, dz = 0 \text{ in } s, \\ \partial_z v = 0 \text{ on } s, \qquad v = 0 \text{ on } b \cup l, \\ v|_{t=0} = v_0. \end{cases}$$

The vertical velocity w will be given by

$$w = \int_z^0 \partial_x v.$$

We remark that system (2) concerns only the horizontal component of the velocity.

3. Main results. This section is devoted to the functional setting of the hydrostatic equations (2) and to the main results. We introduce the space

$$\mathcal{V} = \left\{ \varphi \in C^\infty_{b,l}(\overline{\Omega}) : \langle \varphi \rangle = 0 \text{ in } s \right\},$$

where $\langle \varphi \rangle = \int_{-h}^0 \varphi\, dz$ and $\mathcal{C}^\infty_{b\,l}(\overline{\Omega})$ denotes the space of all smooth ($\mathcal{C}^\infty$) functions on $\overline{\Omega}$ that vanish in a neighborhood of $l \cup b$. Then the space H (resp., V) is the closure of $\mathcal{V}$ in $L^2(\Omega)$ (resp., $H^1(\Omega)$). We can easily check that

$$H = \left\{ \varphi \in L^2(\Omega) : \langle \varphi \rangle = 0 \text{ in } s \right\}, \quad V = \left\{ \varphi \in H^1(\Omega) : \langle \varphi \rangle = 0 \text{ in } s,\, \varphi = 0 \text{ on } b \cup l \right\}.$$

The goal of this paper is to prove the following theorems.

THEOREM 1. *Let us assume $v_0 \in H$ with $\partial_z v_0 \in L^2(\Omega)$ and $h \in W^{2,\infty}(0,1)$ with $h \geq c > 0$. Then there exists a unique global solution v of system (2) such that*

$$v \in L^2(0,\infty; V) \cap L^\infty(0,\infty; H)$$

and

$$\partial_z v \in L^2(0,\infty; H^1(\Omega)) \cap L^\infty(0,\infty; L^2(\Omega)). \qquad \square$$

The definition of a global solution is a weak solution in the classical way (see, for instance, [6]) such that $\partial_z v$ has a weak regularity.

Remark. It is important to note that, in Theorems 1 and 2, if $h'' < 0$ (which means that Ω is convex), then h'' must not be assumed to be bounded. See the estimate (39) below.

Moreover, we can prove the following regularity result.

THEOREM 2. *Let us assume $v_0 \in V$ and $h \in W^{2,\infty}(0,1)$ with $h \geq c > 0$. Then there exists a unique global strong solution v of system* (2) *such that*

$$\begin{aligned} &v \in L^2(0,\infty; H^2(\Omega) \cap V) \cap L^\infty(0,\infty; V), \\ &\partial_t v \in L^2(0,\infty; H), \\ &\partial_x p \in L^2(0,\infty; L^q(\Omega)) \text{ for all } q < \infty. \end{aligned}$$ ∎

A global strong solution is a weak solution with the previous extra regularities, the pressure being given by the de Rham theorem as usual for the hydrostatic equation; see, for instance, [14].

Using the previous theorem and the energy estimates, we can prove the following exponential decay in time of the solution.

COROLLARY 3. *Let us assume $v_0 \in V$ and $h \in W^{2,\infty}(0,1)$ with $h \geq c > 0$. Then the energy of the unique global strong solution v of system* (2) *decays exponentially fast in time. This means that there exists t_0 such that for all $t \geq t_0$,*

$$\|\nabla v(t)\|_{(L^2(\Omega))^2} \leq c_1 \|\nabla v_0\|_{(L^2(\Omega))^2} \exp(-c_2 t),$$

with c_1 and c_2 two nonnegative constants. ∎

At the end, we consider the domain $\Omega = (0,1) \times (0,\pi)$. We define the space $L^2_x H^\alpha_z$ by

$$L^2_x H^\alpha_z = \left\{ v = \sum_{k=1}^{\infty} a_k(x) \varphi_k(z) : \left(\sum_{k=1}^{\infty} \|\lambda_k^{\alpha/2} a_k(x)\|^2_{L^2(0,1)} \right)^{1/2} < \infty \right\},$$

with $\alpha = \pm 1$ and $\{\varphi_k, \lambda_k\}$ the L^2_z orthogonal basis associated to the Stokes operator

$$\begin{cases} \partial_z^2 \varphi_k - \partial_x p_k + \lambda_k^2 \varphi_k = 0, \qquad \partial_z p_k = 0, \\ \displaystyle\int_0^\pi \varphi_k \, dz = 0, \\ \partial_z \varphi_k|_{z=\pi} = 0, \quad \varphi_k|_{z=0} = 0. \end{cases}$$

We prove the following result.

THEOREM 4. *Let us assume $v_0 \in L^2_x H^{1/2}_z$. Then there exists a unique global weak solution v of the system* (2) *such that*

$$\partial_z v \in L^4(0,\infty; L^2(\Omega)),$$

the boundary condition on $\partial_z v$ being satisfied in a weak sense. ∎

We divide the proofs into three sections. Section 4 is devoted to the case of a domain with constant depth ($h = 1$), for didactic purposes. We will also prove Theorems 1 and 2 and Corollary 3 in this section. Section 5 concerns the case where h is a nonconstant function. We will only prove the compatibility condition and the

energy estimate on $\partial_z v$, the rest being similar to the case of a constant depth. In the last section, we look at the uniqueness result and prove that $v_0 \in L^2_x H^{1/2}_z$ ensures uniqueness.

Remark. We can assume to have a traction condition $\partial_z v = f$ on s, as was done in [6]. Then it is possible to establish the same kind of global existence and uniqueness result with some regularity assumptions on f.

Remark. If we consider the homogeneous Chezye condition $\partial_z v = 0$ on b instead of the Dirichlet condition, all previous results remain true.

It would be interesting to consider a depth vanishing on the shore, i.e., $h(0) = h(1) = 0$. This will be done in a forthcoming work since it seems difficult to control the left-hand side of (39), which gives the weak regularity of $\omega = \partial_z v$.

4. Domain with a constant depth. In this section we consider $h = 1$. The system (2) is equivalent to the systems

$$\begin{cases} \partial_t \omega - \Delta \omega + v \partial_x \omega + w \partial_z \omega = 0, \\ \omega = 0 \text{ on } s \cup l, \\ \omega|_{t=0} = \partial_z v_0 \end{cases} \tag{3}$$

and

$$\begin{cases} \partial_z^2 \Psi = \omega, \\ \Psi|_{\partial\Omega} = 0, \\ \partial_z \Psi|_b = 0, \end{cases} \tag{4}$$

with $v = \partial_z \Psi$. The first system comes from the derivative, with respect to z, of the momentum equation satisfied by v using the fact that $\partial_z p = 0$. The second one comes from the fact that $(v, w) = (\partial_z \Psi, -\partial_x \Psi)$. The boundary condition on Ψ on s and b comes from the fact that $w = 0$ on s and b, $v = 0$ on ℓ, $w = -\partial_x \Psi$, and $v = \partial_z \Psi$.

Remark. We note that normally $\Psi|_{\partial\Omega}$ must be equal to a constant $c(t)$. We choose $c(t) = 0$ since u is defined from Ψ by its derivative with respect to z. We then choose a particular stream function.

The existence proof is classically obtained using a Faedo–Galerkin method on ω after finding an appropriate boundary condition on $\omega = \partial_z v$.

A compatibility condition. The first system is underdetermined and the second one is overdetermined. Let us show how to obtain a compatibility condition between ω and Ψ to ensure existence of global strong solution.

A simple proof. At first, let us integrate the momentum equation $(2)_1$ with respect to z from -1 to 0. We get, using the boundary condition satisfied by v and the fact that $\int_{-1}^{0} \partial_x p dz = \partial_x p$, the following equality:

$$\partial_x p = -\omega|_{z=-1} - \int_{-1}^{0} \partial_x |v|^2 \, dz. \tag{5}$$

Moreover we have $\partial_x p|_{z=-1} = \partial_x p$; therefore by taking the trace on the bottom of equation $(2)_1$, we get

$$\partial_x p = \partial_z \omega|_{z=-1}. \tag{6}$$

Using (5)–(6), this gives

$$\partial_z\omega|_{z=-1} + \omega|_{z=-1} + 2\int_{-1}^{0} w\omega dz = 0.$$

This boundary condition will be used to ensure an energy estimate on the vorticity ω.

Another way to prove the compatibility condition. We integrate equation (4) with respect to z, obtaining

$$\Psi_z = \int_{-1}^{z} \omega(t,x,\xi)\,d\xi$$

since $\Psi_z = 0$ on b. Therefore

$$\Psi = \int_{-1}^{z}\left(\int_{-1}^{y}\omega(t,x,\xi)\,d\xi\right)dy = \int_{-1}^{z}(z-\xi)\omega(t,x,\xi)\,d\xi$$

since $\Psi = 0$ on b. Thus to ensure $\Psi|_{z=0} = 0$, we imply

$$\int_{-1}^{0}\xi\omega(t,x,\xi)\,d\xi = 0.$$

So we have to provide

$$\int_{-1}^{0} z\omega\,dz = 0 \text{ for all } t > 0.$$

This gives

$$\int_{-1}^{0} z\partial_t\omega\,dz = 0$$

and

$$\int_{-1}^{0} z\partial_x^2\omega\,dz = 0.$$

Therefore, using equation $(3)_1$, we find

$$\int_{-1}^{0} z\big(\partial_z^2\omega - \partial_z(\partial_x|v|^2) + \partial_z^2(w\,v)\big)\,dz = 0,$$

and thus, integrating by parts, since

$$\int_{-1}^{0} z\partial_z^2(w\,v)\,dz = 0,$$

we get

$$\partial_z\omega|_{z=-1} + \omega|_{z=-1} + 2\int_{-1}^{0} w\omega\,dz = 0. \tag{7}$$

This gives the result. □

A priori estimates. Let us give the a priori estimates which allow us to prove the global existence result.

Energy estimate on v. Choosing v as a test function in $(2)_1$, we get

$$\frac{1}{2}\frac{d}{dt}\|v\|^2_{L^2(\Omega)} + \|\nabla v\|^2_{(L^2(\Omega))^2} \leq 0. \tag{8}$$

Energy estimate on ω. We choose now ω as a test function in $(3)_1$. Using the compatibility condition (7), we get

$$\frac{d}{dt}\|\omega\|^2_{L^2(\Omega)} + \|\nabla\omega\|^2_{(L^2(\Omega))^2} \leq \int_0^1 |\omega|_{z=-1}|^2\,dx + 2I, \tag{9}$$

with

$$I = \left| \int_0^1 \left(\omega|_{z=-1} \int_{-1}^0 w\omega dz \right) dx \right|. \tag{10}$$

Let us prove that this inequality joint with the weak regularity of v will give the weak estimate on ω. At first we remark that

$$\left(\omega|_{z=-1}\right)^2 = -2\int_{-1}^0 \omega\partial_z\omega\,dz \leq c\|\omega\|_{L^2_z}\,\|\partial_z\omega\|_{L^2_z},$$

so

$$\left|\omega|_{z=-1}\right| \leq c\|\omega\|^{1/2}_{L^2_z}\|\partial_z\omega\|^{1/2}_{L^2_z}. \tag{11}$$

Thus the first term in the right-hand side of (9) is bounded as follows:

$$\int_0^1 |\omega|_{z=-1}|^2\,dx \leq \varepsilon\|\nabla\omega\|^2_{(L^2(\Omega))^2} + c\|\omega\|^2_{L^2(\Omega)}. \tag{12}$$

Moreover we have

$$\|\omega\|^2_{L^\infty_x L^2_z} \leq 2\int_0^1\int_{-1}^0 |\omega\partial_x\omega|dzdx \leq c\|\omega\|_{L^2(\Omega)}\|\partial_x\omega\|_{L^2(\Omega)};$$

then

$$\|\omega\|^{1/2}_{L^\infty_x L^2_z} \leq c\|\omega\|^{1/4}_{L^2(\Omega)}\|\partial_x\omega\|^{1/4}_{L^2(\Omega)}. \tag{13}$$

Therefore, using (10), (11), and (13), we get

$$I \leq c\|\omega\|^{1/4}_{L^2(\Omega)}\|\partial_x\omega\|^{1/4}_{L^2(\Omega)} \int_0^1 \left(\|\partial_z\omega\|^{1/2}_{L^2_z} \left| \int_{-1}^0 w\omega dz \right| \right) dx. \tag{14}$$

Since $\omega = \partial_z v$, we get

$$\left|\int_{-1}^0 w\omega\,dz\right| = \left|\int_{-1}^0 w\partial_z v\,dz\right| = \left|\int_{-1}^0 v\partial_z w\,dz\right| \leq \|v\|_{L^2_z}\|\partial_z w\|_{L^2_z}. \tag{15}$$

Moreover

$$\|\partial_z w\|^2_{L^2_z} = -\int_{-1}^0 w\partial_z^2 w\,dz = \int_{-1}^0 w\partial_x\omega\,dz \leq \|w\|_{L^2_z}\|\partial_x\omega\|_{L^2_z}.$$

So, using (15), we get

$$\left|\int_{-1}^{0} w\omega\right| \le \|v\|_{L_z^2}\|w\|_{L_z^2}^{1/2}\|\partial_x\omega\|_{L_z^2}^{1/2}. \tag{16}$$

This implies, using (14) and (16), that

$$I \le c\|\omega\|_{L^2(\Omega)}^{1/4}\|\partial_x\omega\|_{L^2(\Omega)}^{1/4}\int_0^1 \left(\|\partial_z\omega\|_{L_z^2}^{1/2}\|\partial_x\omega\|_{L_z^2}^{1/2}\|v\|_{L_z^2}\|w\|_{L_z^2}^{1/2}\right) dx$$

and thus

$$I \le c\|\omega\|_{L^2(\Omega)}^{1/4}\|\partial_x\omega\|_{L^2(\Omega)}^{1/4}\int_0^1 \left(\|\nabla\omega\|_{L_z^2}\|v\|_{L_z^2}\|w\|_{L_z^2}^{1/2}\right) dx. \tag{17}$$

Using the Hölder inequality on (17), we find

$$I \le c\|\omega\|_{L^2(\Omega)}^{1/4}\|\nabla\omega\|_{(L^2(\Omega))^2}^{1/4}\|\nabla\omega\|_{(L^2(\Omega))^2}\|w\|_{L^2(\Omega)}^{1/2}\left(\int_0^1 \|v\|_{L_z^2}^4\, dx\right)^{1/4}. \tag{18}$$

Next, we bound

$$\int_0^1 \|v\|_{L_z^2}^4\, dx \le \|v\|_{L^2(\Omega)}^2\|v\|_{L_x^\infty L_z^2}^2$$

and

$$\|v\|_{L_x^\infty L_z^2}^2 \le c\|v\|_{L^2(\Omega)}\|\partial_x v\|_{L^2(\Omega)}.$$

So we get

$$I \le c\|\omega\|_{L^2(\Omega)}^{1/4}\|\nabla\omega\|_{(L^2(\Omega))^2}^{5/4}\|v\|_{L^2(\Omega)}^{1/4}\|\partial_x v\|_{L^2(\Omega)}^{1/4}\|v\|_{L^2(\Omega)}^{1/2}\|w\|_{L^2(\Omega)}^{1/2},$$

which implies

$$I \le \varepsilon\|\nabla\omega\|_{(L^2(\Omega))^2}^2 + c\|v\|_{L^2(\Omega)}^2\|\omega\|_{L^2(\Omega)}^{2/3}\|\partial_x v\|_{L^2(\Omega)}^{2/3}\|w\|_{L^2(\Omega)}^{4/3}.$$

This gives, using inequalities (9) and (12),

$$\frac{d}{dt}\|\omega\|_{L^2(\Omega)}^2 + \frac{1}{2}\|\nabla\omega\|_{(L^2(\Omega))^2}^2 \le c\|v\|_{L^2(\Omega)}^2\|\partial_x v\|_{L^2(\Omega)}^{2/3}\|w\|_{L^2(\Omega)}^{4/3}\|\omega\|_{L^2(\Omega)}^{2/3} + c\|\omega\|_{L^2(\Omega)}^2, \tag{19}$$

with c independent on t.

Weak strong solution. The first inequality, (8), gives, if $v_0 \in H$, that

$$v \in L^2(0,\infty;V) \cap L^\infty(0,\infty;H).$$

Assuming that $v_0 \in L^2(\Omega)$ and $\partial_z v_0 \in L^2(\Omega)$, we see that

$$\|v\|_{L^2(\Omega)}^2\|\partial_x v\|_{L^2(\Omega)}^{2/3}\|w\|_{L^2(\Omega)}^{4/3} \in L^1(0,\infty).$$

Inequality (19) will give the weak regularity on $\partial_z v$, meaning that

$$\partial_z v \in L^\infty(0,\infty;L^2(\Omega)) \cap L^2(0,\infty;H^1(\Omega)).$$

Indeed $\|\omega\|^2_{L^2(\Omega)} = \|\partial_z v\|^2_{L^2(\Omega)} \in L^1(0,\infty)$. The verification is straightforward, using the weak regularity of v, the exponential decaying property of $\|v(t)\|_{L^2(\Omega)}$ (which will be proved later on), and the relation between ω, $\nabla\omega$, and v. Indeed we have $v \in L^\infty(0,\infty;L^2(\Omega))$ and

$$\int_0^\infty \|\partial_x v\|^{2/3}_{L^2(\Omega)} \|w\|^{4/3}_{L^2(\Omega)}\, dt \le \left(\int_0^\infty \|v_x\|^2_{L^2(\Omega)}\, dt\right)^{1/3} \left(\int_0^\infty \|w\|^2_{L^2(\Omega)}\, dt\right)^{2/3}.$$

Remark. Therefore (5) implies that $\partial_x p \in L^2(0,\infty;L^1_x L^\infty_z)$.

Uniqueness. This energy estimate on $\partial_z v$ allows us to prove the existence of a weak strong solution when $v_0 \in H$ and $\partial_z v_0 \in L^2(\Omega)$; this means a solution such that

$$v \in L^2(0,\infty;V) \cap L^\infty(0,\infty;H),$$
$$\partial_z v \in L^2(0,\infty;H^1(\Omega)) \cap L^\infty(0,\infty;L^2(\Omega)).$$

The uniqueness follows from the fact that if we consider two weak solutions, one of which has the regularity $\partial_z v \in L^4(0,T;L^2(\Omega))$, then they coincide; see [6]. Denoting v_1 and v_2 as the two weak solutions, with v_2 satisfying the extra regularity and $w_1 = -\int_0^z v_1$ and $w_2 = -\int_0^z v_2$ making the difference of the two momentum equations, this result is easy to check by classical energy estimates using anisotropic estimates on the nonlinear quantity

$$I_1 = \int_\Omega (v_1 - v_2)^2 \partial_x v_2 + (w_1 - w_2)\partial_z v_2 (v_1 - v_2).$$

More precisely, we use the fact that

$$I_1 \le \|v_1 - v_2\|_{L^2_x L^\infty_z} \|v_1 - v_2\|_{L^\infty_x L^2_z} \|\partial_x v_2\|_{L^2} + \|v_1 - v_2\|_{L^\infty_x L^2_z} \|\partial_z v_2\|_{L^2} \|w_1 - w_2\|_{L^2_x L^\infty_z},$$

and therefore

$$I_1 \le \frac{\nu}{2} \int_\Omega (\partial_x (v_1 - v_2))^2 + c_1 (\|\partial_x v_2\|^2_{L^2(\Omega)} + \|\partial_z v_2\|^4_{L^2(\Omega)}) \|v_1 - v_2\|^2_{L^2(\Omega)}.$$

Therefore we obtain the first global existence and uniqueness result with the Dirichlet boundary condition at the bottom for a domain with a constant depth. We do not have to assume a smallness assumption on the data. We will prove in the next section how to obtain the same result for a domain with $h \ge c > 0$ and $h'' \in W^{2,\infty}(0,1)$.

Exponential decay in time on $\|v\|_{L^2(\Omega)}$ and $\|\omega\|_{L^2(\Omega)}$. Equation (8) gives the exponential decay in time of $\|v\|^2_{L^2(\Omega)}$. Therefore, denoting $y = \|\omega\|^2_{L^2(\Omega)}$ and $f = \|\partial_x v\|^{2/3}_{L^2(\Omega)} \|w\|^{4/3}_{L^2(\Omega)} \in L^1(0,\infty)$, equation (19) gives

$$y' + cy \le c \exp(-ct) f(t) y^{\frac{1}{3}} + c\|\omega\|^2_{L^2(\Omega)}.$$

Now, using the inequality $\|\omega\|^2_{L^2(\Omega)} \le \|v\|_{L^2(\Omega)} \|\partial_z \omega\|_{L^2(\Omega)}$, we obtain

$$c\|\omega\|^2_{L^2(\Omega)} \le c y^{\frac{1}{3}} \exp(-ct) g(t),$$

with $g(t) \in L^1(0,\infty)$ (since for all $\alpha > 0$, $\|v\|^\alpha_{L^2(\Omega)} \in L^1(0,\infty)$ and $\|\partial_z \omega\|_{L^2(\Omega)} \in L^2(0,\infty)$). Thus we have

$$y' + cy \le c \exp(-ct) h(t) y^{\frac{1}{3}},$$

with $h(t) = f(t) + g(t) \in L^1(0,\infty)$, and therefore $y = \|\omega\|^2_{L^2(\Omega)}$ decays exponentially fast in time.

Regularity. We can prove, if we assume $v_0 \in V$, that there exists a unique weak solution such that

$$v \in L^2(0,\infty;H^2(\Omega)\cap V)\cap L^\infty(0,\infty;V), \qquad \partial_t v \in L^2(0,\infty;H),$$
$$\partial_x p \in L^2(0,\infty;L^q(\Omega)) \text{ for all } q<\infty.$$

Let us recall that if $v_0 \in H$ with $\partial_z v_0 \in L^2(\Omega)$, then we have weak estimates on $\omega = \partial_z v$. Using $\partial_t v$ as a test function, we get $\partial_t v \in L^2(0,\infty;L^2(\Omega))$ and $v \in L^\infty(0,\infty;H^1(\Omega))$. Therefore (5) gives $\partial_x p \in L^2(0,\infty;L^2(\Omega))$. Then we obtain by regularity $v \in L^2(0,\infty;H^2(\Omega))$. Now using the regularity on v, this gives the regularity on $\partial_x p$ with the help of (5).

Regularity on $\partial_x p$. Since $v \in L^\infty(0,\infty;V)$, we get $v \in L^\infty(0,\infty;L^\infty_x L^2_z)$. On the one hand, $\partial_x v \in L^2(0,\infty;H^1(\Omega))$, and then $\partial_x v \in L^2(0,\infty;L^\infty_x L^2_z)$. Then $\int_{-1}^0 \partial_x |v|^2\,dz \in L^2(0,\infty;L^\infty_x)$. This gives

$$\int_{-1}^0 \partial_x |v|^2\,dz \in L^2(0,\infty;L^\infty(\Omega)).$$

On the other hand $\omega = \partial_x v \in L^2(0,\infty;H^1(\Omega))$, and then $\omega|_{z=-1} \in L^2(0,\infty;H^{\frac12}(b))$. Since $H^{\frac12}(b) \subset L^q(b)$ for all $q \in [1,+\infty[$ $(d=2)$, this gives, for all $q \in [1,+\infty[$, $\omega|_{z=-1} \in L^2(0,\infty;L^q(b))$; then

$$\omega|_{z=-1} \in L^2(0,\infty;L^q(\Omega)).$$

Therefore, using (5), this ends the proof.

Regularity on $\partial_t v$ and ∇v. For the reader's convenience, let us give the estimate on $\partial_t v$ and on ∇v, for which we will use the regularity obtained on $\partial_z v$. We multiply the momentum equation satisfied by v by $\partial_t v$, giving

$$\int_\Omega |\partial_t v|^2 + \frac12 \frac{d}{dt}\int_\Omega |\nabla v|^2 + \int_\Omega (v\partial_x v\partial_t v + w\partial_z v\partial_t v) = 0. \tag{20}$$

Let us estimate the two last terms. We have

$$\left|\int_\Omega v\partial_x v\partial_t v\right| \le \varepsilon \|\partial_t v\|^2_{L^2(\Omega)} + c\|v\partial_x v\|^2_{L^2(\Omega)}. \tag{21}$$

Since

$$\|v\partial_x v\|^2_{L^2(\Omega)} \le \|\partial_x v\|^2_{L^\infty_z L^2_x}\|v\|^2_{L^2_z L^\infty_x},$$

we get, using (13),

$$\|v\partial_x v\|^2_{L^2(\Omega)} \le c\|\partial_x\omega\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)}\|\partial_x v\|^2_{L^2(\Omega)}. \tag{22}$$

Let us now look at the last term of (20). We have

$$\left|\int_\Omega w\partial_z v\partial_t v\right| \le \varepsilon\|\partial_t v\|^2_{L^2(\Omega)} + c\|w\omega\|^2_{L^2(\Omega)}. \tag{23}$$

Since

$$\|w\omega\|^2_{L^2(\Omega)} \le \int_0^1 \left(\|w\|^2_{L^\infty_z}\int_{-1}^0 |\omega|^2\,dz\right)dx,$$

we get

$$\|w\omega\|^2_{L^2(\Omega)} \le \|\omega\|^2_{L^\infty_x L^2_z}\|\partial_x v\|^2_{L^2(\Omega)},$$

and therefore

$$\|w\omega\|^2_{L^2(\Omega)} \le \|\omega\|_{L^2(\Omega)}\|\partial_x\omega\|_{L^2(\Omega)}\|\partial_x v\|^2_{L^2(\Omega)}. \tag{24}$$

Now using (20)–(24), we find

$$\int_\Omega |\partial_t v|^2 + \frac{d}{dt}\int_\Omega |\nabla v|^2 \le c(t)\int_\Omega |\nabla v|^2, \tag{25}$$

with

$$c(t) = \|\partial_x\omega\|_{L^2(\Omega)}(\|\omega\|_{L^2(\Omega)} + \|v\|_{L^2(\Omega)}) \in L^1(0,\infty).$$

This gives $\partial_t v \in L^2(0,\infty;L^2(\Omega))$ and $v \in L^\infty(0,\infty;H^1(\Omega))$ if $v_0 \in H^1(\Omega)$.

Let us now look at the regularity $v \in L^2(0,\infty;H^2(\Omega))$. We have

$$-\Delta v = -\partial_t v - \partial_x p - v\partial_x v - w\partial_z v.$$

Using the regularity on $\partial_t v$, v, and $\partial_x p$, we get an elliptic equations with a right-hand side in $L^2(0,\infty;L^2(\Omega))$. Therefore, using [14], $v \in L^2(0,\infty;H^2(\Omega))$. Now using the expression (5) of $\partial_x p$, we conclude that $\partial_x p \in L^2(0,\infty;L^q(\Omega))$ for all $q < \infty$.

Exponential decay in time on $\|\nabla v\|_{(L^2(\Omega))^2}$. Using the energy inequalities (19), (25), we can prove the exponential decay in time of $\|\nabla v\|^2_{(L^2(\Omega))^2}$ since $\|v\|^2_{L^2(\Omega)}$ and $\|\omega\|^2_{L^2(\Omega)}$ decay exponentially fast. Let us give the proof for the reader's convenience. We have

$$\int_\Omega |\partial_t v|^2 + \frac{d}{dt}\int_\Omega |\nabla v|^2 \le c(t)\int_\Omega |\nabla v|^2, \tag{dec2}$$

with

$$c(t) = \|\partial_x\omega\|_{L^2(\Omega)}(\|\omega\|_{L^2(\Omega)} + \|v\|_{L^2(\Omega)})$$

and $\|v\|^2_{L^2(\Omega)}$ and $\|\omega\|^2_{L^2(\Omega)}$ with exponential decaying properties. Therefore, there exists $(c_1,c_2) \in (\mathbb{R}_+)^2$ such that

$$\int_\Omega |\partial_t v|^2 + \frac{d}{dt}\int_\Omega |\nabla v|^2 \le c_1 e^{-c_2 t}\|\partial_x\omega\|_{L^2(\Omega)}\int_\Omega |\nabla v|^2. \tag{dec3}$$

On the other hand, we have

$$\frac{d}{dt}\|v\|^2_{L^2(\Omega)} + \|\nabla v\|^2_{(L^2(\Omega))^2} \le 0; \tag{dec4}$$

then

$$\|\nabla v\|^2_{(L^2(\Omega))^2} \le \int_\Omega |\partial_t v|^2 + \int_\Omega |v|^2.$$

By reinjecting in (dec3), we get

$$\int_\Omega |\nabla v|^2 + \frac{d}{dt}\int_\Omega |\nabla v|^2 \le c_1 e^{-c_2 t}\|\partial_x\omega\|_{L^2(\Omega)}\int_\Omega |\nabla v|^2 + \int_\Omega |v|^2.$$

Since $\int_\Omega |v|^2 \le ce^{-c_3 t}$ with $c_3 \le 1$, then

$$B(t)\int_\Omega |\nabla v|^2 + \frac{d}{dt}\int_\Omega |\nabla v|^2 \le ce^{-c_3 t},$$

where $B(t) = 1 - c_1 e^{-c_2 t}\|\partial_x \omega\|_{L^2(\Omega)}$. Then we get

$$\frac{d}{dt}\left(\int_\Omega |\nabla v|^2 e^{\int_0^t B(\tau)\,d\tau}\right) \le ce^{\int_0^t B(\tau)\,d\tau - c_3 t}.$$

Estimation of $\int_0^t B(\tau)\,d\tau$. We have, for all $t > 0$,

$$\begin{aligned}\int_0^t e^{-c_2\tau}\|\partial_x\omega\|_{L^2(\Omega)}\,d\tau &\le \left(\int_0^t e^{-2c_2\tau}\,d\tau\right)^{\frac12}\left(\int_0^t \|\partial_x\omega\|^2_{L^2(\Omega)}\,d\tau\right)^{\frac12}\\ &\le \frac{1}{\sqrt{2c_2}}(1-e^{-2c_2 t})^{\frac12}\left(\int_0^t \|\partial_x\omega\|^2_{L^2(\Omega)}\,d\tau\right)^{\frac12}.\end{aligned}$$

Then for $t \ge 1$,

$$\int_0^t e^{-c_2\tau}\|\partial_x\omega\|_{L^2(\Omega)}\,d\tau \le te^{-c_2\xi}\left(\int_0^\infty \|\partial_x\omega\|^2_{L^2(\Omega)}\,d\tau\right)^{\frac12},$$

with $\lim_{t\to\infty}\xi = +\infty$; and then, for $t \ge t_0$,

$$\int_0^t e^{-c_2\tau}\|\partial_x\omega\|_{L^2(\Omega)}\,d\tau \le \frac{c_3}{2c_1}t.$$

In conclusion we get, for all $t \ge t_0$,

$$\left(1-\frac{c_3}{2}\right)t \le \int_0^t B(\tau)\,d\tau \le t.$$

This gives, for all $t \ge t_0$,

$$\frac{d}{dt}\left(\int_\Omega |\nabla v|^2 e^{\int_0^t B(\tau)\,d\tau}\right) \le ce^{(1-c_3)t};$$

then

$$\int_\Omega |\nabla v|^2(t) \le c\|\nabla v(t_0)\|^2_{L^2(\Omega)^2}(e^{(1-c_3)t}-1)e^{(-1+\frac{c_3}{2})t} \le c\|\nabla v(t_0)\|^2_{L^2(\Omega)^2}e^{-\frac{c_3}{2}t}$$

since $c_3 \le 1$. Using the fact that ∇v is bounded in $(L^2(0,T;L^2(\Omega))^9)$, this ends the proof. ∎

5. A domain with a nonconstant depth. Let us now consider a domain with a nonconstant depth, and let us give the compatibility condition, which will be more complicated than that for a domain with constant depth.

Before beginning to search for such compatibility condition, let us recall for the reader's convenience the definition of the anisotropic spaces $L^\infty_x L^2_z$.

Definition of the anisotropic space. A function belongs to $L^\infty_x L^2_z$ if

$$u(x,.) \in L^2(-h(x),0) \quad \text{and} \quad \|u(x,.)\|_{L^2(-h(x),0)} \in L^\infty.$$

Moreover its norm is given by

$$\|u\|_{L^\infty_x L^2_z} = \sup_{x\in(0,1)}\Big(\|u(x,.)\|_{L^2(-h(x),0)}\Big).$$

The compatibility condition. We will use the first kind of proof given in section 1 in the case of a constant depth. At first, let us integrate the momentum equation on v from $-h$ to 0 with x fixed. We get, using the boundary condition satisfied by v and the fact that $\int_{-h}^{0} \partial_x p \, dz = h\partial_x p$, the following equality:

$$h\partial_x p - \int_{-h}^{0} \partial_x^2 v \, dz + \omega|_{z=-h} + \int_{-h}^{0} (v\partial_x v + w\partial_z v) \, dz = 0. \tag{26}$$

Let us calculate the term on $\partial_x^2 v$. After some calculations and using that

$$\partial_x\big(v(x,-h(x))\big) = \big(\partial_x v\big)(x,-h(x)) - h' \, \big(\partial_z v\big)(x,-h(x)) \tag{27}$$

and $v = 0$ on b, we get

$$\int_{-h}^{0} \partial_x^2 v \, dz = -(h')^2 \, \omega|_b. \tag{28}$$

We have, using the divergence-free condition on $u = (v, w)$ and the boundary condition satisfied by w,

$$\int_{-h}^{0} (v\partial_x v + w\partial_z v) \, dz = \int_{-h}^{0} \partial_x |v|^2 \, dz = 2 \int_{-h}^{0} w\omega \, dz. \tag{29}$$

If we use (26), (28), and (29), we get

$$h\partial_x p + (1 + (h')^2)\omega|_b + 2 \int_{-h}^{0} \omega w \, dz = 0. \tag{30}$$

Let us now take the trace on the bottom of the momentum equation. Since $(v, w) = 0$ on b, we get

$$(\partial_t v + v\partial_x v + w\partial_z v)|_b = 0.$$

It remains that

$$\partial_x p = (\partial_x^2 v + \partial_z^2 v)|_b. \tag{31}$$

Let us calculate the right-hand side in terms of ω. We have by definition of ω

$$(\partial_z^2 v)|_b = (\partial_z \omega)|_b. \tag{32}$$

Let us look at the $\partial_x^2 v$. We have

$$\partial_x\big((\partial_x v)(x,-h(x))\big) = \big(\partial_x^2 v\big)(x,-h(x)) - h'\big(\partial_x \partial_z v\big)(x,-h(x)).$$

Therefore, using (27), we get

$$(\partial_x^2 v)(x,-h(x)) = \partial_x\big(h' \, \omega(x,-h(x))\big) + h'(\partial_x \omega)(x,-h(x)).$$

Thus we get

$$(\partial_x^2 v)|_b = 2h'(\partial_x \omega)|_b - (h')^2(\partial_z \omega)|_b + h''\omega|_b. \tag{33}$$

Thus using (31), (32), and (33), we get

$$\partial_x p = 2h'(\partial_x\omega)|_b + (1-(h')^2)(\partial_z\omega)|_b + h''\omega|_b. \tag{34}$$

Finally, using (26) and (34), we get the following compatibility condition:

$$\big(hh'' + \big(1+(h')^2\big)\big)\omega|_b + 2\int_{-h}^{0}(\omega w)\,dz + 2hh'(\partial_x\omega)|_b + h\big(1-(h')^2\big)(\partial_z\omega)|_b = 0. \tag{35}$$

Let us rewrite this condition in terms of the normal derivative and the tangential derivative of ω. Recall that

$$\partial\omega/\partial n = -\big(h'(\partial_x\omega)|_b + (\partial_z\omega)|_b\big)/\big(1+(h')^2\big)^{1/2}$$

and

$$\partial\omega/\partial\tau = \big((\partial_x\omega)|_b - h'(\partial_z\omega)|_b\big)/\big(1+(h')^2\big)^{1/2}.$$

Then the compatibility condition may be written as

$$-h\partial\omega/\partial n + hh'\partial\omega/\partial\tau + (1+h'^2)^{-1/2}\left(\big(hh'' + \big(1+(h')^2\big)\big)\omega|_b + 2\int_{-h}^{0}(\omega w)\,dz\right) = 0. \tag{36}$$

A priori estimates. Let us give the a priori estimates, which allows us to prove the global existence result.

Energy estimate on v. Taking v as a test function on the momentum equation, we get, as in the constant depth assumption,

$$\frac{d}{dt}\|v\|^2_{L^2(\Omega)} + \|\nabla v\|^2_{(L^2(\Omega))^2} \le 0. \tag{37}$$

Energy estimate on ω. Let us multiply by ω the equation satisfied by ω; we then get

$$\frac{d}{dt}\|\omega\|^2_{L^2(\Omega)} + \|\nabla\omega\|^2_{(L^2(\Omega))^2} = \int_b \frac{\partial\omega}{\partial n}\omega. \tag{38}$$

Let us remark that the compatibility condition (36) allows us to replace the normal derivative by the tangential derivative and $\omega|_b$. We get

$$\begin{aligned}\int_b \frac{\partial\omega}{\partial n}\omega = \int_b h'\frac{\partial\omega}{\partial\tau}\omega + \int_b (1+h'^2)^{-1/2}\left(h'' + \frac{1+(h')^2}{h}\right)|\omega|^2 \\ + 2\int_b \frac{1}{h\sqrt{1+h'^2}}\left(\int_{-h}^{0}(\omega w)\,dz\right)\omega,\end{aligned}$$

and using the fact that

$$\int_b h'\frac{\partial\omega}{\partial\tau}\omega = -\frac{1}{2}\int_b \frac{h''}{(1+h'^2)^{1/2}}|\omega|^2,$$

we finally obtain

$$\frac{d}{dt}\|\omega\|^2_{L^2(\Omega)} + \|\nabla\omega\|^2_{(L^2(\Omega))^2} \le \int_b \left(\frac{\sqrt{1+h'^2}}{h} + \frac{h''}{2\sqrt{1+h'^2}}\right)|\omega|^2 + 2\int_b \frac{1}{h\sqrt{1+h'^2}}\left(\int_{-h}^0 (\omega w)\,dz\right)\omega. \tag{39}$$

Remark. If we choose $h = 1$ in (39), we get exactly inequality (10) found in section 2.

Assuming that $h \in W^{2,\infty}(0,1)$, the first term in the right-hand side of (39) is bounded as follows:

$$\int_b \left(\frac{\sqrt{1+h'^2}}{h} + \frac{h''}{2\sqrt{1+h'^2}}\right)|\omega|^2 \le c\|\omega\|_{L^2(\Omega)}\|\nabla\omega\|_{(L^2(\Omega))^2},$$

with c depending only on h. The second term in the right-hand side of (39) is bounded, as was done in the case of constant depth, assuming that $h \ge c > 0$. This gives the global existence and uniqueness mentioned in Theorem 1.

The regularity and exponential decay in time given in Theorem 2 and Corollary 3 follow as in the case of constant depth. ∎

Remark. If $h'' < 0$, meaning Ω is convex, then we do not have to assume that $h \in W^{2,\infty}(0,1)$ since the term containing h'' in the right-hand side is therefore negative.

6. Some remarks on the uniqueness result. The result in this section is in the same spirit as [7], where they consider u_0 in $(H^{1/2}(\Omega))^3$ for the Navier–Stokes equations in the three-dimensional case.

We have recalled in the previous sections that we have uniqueness if we prove the regularity $\partial_z v \in L^4(0,\infty;L^2(\Omega))$. With $v_0 \in H$, we have a weak solution of system (2),

$$v \in L^2(0,\infty;V) \cap L^\infty(0,\infty;H),$$

and if we assume $v_0 \in H$ with $\partial_z v_0 \in L^2(\Omega)$, we get a weak solution of (2) such that

$$v \in L^2(0,\infty;V) \cap L^\infty(0,\infty;H),$$
$$\partial_z v \in L^2(0,\infty;H^1(\Omega)) \cap L^\infty(0,\infty;L^2(\Omega)).$$

This last regularity result is not optimal but just ensures uniqueness. Studying the linear problem, we see by interpolation that it is enough to suppose $v_0 \in L^2_x H^{1/2}_z$ to prove that $\partial_z v \in L^4(0,\infty;L^2(\Omega))$. This is what we want to prove in this section on the nonlinear system.

For didactic considerations, we begin by study the nonlinear problem with $\partial_z v|_{z=\pi} = 0$ and $\partial_z v|_{z=0} = 0$.

Nonlinear problem with $\partial_z v|_{z=\pi} = 0$ and $\partial_z v|_{z=0} = 0$. We write the initial data v_0 as $v_0 = \sum_{n=1}^\infty b_n(x)\cos(nz)$ and look at a solution v on the form $v = \sum_{n=1}^\infty a_n \cos(nz)$ with $a_n = a_n(x,t)$. Let us multiply equation $(2)_1$ satisfied by v by $\overline{\partial_z v} = \sum_{n=1}^\infty n a_n \cos(nz)$, and let us prove that it will give us the estimate on $\partial_z^{1/2} v$ if $v_0 \in L^2_x H^{1/2}_z$, meaning that

$$\partial_z^{1/2} v \in L^2(0,\infty;H^1(\Omega)) \cap L^\infty(0,\infty;L^2(\Omega)).$$

This will give

$$\partial_z v \in L^2(0,\infty;L^2_x H^{1/2}_z) \cap L^\infty(0,\infty;L^2_x H^{-1/2}_z),$$

and thus, by interpolation,

$$\partial_z v \in L^4(0,\infty;L^2(\Omega)).$$

This ensures the uniqueness.

We multiply equation $(2)_1$ by $\overline{\partial_z v}$ and obtain, since $\int_0^\pi \overline{\partial_z v}\,dz = 0$ and $\partial_z p = 0$,

$$\int_\Omega \partial_t v\overline{\partial_z v} + \int_\Omega \nabla v \cdot \nabla\overline{\partial_z v} + \int_\Omega v\partial_x v\overline{\partial_z v} + \int_\Omega w\partial_z v\overline{\partial_z v} = 0.$$

The third term reads

$$\int_\Omega v\partial_x v\overline{\partial_z v} = \sum_{l,m,n\geq 0}\int_0^1 a_l a'_m n\, a_n \int_0^\pi \cos(lz)\cos(mz)\cos(nz).$$

The only remaining terms are such that $l+m-n=0$, $l-m-n=0$, $l-m+n=0$. Therefore every time we can upperbound n by $n^{1/2}(l^{1/2}+m^{1/2})$. This will give

$$\begin{aligned}\left|\int_\Omega v\partial_x v\overline{\partial_z v}\right| &\leq \|\partial_z^{1/2} v\|_{L^2(\Omega)}\|\nabla\partial_z^{1/2} v\|_{(L^2(\Omega))^2}\|\partial_x v\|_{L^2(\Omega)} \\ &\quad + \|v\|^{1/2}_{L^2(\Omega)}\|\partial_z v\|^{1/2}_{L^2(\Omega)}\|\nabla\partial_z^{1/2} v\|^{3/2}_{(L^2(\Omega))^2}\|\partial_z^{1/2} v\|^{1/2}_{L^2(\Omega)}.\end{aligned} \tag{40}$$

We prove in the same manner that the fourth term is controlled as follows:

$$\begin{aligned}\left|\int_\Omega w\partial_z v\overline{\partial_z v}\right| &\leq \|\partial_x\partial_z^{1/2} v\|^{3/2}_{L^2(\Omega)}\|v\|^{1/2}_{L^2(\Omega)}\|\partial_z v\|^{1/2}_{L^2(\Omega)}\|\partial_z^{1/2} v\|^{1/2}_{L^2(\Omega)} \\ &\quad + \|\partial_x v\|_{L^2(\Omega)}\|\partial_z^{1/2} v\|_{L^2(\Omega)}\|\nabla\partial_z^{1/2} v\|_{(L^2(\Omega))^2}.\end{aligned} \tag{41}$$

These inequalities will provide us with the estimate

$$\frac{d}{dt}\|\partial_z^{1/2} v\|^2_{L^2(\Omega)} + \|\nabla\partial_z^{1/2} v\|^2_{(L^2(\Omega))^2} \leq c(t)\|\partial_z^{1/2} v\|^2_{L^2(\Omega)}, \tag{42}$$

with $c \in L^1(0,\infty)$. Therefore, assuming the regularity $v_0 \in L^2_x H^{1/2}_z$, we get the result.

Proof of estimate (41). We have

$$\int_\Omega w\partial_z v\overline{\partial_z v} = -\sum_{l,m,n\geq 0}\int_0^1 \frac{a'_l}{l} n a_n m\, a_m \int_0^\pi \sin(lz)\sin(nz)\cos(mz).$$

We recall that

$$\begin{aligned}\sin(lz)\sin(nz)\cos(mz) = \frac{1}{4}\big(&\cos\big((l-n-m)z\big) + \cos\big((l-n+m)z\big) \\ &- \cos\big((l+n-m)z\big) - \cos\big((l+n+m)z\big)\big).\end{aligned}$$

The only terms which do not vanish are the ones such that $l-n-m=0$, $l-n+m=0$, $l+n-m=0$. This gives

$$\begin{aligned}\left|\int_\Omega w\partial_z v\overline{\partial_z v}\right| &= \frac{\pi}{4}\sum_{l-n-m=0}\int_0^1 \frac{a_l'}{l}na_nma_m + \frac{\pi}{4}\sum_{l-n+m=0}\int_0^1 \frac{a_l'}{l}na_nma_m \\ &\quad - \frac{\pi}{4}\sum_{l+n-m=0}\int_0^1 \frac{a_l'}{l}na_nma_m \\ &= \frac{\pi}{4}\sum_{l-n-m=0}\int_0^1 \frac{|a_l'|}{l}na_nma_m.\end{aligned}$$

Let us now estimate the right-hand side. We have, since $m \le m^{1/2}(l^{1/2}+n^{1/2})$,

$$\begin{aligned}\sum_{l-n-m=0}\int_0^1 \frac{a_l'}{l}na_nma_m &\le \frac{\pi}{4}\sum_{l-n-m=0}\int_0^1 \frac{|a_l'|}{l^{1/2}}n|a_n|m^{1/2}|a_m| \\ &\quad + \frac{\pi}{4}\sum_{l-n-m=0}\int_0^1 \frac{|a_l'|}{l}n^{3/2}|a_n|m^{1/2}|a_m|.\end{aligned}$$

Moreover, since when $l-n-m=0$, $n\le l$,

$$\begin{aligned}&\frac{\pi}{4}\sum_{l-n-m=0}\int_0^1 \frac{|a_l'|}{l^{1/2}}n|a_n|m^{1/2}|a_m| \le \frac{\pi}{4}\sum_{l-n-m=0}\int_0^1 l^{1/2}|a_l'||a_n|m^{1/2}|a_m| \\ &\le \sum_{l,n,m}\int_0^1\int_0^\pi l^{1/2}|a_l'|\sin(lz)|a_n|\sin(nz)m^{1/2}|a_m|\cos(mz) \\ &\le \left\|\sum_l l^{1/2}|a_l'|\sin lz\right\|_2 \left\|\sum_n |a_n|\sin nz\right\|_{L_x^2L_z^\infty} \left\|\sum_m m^{1/2}|a_m|\cos mz\right\|_{L_x^\infty L_z^2} \\ &\le \|\partial_x\partial_z^{1/2}v\|_2 \left\|\sum_n |a_n|\sin nz\right\|_2^{1/2} \left\|\sum_n n|a_n|\sin nz\right\|_2^{1/2} \\ &\qquad \left\|\sum_m m^{1/2}|a_m|\cos mz\right\|_2^{1/2} \left\|\sum_m m^{1/2}|a_m'|\cos mz\right\|_2^{1/2} \\ &\le \|\partial_x\partial_z^{1/2}v\|_{L^2(\Omega)}^{3/2}\|v\|_{L^2(\Omega)}^{1/2}\|\partial_z v\|_{L^2(\Omega)}^{1/2}\|\partial_z^{1/2}v\|_{L^2(\Omega)}^{1/2}.\end{aligned}$$

Moreover, since if $l-n-m=0$, $n<l$, we have

$$\begin{aligned}&\frac{\pi}{4}\sum_{l-n-m=0}\int_0^1 \frac{|a_l'|}{l}n^{3/2}|a_n|m^{1/2}|a_m| \le \frac{\pi}{4}\sum_{l-n-m=0}\int_0^1 |a_l'|n^{1/2}a_nm^{1/2}a_m \\ &\le \sum_{l,n,m}\int_0^1\int_0^\pi |a_l'|\sin(lz)n^{1/2}|a_n|\sin(nz)m^{1/2}|a_m|\cos(mz) \\ &\le \sum_{l,n,m}\int_0^1\int_0^\pi |a_l'|\sin(lz)n^{1/2}|a_n|\sin(nz)m^{1/2}|a_m|\cos(mz) \\ &\le \|\partial_x v\|_2 \left\|\sum_n n^{1/2}|a_n|\sin nz\right\|_{L_x^2L_z^\infty} \left\|\sum_m m^{1/2}|a_m|\cos mz\right\|_{L_x^\infty L_z^2}\end{aligned}$$

$$\leq \|\partial_x v\|_2 \left\| \sum_n n^{1/2}|a_n| \sin nz \right\|_2^{1/2} \left\| \sum_n n^{3/2}|a_n| \sin nz \right\|_2^{1/2}$$
$$\left\| \sum_m m^{1/2}|a_m| \cos mz \right\|_2^{1/2} \left\| \sum_m m^{1/2}|a'_m| \cos mz \right\|_2^{1/2}$$
$$\leq \|\partial_x v\|_{L^2(\Omega)} \|\partial_z^{1/2} v\|_{L^2(\Omega)} \|\nabla \partial_z^{1/2} v\|_{(L^2(\Omega))^2},$$

and in conclusion,

$$\left| \int_\Omega w \partial_z v \overline{\partial_z v} \right| \leq \|\partial_x \partial_z^{1/2} v\|_{L^2(\Omega)}^{3/2} \|v\|_{L^2(\Omega)}^{1/2} \|\partial_z v\|_{L^2(\Omega)}^{1/2} \|\partial_z^{1/2} v\|_{L^2(\Omega)}^{1/2}$$
$$+ \|\partial_x v\|_{L^2(\Omega)} \|\partial_z^{1/2} v\|_{L^2(\Omega)} \|\nabla \partial_z^{1/2} v\|_{(L^2(\Omega))^2}.$$

Summary. (i) We have

$$\left\| \sum_n |a_n| \sin nz \right\|_2 = \left\| \sum_n |a_n| \cos nz \right\|_2.$$

(ii) If $f(0) = 0$, we have $|f(x)| = \int_0^x \partial_x |f|^2 \leq c \|\partial_x f\|_2^{1/2} \|f\|_2^{1/2}$, and then

$$\|f(x)\|_\infty \leq c \|\partial_x f\|_2^{1/2} \|f\|_2^{1/2}.$$

Let us now look at the Dirichlet boundary condition on the bottom.

Nonlinear problem with $\partial\ v|_{\ =}\ = 0$ and Dirichlet condition $v|_{\ =0} = 0$. Let us look at the following eigenvalues problem:

$$\begin{cases} \partial_z^2 v_k - \partial_x p_k + \lambda_k^2 v_k = 0, \qquad \partial_z p_k = 0, \\ \displaystyle\int_0^\pi v_k \, dz = 0, \\ \partial_z v_k|_{z=\pi} = 0, \quad v_k|_{z=0} = 0. \end{cases}$$

Writing the system on ω_k with the compatibility condition $(\partial_z \omega_k + \frac{1}{\pi}\omega_k)|_{z=0} = 0$, we find (v_k, λ_k) given by

$$v_k = \frac{1}{\sin(\lambda_k \pi)} \Big(\cos\big(\lambda_k (z - \pi)\big) - \cos(\lambda_k \pi) \Big),$$

with λ_k satisfying

$$\tan(\lambda_k \, \pi) = \lambda_k \pi.$$

Let us remark that $(v_k)_k$ is an orthogonal basis of L_z^2. Then we search the weak solution v of system (2) on the form $v = \sum_{n \geq 0} a_n v_n$ with $a_n = a_n(t, x)$, and we multiply the momentum equation by $\overline{\partial_z v} = \sum_{n=1}^\infty \lambda_n a_n(t, x) v_n(z)$. After some computations, we prove similar estimates to (40) and (41). We get an estimate similar to (42). This gives the results on the regularity and therefore the uniqueness. ∎

REFERENCES

[1] O. Besson and M. Laydi, *Some estimates for the anisotropic Navier–Stokes equations and for the hydrostatic approximation*, M^2AN Math. Model. Numer. Anal., 26 (1992), pp. 855–865.

[2] Y. Brenier, *Homogeneous hydrostatic flows with convex velocity profiles*, Nonlinearity, 12 (1999), pp. 495–512.

[3] D. Bresch, F. Guillén-González, N. Masmoudi, and M. A. Rodríguez-Bellido, *On the uniqueness for the two-dimensional primitive equations*, Differential Integral Equations, 16 (2003), pp. 77–94.

[4] G. P. Galdi, *An introduction to the Navier-Stokes initial boundary value problem*, in Fundamental Direction in Mathematical Fluid Mechanics, G. P. Galdi, J. H. Heywood, and R. Rannacher, eds., Adv. Math. Fluid Mech., Birkhäuser, Basel, 2000, pp. 1–70.

[5] E. Grenier, *On the derivation of homogeneous hydrostatic equations*, M^2AN Math. Model. Numer. Anal., 33 (1999), pp. 965–970.

[6] F. Guillén-González, N. Masmoudi, and M. A. Rodríguez-Bellido, *Anisotropic estimates and strong solutions of the primitive equations*, Differential Integral Equations, 14 (2001), pp. 1381–1408.

[7] F. Fujita and T. Kato, *On the non-stationary Navier-Stokes system*, Rend. Sem. Mat. Univ. Padova, 33 (1962), pp. 243–260.

[8] G. N. Ivey and J. C. Patterson, *A model of the vertical mixing in Lake Erie in summer*, Limnol. Oceanogr., 29 (1984), pp. 553–563.

[9] R. Lewandowski, *Analyse Mathématique et Océanographie*, Masson, Paris, 1997.

[10] J.-L. Lions, R. Temam, and S. Wang, *New formulation of the primitive equations of the atmosphere and applications*, Nonlinearity, 5 (1992), pp. 237–288.

[11] J.-L. Lions, R. Temam, and S. Wang, *On the equations of the large scale ocean*, Nonlinearity, 5 (1992), pp. 1007–1053.

[12] R. Temam and M. Ziane, *Navier-Stokes in thin spherical domains*, in Optimization Methods in Partial Differential Equations, Contemp. Math. 209, AMS, Providence, RI, 1997, pp. 281–314.

[13] R. Temam and M. Ziane, *Navier-Stokes equations in three-dimensional thin domains with various boundary conditions*, Adv. Differential Equations, 1 (1996), pp. 499–546.

[14] M. Ziane, *Regularity results for the stationary primitive equations of the atmosphere and the ocean*, Nonlinear Anal., 28 (1997), pp. 289–313.

SIAM J. MATH. ANAL.
Vol. 36, No. 3, pp. 815–855

PERIODIC SOLUTIONS OF THE KORTEWEG–DE VRIES EQUATION DRIVEN BY WHITE NOISE*

A. DE BOUARD†, A. DEBUSSCHE‡, AND Y. TSUTSUMI§

Abstract. We consider a Korteweg–de Vries equation perturbed by a noise term on a bounded interval with periodic boundary conditions. The noise is additive, white in time, and "almost white in space." We get a local existence and uniqueness result for the solutions of this equation. In order to obtain the result, we use the precise regularity of the Brownian motion in Besov spaces, and the method which was introduced by Bourgain, but based here on Besov spaces.

Key words. Korteweg–de Vries equation, stochastic partial differential equations, white noise, Besov spaces

AMS subject classifications. 35Q53, 60H15, 76B35

DOI. 10.1137/S0036141003425301

1. Introduction.

The Korteweg–de Vries (KdV) equation, which models the propagation of unidirectional weakly nonlinear waves in an infinite channel, is an ideal model, and it is natural to consider perturbations of this model. In this direction, stochastic perturbations of this equation were introduced in [5], [12], [19] to model the propagation of weakly nonlinear waves in a noisy plasma.

Here, we consider as in [2], [3] a KdV equation with a stochastic perturbation which is Gaussian and of white noise–type in time. Contrary to the previous works [2] and [3], we will set the equation on a bounded space interval with periodic boundary conditions. Although the derivation of the KdV equation is usually done with $x \in \mathbb{R}$, there is no reason to confine oneself to localized solutions. It is also well known that the KdV equation possesses spatially periodic traveling waves solutions. The study of the periodic boundary conditions case is also of importance when dealing with numerical computations, since these are necessarily performed on a bounded interval.

Our aim in the present paper is to study the Cauchy problem for a stochastic KdV equation with an additive noise as previously described, and which has spatial correlations "as rough" as our techniques allow, the aim being to stay as close as possible to the space-time white noise.

The equation is then written as

$$\partial_t u + \partial_x^3 u + u\partial_x u = \phi \frac{\partial^2 B}{\partial t \partial x}, \tag{1.1}$$

where u is a random process defined for $(t, x) \in \mathbb{R}^+ \times \mathbb{T}$, $\mathbb{T}$ being a one-dimensional torus, and ϕ is a bounded linear operator on $L^2(\mathbb{T})$ that will be described in more detail later. Also, B is a two parameter Brownian motion on $\mathbb{R}^+ \times \mathbb{T}$, that is, a zero

*Received by the editors March 30, 2003; accepted for publication (in revised form) December 12, 2003; published electronically October 14, 2004.
http://www.siam.org/journals/sima/36-3/42530.html
†CNRS et Université Paris-Sud, UMR 8628, Bât. 425, Université de Paris-Sud, 91405 Orsay Cedex, France (anne.debouard@math.u-psud.fr).
‡ENS de Cachan, Antenne de Bretagne, Campus de Ker Lann, Av. R. Schuman, 35170 Bruz, France (arnaud.debussche@bretagne.ens-cachan.fr).
§Mathematical Institute, Tohoku University, Sendai 980-8578, Japan (tsutsumi@math.tohoku.ac.jp).

mean Gaussian process whose correlation function is given by

$$\mathbb{E}(B(t,x)B(s,y)) = (t \wedge s)(x \wedge y)$$

for $t, s \geq 0$, $x, y \in \mathbb{T}$.

Note that in the case where ϕ is defined by a kernel $k(x,y)$, the correlation function of the noise is

$$\mathbb{E}\left(\phi\frac{\partial^2 B}{\partial t \partial x}(t,x)\phi\frac{\partial^2 B}{\partial t \partial x}(s,y)\right) = c(x,y)\delta_{t-s},$$

with δ the Dirac δ-function and

$$c(x,y) = \int_{\mathbb{T}} k(x,z)k(y,z)dz.$$

In this formalism, the case $\phi = Id$, i.e., $c(x,y) = \delta(x-y)$, corresponds to the space-time white noise. This is the case we would like to treat. However, our result needs a slightly more restrictive assumption, and we are only able to treat a noise which is "almost" delta correlated in space.

Except in [2] and [3], equations of the type (1.1) have essentially been studied by using inverse scattering theory (and only in cases where the noise is space independent) or by perturbation arguments near the integrable case (see [12], [16], [21], [22]).

A great deal of attention has been paid to the (deterministic) KdV equation on the real line (see [1], [4], [13], [18]) and improvements made on the regularity needed on the initial value to get local existence of solutions have occurred step by step. On the other hand, for the periodic case, up to the famous work of Bourgain on the KdV equation (see [4]), existence results in $H^s(\mathbb{T})$ were restricted to the case $s > 3/2$. Then, using functions spaces based on the linear group, Bourgain was able to prove global well-posedness in $L^2(\mathbb{T})$. Making use of the same spaces, and improving the nonlinear estimate, Kenig, Ponce, and Vega (see [15]) proved local well-posedness in $H^s(\mathbb{T})$ for $s > -1/2$ (see Colliander et al. [7] for $s = -1/2$). After that, using a splitting into high and low Fourier frequencies of the solution, together with almost conserved quantities and rescaling arguments, Colliander et al. [7] were able to prove global existence in $H^s(\mathbb{T})$ for $s \geq -1/2$.

Using Bourgain-type spaces, we were able in [3] to prove local existence of solutions for (1.1) in the real line case, when the noise is a "localized space-time white noise," that is, when its correlation function has the form

$$\mathbb{E}\left(\phi\frac{\partial^2 B}{\partial t \partial x}(t,x)\phi\frac{\partial^2 B}{\partial t \partial x}(s,y)\right) = k(x)k(y)\delta_{x-y}\delta_{t-s},$$

k being an L^2 function. It is indeed hopeless in the real line case to be able to get even local existence of solutions in $H^s(\mathbb{R})$, with a pure space-time white noise. The obstruction is not due to the lack of spatial regularity of the noise, but rather to its homogeneity (see [3]). In the periodic case, however, there is no such obstruction, and we are able to treat homogeneous noises, i.e., noises whose spatial correlation function depends only on $x - y$ (or such that ϕ is a convolution operator); also, thanks to the use of Bourgain's method adapted to Besov spaces, we are able to treat noises which have spatial correlations in H^s, $s > -1/2$. The main difficulty encountered in the application of Bourgain's method in our case is that it needs time regularity of order $1/2$. However, it is well known that this regularity does not hold for Brownian motions

unless Besov spaces are considered. This is why we use this method in the context of Besov spaces—see below for details. The problem of global existence of solutions for such noises in spaces with negative regularity is not considered here, but could probably be handled with the use of the method previously mentioned [7].

Before stating our result precisely, we introduce some notation and assumptions.

We consider $\tilde{W}(t) = \frac{\partial B}{\partial x}$ a cylindrical Wiener process on $L^2(\mathbb{T})$ which may be written as $\tilde{W}(t) = \sum_{j\in\mathbb{N}} \beta_j e_j$, where $(e_j)_{j\in\mathbb{N}}$ is a complete orthonormal system in $L^2(\mathbb{T})$, and $(\beta_j)_{j\in\mathbb{N}}$ is a sequence of mutually independent real-valued Brownian motions in a fixed probability space $(\Omega, \mathcal{F}, \mathbb{P})$ associated with a filtration $(\mathcal{F}_t)_{t\geq 0}$.

The process $W = \phi\tilde{W}$ is then a $\phi\phi^*$-Wiener process (recall that ϕ is a linear bounded operator in $L^2(\mathbb{T})$), that is, $(W(t))_{t\geq 0}$ is a Gaussian process with law $(\mathcal{N}(0, t\phi\phi^*)_{t\geq 0})$.

We then consider (1.1) in its Itô form,

$$du + (\partial_x^3 u + u\partial_x u)dt = dW, \quad x \in \mathbb{T},\ t \geq 0, \tag{1.2}$$

supplemented with the initial condition

$$u(0, x) = u_0(x), \quad x \in \mathbb{T}. \tag{1.3}$$

Consider the Fourier transform

$$\hat{f}(n) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{T}} e^{inx} f(x)dx$$

for functions f defined on $\mathbb{T}$, and for $s \in \mathbb{R}$, let $H^s(\mathbb{T})$ be the Sobolev space of functions f such that the norm

$$|f|_{H^s(\mathbb{T})} := \left(\sum_{n\in\mathbb{Z}} (1 + n^2)^s |\hat{f}(n)|^2 \right)^{1/2}$$

is finite. We also define, for $s \in \mathbb{R}$, the Besov space $B^s_{2,1}(\mathbb{T})$ as the space of functions f defined on $\mathbb{T}$ for which the norm

$$|f|_{B^s_{2,1}(\mathbb{T})} = |\hat{f}(0)| + \sum_{n\in\mathbb{N}} 2^{sn} \left(\sum_{2^{n-1}\leq |n'|\leq 2^{n+1}} |\hat{f}(n')|^2 \right)^{1/2}$$

is finite.

Let $U(t) = e^{-t\partial_x^3}$ be the group associated with the linear equation on $L^2(\mathbb{T})$, that is, $v(t) = U(t)u_0$ satisfies

$$\begin{cases} \partial_t v + \partial_x^3 v = 0, \\ v(0, x) = u_0(x), \quad x \in \mathbb{T}. \end{cases}$$

Then the solution of

$$\begin{cases} dw + \partial_x^3 w\, dt = dW, \\ w(0, x) = 0, \quad x \in \mathbb{T}, \end{cases}$$

is given by the stochastic convolution

$$w(t) = \int_0^t U(t - s)dW(s). \tag{1.4}$$

Note that $U(t)$ is a unitary group on $H^s(\mathbb{T})$ for any $s \in \mathbb{R}$, so that $w(t)$ lies in $H^s(\mathbb{T})$ almost surely if and only if $\phi\phi^*$ has finite trace from $L^2(\mathbb{T})$ into $H^s(\mathbb{T})$. This clearly holds in the case where ϕ is the identical operator on $L^2(\mathbb{T})$ if and only if $s < -1/2$.

The difficulty in the use of Bourgain's spaces here is the smoothness in time. Indeed, let $Y^{s,b}$ be the space of functions f such that $U(-t)f(t,\cdot) \in H^{s,b}$, where $H^{s,b}$ is a space-time Sobolev space, s being the regularity in space, and b the regularity in time (see [15] for a precise definition of $Y^{s,b}$). Then, as was proved in [15], the only possible value of b for which a bilinear estimate holds, which allows us to handle the nonlinear term $\partial_x(u^2)$ in the KdV equation using a straightforward iteration scheme, in the periodic case, is $b = 1/2$. Writing then the expression of $w(t)$ defined by (1.4) as

$$w(t) = \sum_{j\in\mathbb{N}} \int_0^t U(t-s)(\phi e_j) d\beta_j(s),$$

one can compute the spatial Fourier transform of $h(t) = U(-t)w(t)$:

$$\hat{h}(t,n) = \sum_{j\in\mathbb{N}} \int_0^t e^{isn^3} \widehat{\phi e_j}(n) d\beta_j(s).$$

But there is no hope that this term lives in $H^{1/2}[0,T]$ in the time variable, because the Brownian motions β_j do not. Indeed,

$$\begin{aligned}
&\mathbb{E}\left(|\hat{h}(t,n)|^2_{H^{1/2}_t}\right) \\
&\quad = \sum_{j\in\mathbb{N}} |\widehat{\phi e_j}(n)|^2 \mathbb{E}\left|\int_0^t e^{isn^3} d\beta_j(s)\right|^2_{H^{1/2}_t} \\
&\quad = \sum_{j\in\mathbb{N}} |\widehat{\phi e_j}(n)|^2 \left\{ \mathbb{E}\int_0^T \left|\int_0^t e^{isn^3} d\beta_j(s)\right|^2 dt \right. \\
&\qquad\qquad \left. + \mathbb{E}\int\int_{(0,T)^2} \frac{|\int_0^{t_1} e^{isn^3} d\beta_j(s) - \int_0^{t_2} e^{isn^3} d\beta_j(s)|^2}{|t_1 - t_2|^2} dt_1 dt_2 \right\}.
\end{aligned}$$

The first term in the right-hand side above is obviously equal to $\frac{T^2}{2}\sum_{j\in\mathbb{N}} |\widehat{\phi e_j}(n)|^2$, while the contribution of each j to the second term in the right-hand side above is infinite, due to the fact that

$$\mathbb{E}\left|\int_{t_1}^{t_2} e^{isn^3} d\beta_j(s)\right|^2 = |t_2 - t_1|.$$

However, $H^{1/2}$ is a limiting case concerning the regularity of the Brownian motion, as far as Sobolev spaces are concerned. It is then natural to try to replace Sobolev spaces here by other spaces which describe more precisely the regularity in time of the Brownian motions. This is exactly what we will do here, using Besov spaces instead of Sobolev spaces in time. Indeed, it is known (see [6], [17]) that the Brownian motion lies almost surely in $B^{1/2}_{p,q}([0,T])$ if and only if $1 \le p < +\infty$ and $q = +\infty$. Trying to derive some bilinear estimate which would allow us to handle in the same time both $w(t)$ defined by (1.4) and the nonlinear term,

$$\int_0^t U(t-s)(\partial_x(u^2)(s)) ds,$$

we were led to consider also Besov spaces in the space variable.

We now turn to give precise definitions of these spaces. We denote by $\langle \cdot, \cdot \rangle$ the L^2 space-time duality product, that is,

$$\begin{aligned}\langle f, g \rangle &= \int_{\mathbb{T}} \int_{\mathbb{R}} f(t,x)\overline{g(t,x)} dt\, dx \\ &= \sum_{n \in \mathbb{Z}} \int_{\mathbb{R}} \hat{f}(\tau, n)\overline{\hat{g}(\tau, n)} d\tau\end{aligned}$$

by the Plancherel formula; here, and in all that follows, we denote by $\hat{f}$ (resp., $\hat{g}$) the Fourier transform of f (resp., g) with respect to both variables. We also use the notation $\langle \tau \rangle = (1 + |\tau|^2)^{1/2}$ for $\tau \in \mathbb{R}$. The spaces that we will use are defined as follows. Consider first functions f defined on $\mathbb{R} \times \mathbb{T}$ such that $f(\cdot, x) \in \mathcal{S}'(\mathbb{R})$ for any $x \in \mathbb{T}$, and such that $\hat{f}(\tau, 0) = 0$ for any $\tau \in \mathbb{R}$.

We denote by $X^{s,b}_{1,1}$ the space of such functions f for which in addition the norm

$$\begin{aligned}|f|_{X^{s,b}_{1,1}} = &\sum_{n=0}^{\infty} 2^{sn} \sum_{k=0}^{\infty} \left(\sum_{2^{n-1} \le |n'| \le 2^{n+1}} \int_{2^{k-1}}^{2^{k+1}} |\langle \tau - n'^3 \rangle^b \hat{f}(\tau, n')|^2 d\tau \right)^{1/2} \\ &+ \sum_{n=0}^{\infty} 2^{sn} \left(\sum_{2^{n-1} \le |n'| \le 2^{n+1}} \int_{0}^{1} |\langle \tau - n'^3 \rangle^b \hat{f}(\tau, n')|^2 d\tau \right)^{1/2}\end{aligned}$$

is finite. In the same way, we will denote by $X^{s,b}_{1,\infty}$ the space of such functions f for which in addition the norm

$$\begin{aligned}|f|_{X^{s,b}_{1,\infty}} = &\sum_{n=0}^{\infty} 2^{sn} \sup_{k \in \mathbb{N}} \left(\sum_{2^{n-1} \le |n'| \le 2^{n+1}} \int_{2^{k-1}}^{2^{k+1}} |\langle \tau - n'^3 \rangle^b \hat{f}(\tau, n')|^2 d\tau \right)^{1/2} \\ &+ \sum_{n=0}^{\infty} 2^{sn} \left(\sum_{2^{n-1} \le |n'| \le 2^{n+1}} \int_{0}^{1} |\langle \tau - n'^3 \rangle^b \hat{f}(\tau, n')|^2 d\tau \right)^{1/2}\end{aligned}$$

is finite.

The basic space in which we will solve the Cauchy problem for the stochastic KdV equation is $X^{s,b}_{1,1}$. However, we will make use, at intermediate steps, of other spaces of the same type: $\widetilde{X}^{s,b}_{1,1}$ (resp., $\widetilde{X}^{s,b}_{1,\infty}$) is the space of functions f such that $f(t, \cdot) = U(t)g(t, \cdot)$ with g in the "space-time Besov space" $(B^{s,b}_{2,1})_{x,t}$ (resp., $(B^{s}_{2,1})_x(B^{b}_{2,\infty})_t$), where $(B^{s,b}_{2,1})_{x,t}$ is defined by the norm

$$\begin{aligned}|f|_{(B^{s,b}_{2,1})_{x,t}} = &\sum_{n=0}^{\infty} \sum_{k=0}^{\infty} 2^{sn+kb} \left(\sum_{2^{n-1} \le |n'| \le 2^{n+1}} \int_{2^{k-1}}^{2^{k+1}} |\hat{f}(\tau, n')|^2 d\tau \right)^{1/2} \\ &+ \sum_{n=0}^{\infty} 2^{sn} \left(\sum_{2^{n-1} \le |n'| \le 2^{n+1}} \int_{0}^{1} |\hat{f}(\tau, n')|^2 d\tau \right)^{1/2}\end{aligned}$$

and $(B^s_{2,1})_x(B^b_{2,\infty})_t$ is defined by the norm

$$|f|_{(B^s_{2,1})_x(B^b_{2,\infty})_t} = \sum_{n=0}^{\infty} \sup_{k\in\mathbb{N}} 2^{sn+kb} \left(\sum_{2^{n-1}\le|n'|\le 2^{n+1}} \int_{2^{k-1}}^{2^{k+1}} |\hat f(\tau,n')|^2 d\tau \right)^{1/2} + \sum_{n=0}^{\infty} 2^{sn} \left(\sum_{2^{n-1}\le|n'|\le 2^{n+1}} \int_0^1 |\hat f(\tau,n')|^2 d\tau \right)^{1/2}.$$

Remark 1.1. Note that the spaces $X^{s,b}_{1,1}$ and $\widetilde X^{s,b}_{1,1}$ are different and there is no inclusion relation between them: an alternative definition of the norm in $\widetilde X^{s,b}_{1,1}$ is

$$|f|_{\widetilde X^{s,b}_{1,1}} = \sum_{n=0}^{\infty} 2^{sn} \sum_{k=0}^{\infty} \left(\sum_{2^{n-1}\le|n'|\le 2^{n+1}} \int_{2^{k-1}\le|\tau-n'^3|\le 2^{k+1}} |\langle \tau-n'^3\rangle^b \hat f(\tau,n')|^2 d\tau \right)^{1/2} + \sum_{n=0}^{\infty} 2^{sn} \left(\sum_{2^{n-1}\le|n'|\le 2^{n+1}} \int_{|\tau-n'^3|\le 1} |\langle \tau-n'^3\rangle^b \hat f(\tau,n')|^2 d\tau \right)^{1/2};$$

here, the dyadic decomposition is made on $|\tau - n'^3|$ and not on $|\tau|$. However, embeddings do hold between these spaces with some small loss of space regularity, as is stated in Lemma 1.6, at the end of this section.

Since all those definitions have to be used only locally in time, we will actually consider, for $T \ge 0$ fixed, the spaces $X^{s,b,T}_{1,1}$ and $X^{s,b,T}_{1,\infty}$ of restrictions on $[0,T]$ of functions of $X^{s,b}_{1,1}$ (resp., $X^{s,b}_{1,\infty}$). They are endowed with the natural norm

$$|f|_{X^{s,b,T}_{1,1}} = \inf\left\{ |\tilde f|_{X^{s,b}_{1,1}},\ \tilde f \in X^{s,b}_{1,1} \text{ and } f = \tilde f_{|_{[0,T]}} \right\},$$

and the equivalent for $X^{s,b,T}_{1,\infty}$.

To handle the integral estimate in Duhamel's formula, we will need to make use, as is classical, of another space which is defined as the space of zero (spatial) mean functions with finite corresponding norm, where

$$|f|_{Y_s} = \sum_{n=0}^{\infty} 2^{sn} \left(\sum_{2^{n-1}\le|n'|\le 2^{n+1}} \left(\int_{\mathbb{R}} \frac{|\hat f(\tau,n')|}{\langle \tau-n'^3\rangle} d\tau \right)^2 \right)^{1/2}.$$

A local space $Y_{s,T}$ is also defined, in the same way as for $X^{s,1/2,T}_{1,1}$.

Throughout the paper, we will use the notation $|n'| \sim 2^n$ for $2^{n-1} \le |n'| \le 2^{n+1}$, and $|\tau| \sim 2^k$ for $2^{k-1} \le |\tau| \le 2^{k+1}$ if $k \ge 1$ and $|\tau| \le 2$ if $k = 0$.

As previously mentioned, we will be led to assume[1] that the operator ϕ is a Hilbert–Schmidt operator (or equivalently that $\phi\phi^*$ has finite trace) from $L^2(\mathbb{T})$ into $H^s(\mathbb{T})$ for some negative s with $s > -1/2$. We will denote by $L^{0,s}_2$ the space of such operators, which is endowed by its natural norm,

$$\|\phi\|_{L^{0,s}_2} = \left(\sum_{i\in\mathbb{N}} |\phi e_i|^2_{H^s} \right)^{1/2},$$

[1]Note that this assumption excludes the identical operator on $L^2(\mathbb{T})$.

where $(e_i)_{i\in\mathbb{N}}$ is any complete orthonormal system in $L^2(\mathbb{T})$. For convenience, in all that follows, we take as $(e_i)_{i\in\mathbb{N}}$ the usual complete orthonormal system of $L^2(\mathbb{T})$ given by

$$e_{2k}(x) = \frac{1}{\sqrt{\pi}}\cos kx, \quad k \geq 1, \quad e_0(x) = \frac{1}{\sqrt{2\pi}},$$
$$e_{2k+1}(x) = \frac{1}{\sqrt{\pi}}\sin kx.$$

We consider the mild form of (1.2), (1.3), that is,

$$u(t) = U(t)u_0 - \frac{1}{2}\int_0^t U(t-s)\partial_x(u^2(s))ds + \int_0^t U(t-s)dW(s). \tag{1.5}$$

Our main result, which concerns local existence in a situation where W is arbitrarily close to a cylindrical Wiener process, is the following.

THEOREM 1.2. *Assume that* $\operatorname{Im}\phi \subset \operatorname{span}\{e_j,\ j \geq 1\}$ *and that* $\phi \in L_2^{0,s}$ *for some* s *with* $s > -1/2$. *Let* u_0 *be* $\mathcal{F}_0$*-measurable, with* u_0 *in the Besov space* $B^\sigma_{2,1}(\mathbb{T})$ *almost surely for some* σ *with* $-1/2 \leq \sigma < s$; *then there is a stopping time* $T_\omega > 0$ *and a unique process* u *solution of the forced KdV equation* (1.5) *which satisfies*

$$u \in C([0,T_\omega]; B^\sigma_{2,1}(\mathbb{T})) \cap X^{\sigma,1/2,T_\omega}_{1,1} \text{ almost surely.}$$

Remark 1.3. The assumption $\operatorname{Im}\phi \subset \operatorname{span}\{e_j,\ j \geq 1\}$ says that the spatial mean of the noise is zero at any time. This assumption is necessary to perform the fixed point procedure, because we work in a space of functions with zero spatial mean. We will actually remove this assumption at the end of the paper (see Proposition 4.3) by changing the unknown function u and the noise. At that place, we will have to deal with a non-Gaussian noise.

Remark 1.4. One can show by using classical arguments and looking more carefully into the proof of Proposition 3.1 (see section 3) that the regularity is preserved in Theorem 1.2, i.e., if $\phi \in L_2^{0,s}$ and $u_0 \in B^\sigma_{2,1}(\mathbb{T})$ with $-1/2 \leq \sigma' \leq \sigma < s$, then the existence times of the solution in $B^{\sigma'}_{2,1}(\mathbb{T})$ and in $B^\sigma_{2,1}(\mathbb{T})$ are the same.

Naturally, when the noise is such that the Wiener process lies in $L^2(\mathbb{T})$, we get a global existence result thanks to the invariance of the L^2 norm for the deterministic equation and the embedding $L^2(\mathbb{T}) \subset B^\sigma_{2,1}(\mathbb{T})$ for any $\sigma < 0$.

THEOREM 1.5. *Assume that, in addition,* $\phi \in L_2^{0,0}$; *if* $u_0 \in L^2(\Omega; L^2(\mathbb{T}))$, *the solution given by Theorem* 1.2 *is globally defined in time and lies in* $L^2(\Omega; L^\infty(0,T; L^2(\mathbb{T})))$ *and in* $C(\mathbb{R}^+; B^\sigma_{2,1}(\mathbb{T}))$ *almost surely for any* $T > 0$ *and* $\sigma < 0$.

As was previously mentioned, Theorem 1.2 allows us to handle a situation arbitrarily close to the space-time white noise case, since this latter case corresponds to $\phi = \mathrm{id}$, which is a Hilbert–Schmidt operator from $L^2(\mathbb{T})$ into $H^s(\mathbb{T})$ for any $s < -1/2$. Theorem 1.2 will be proved by using a fixed point argument in the space $X^{\sigma,1/2,T}_{1,1}$ for T small enough. We need the assumption $s > -1/2$ because we will need that $s > \sigma \geq -1/2$. Indeed, to show that the fixed point works, we will first prove that the stochastic integral lies almost surely in $X^{\sigma,1/2}_{1,\infty}$. At that point, we have already lost some spatial regularity. We then prove a bilinear estimate, allowing us to handle such a term as $\partial_x(fg)$ with $f \in X^{\sigma,1/2}_{1,1}$ and $g \in X^{\sigma,1/2}_{1,\infty}$. To treat the term $\partial_x(g^2)$ in the same space, we again have to sacrifice an arbitrarily small amount of spatial regularity.

It is not difficult to see that when $\phi = \mathrm{id}$, the stochastic integral $w(t)$ given by (1.4) lies almost surely in $X^{-1/2,1/2}_{\infty,\infty}$, where this latest space is defined by changing the norm in the definition of $X^{-1/2,1/2}_{1,\infty}$ in an obvious way. Unfortunately, a bilinear estimate which would handle terms like $\partial_x(g^2)$ in $X^{-1/2,-1/2}_{\infty,\infty}$ with g in $X^{-1/2,1/2}_{\infty,\infty}$ seems to fail.

The paper is organized as follows. In section 2, we prove an estimate which shows that the stochastic integral lives in $X^{\sigma,1/2}_{1,\infty}$ almost surely when ϕ is in $L^{0,s}_2$ with $\sigma < s$ (we will actually prove that the stochastic integral lies in $\widetilde{X}^{\sigma,1/2}_{1,\infty}$, which is enough, thanks to Lemma 1.6 below). This result is based on the works of Cieselskii [6] and Roynette [17], but we will use a different characterization of Besov spaces than in [17].

In section 3, we prove some bilinear estimates which are needed in the proof of Theorem 1.2. The main one is an estimate of $\partial_x(fg)$ in $X^{\sigma,-1/2}_{1,1}$ when $f \in X^{\sigma,1/2}_{1,1}$ and $g \in X^{\sigma,1/2}_{1,\infty}$. Other easier bilinear estimates are proved in that section too.

Section 4 is devoted to the proofs of Theorems 1.2 and 1.5. Once we have the bilinear estimates in hand, together with the estimate on the stochastic integral, it mainly remains to prove that we gain one degree of regularity in time when passing from $\partial_x(fg)$ to $\int_0^t U(t-s)\partial_x(fg)(s)ds$. The proof of this fact has to be done because we do not stand in the usual context of Sobolev spaces, but we deal with Besov spaces. However, the proof closely follows that of the Sobolev case.

We end the present section by giving the lemma relating the spaces $X^{s,b}_{1,1}$ and $\widetilde{X}^{s,b}_{1,\infty}$.

LEMMA 1.6. *For any $s_1 > s_2 > s_3$,*

$$\widetilde{X}^{s_1,b}_{1,1} \subset X^{s_2,b}_{1,1} \subset \widetilde{X}^{s_3,b}_{1,1} \quad \textit{and} \quad \widetilde{X}^{s_1,b}_{1,\infty} \subset X^{s_2,b}_{1,\infty} \subset \widetilde{X}^{s_3,b}_{1,\infty}.$$

Proof. We show only that $\widetilde{X}^{s_1,b}_{1,\infty} \subset X^{s_2,b}_{1,\infty}$, and all the other embeddings are proved similarly. Let $f \in \widetilde{X}^{s_1,b}_{1,\infty}$ and let us decompose the norm of f in $X^{s_2,b}_{1,\infty}$ as

$$\begin{aligned}
|f|_{X^{s_2,b}_{1,\infty}} \leq{}& \sum_{n\in\mathbb{N}} 2^{s_2 n} \sup_{k<3n-4} \left(\sum_{|n'|\sim 2^n} \int_{|\tau|\sim 2^k} \langle \tau - n'^3\rangle^{2b} |\hat{f}(\tau,n')|^2 \, d\tau \right)^{1/2} \\
&+ \sum_{n\in\mathbb{N}} 2^{s_2 n} \sup_{3n-4\leq k\leq 3n+4} \left(\sum_{|n'|\sim 2^n} \int_{|\tau|\sim 2^k} \langle \tau - n'^3\rangle^{2b} |\hat{f}(\tau,n')|^2 \, d\tau \right)^{1/2} \\
&+ \sum_{n\in\mathbb{N}} 2^{s_2 n} \sup_{k>3n+4} \left(\sum_{|n'|\sim 2^n} \int_{|\tau|\sim 2^k} \langle \tau - n'^3\rangle^{2b} |\hat{f}(\tau,n')|^2 \, d\tau \right)^{1/2} \\
\leq{}& I + II + III.
\end{aligned}$$

If $k > 3n+4$, $|n'| \sim 2^n$ and $|\tau| \sim 2^k$, then $\frac{1}{8}|\tau| \leq |\tau - n'^3| \leq \frac{3}{2}|\tau|$; hence we easily have

$$\begin{aligned}
III \leq{}& C \sum_{n\in\mathbb{N}} 2^{s_2 n} \sup_{k\in\mathbb{N}} \left(\sum_{|n'|\sim 2^n} \int_{|\tau - n'^3|\sim 2^k} \langle \tau - n'^3\rangle^{2b} |\hat{f}(\tau,n')|^2 \, d\tau \right)^{1/2} \\
\leq{}& C|f|_{\widetilde{X}^{s_2,b}_{1,\infty}}.
\end{aligned}$$

On the other hand, if $k < 3n-4$, $|n'| \sim 2^n$ and $|\tau| \sim 2^k$, then $2^{3n-4} \leq |\tau - n'^3| \leq 2^{3n+4}$; hence

$$I \leq \sum_{n\in\mathbb{N}} 2^{s_2 n} \left(\sum_{|n'|\sim 2^n} \int_{2^{3n-4}\leq|\tau-n'^3|\leq 2^{3n+4}} \langle \tau - n'^3\rangle^{2b} |\hat{f}(\tau,n')|^2 \, d\tau \right)^{1/2}$$
$$\leq 8|f|_{\widetilde{X}^{s_2,b}_{1,\infty}}.$$

Finally, if $3n-4 \leq k \leq 3n+4$, $|n'| \sim 2^n$, and $|\tau| \sim 2^k$, then $0 \leq |\tau - n'^3| \leq 2^{3n+6}$; hence

$$II \leq \sum_{n\in\mathbb{N}} 2^{s_2 n} \left(\sum_{|n'|\sim 2^n} \sum_{k=0}^{3n+5} \int_{|\tau-n'^3|\sim 2^k} \langle \tau - n'^3\rangle^{2b} |\hat{f}(\tau,n')|^2 \, d\tau \right)^{1/2}$$
$$\leq \sum_{n\in\mathbb{N}} 2^{s_2 n}(3n+5) \sup_{k\in\mathbb{N}} \left(\sum_{|n'|\sim 2^n} \int_{|\tau-n'^3|\sim 2^k} \langle \tau - n'^3\rangle^{2b} |\hat{f}(\tau,n')|^2 \, d\tau \right)^{1/2}$$
$$\leq C \sum_{n\in\mathbb{N}} 2^{s_1 n} \sup_{k\in\mathbb{N}} \left(\sum_{|n'|\sim 2^n} \int_{|\tau-n'^3|\sim 2^k} \langle \tau - n'^3\rangle^{2b} |\hat{f}(\tau,n')|^2 \, d\tau \right)^{1/2}$$

since $s_2 < s_1$. The result follows. □

2. Estimate on the stochastic integral. In this section, we prove an estimate on the stochastic integral—that is, the last term in (1.5)—which will enable us to use a fixed point procedure to solve (1.5) in an appropriate space of functions of the space and time variables. This latest space will actually be of the form $X^{\sigma,1/2}_{1,1}$ for some well chosen σ.

Although, for the sake of clarity, we did not assume that the covariance operator $\phi\phi^*$ of the noise could be random or could depend on the time variable t in Theorem 1.2, we will state here a proposition where ϕ is allowed to depend on both t and ω, but under the condition that the $L_2^{0,\sigma'}$ norm of $\phi(\cdot)$ is bounded in both t and ω. This will indeed be useful in order to prove that our result generalizes to the case where the noise does not have a zero spatial mean value (see Proposition 4.3).

We need to use a cut-off function in the time variable: we consider a function $\theta : \mathbb{R} \longrightarrow \mathbb{R}^+$ such that $\theta(t) \equiv 0$ for $t \leq -1$, and $t \geq 2$, $\theta(t) \equiv 1$ for $t \in [0,1]$, and $\theta \in C_0^\infty(\mathbb{R})$.

Also, to state precisely our estimate on the stochastic integral, we define, for $n \in \mathbb{N}$, the operator Δ_n acting on $L^2(\mathbb{T})$ by

$$\widehat{\Delta_n u}(n') = \mathbb{1}_{\{2^{n-1}\leq|n'|\leq 2^{n+1}\}} \hat{u}(n')$$

for $u \in L^2(\mathbb{T})$ and for any $n' \in \mathbb{Z}$.

We now state our proposition.

PROPOSITION 2.1. *Let $s' \in \mathbb{R}$, and assume that ϕ is predictable and lies in $L^\infty([0,T]\times\Omega; L_2^{0,s'})$ for some T with $0 < T \leq 1$; let θ and Δ_n be as above; then the stochastic integral $w(t)$ defined by (1.4) satisfies, for any $\sigma' < \sigma < s'$, $\theta w \in$*

$L^1(\Omega; X_{1,\infty}^{\sigma',1/2,T})$ *and*

$$\mathbb{E}\left(|\theta w|_{X_{1,\infty}^{\sigma',1/2,T}}\right) \le C(\theta)\sum_{n\in\mathbb{N}}\|\Delta_n\phi\|_{L^\infty([0,T]\times\Omega;L_2^{0,\sigma})} \le C(\theta,\sigma,s')\|\phi\|_{L^\infty([0,T]\times\Omega;L_2^{0,s'})},$$

where $C(\theta)$ is a constant depending only on the function θ.

Proof. We first prove that $\theta w \in L^1(\Omega;\widetilde{X}_{1,\infty}^{\sigma,1/2,T})$ and that

$$\mathbb{E}\left(|\theta w|_{\widetilde{X}_{1,\infty}^{\sigma,1/2,T}}\right) \le C(\theta)\sum_{n\in\mathbb{N}}\|\Delta_n\phi\|_{L^\infty([0,T]\times\Omega;L_2^{0,\sigma})}$$

and then make use of Lemma 1.6.

Let $g(t,\cdot)=\theta(t)\int_0^t U(-s)dW(s)$ so that $\theta(t)w(t)=U(t)g(t,\cdot)$; we also set, for $s\in\mathbb{R}$, $n\in\mathbb{Z}$, and $\ell\in\mathbb{N}$,

$$\varphi_{n,\ell}(s)=\begin{cases}0 & \text{if } s<0 \quad \text{or} \quad s\ge T,\\ \widehat{\phi(s)e_\ell}(n) & \text{if } s\in[0,T],\end{cases}$$

and we assume that each β_ℓ has been extended to a Brownian motion on $\mathbb{R}$, in such a way that the family $(\beta_\ell)_{\ell\in\mathbb{N}}$ is still an independent family. We then have, for any $t\in[0,T]$ and $n\in\mathbb{Z}$,

$$\mathcal{F}_n g(t)(n)=\sum_{\ell\in\mathbb{N}}\theta(t)I_{n,\ell}(t),$$

with $I_{n,\ell}(t)=\int_{-\infty}^t\theta(s)e^{ins^3}\varphi_{n,\ell}(s)\,d\beta_\ell(s)$, $\mathcal{F}_n$ being the Fourier transform in space.

In view of the equivalent definition of the space $\widetilde{X}_{1,\infty}^{\sigma,1/2}$, we have to show that

$$\begin{aligned}&\mathbb{E}\left(\sum_{n=0}^\infty \sup_{k\in\mathbb{N}} 2^{\sigma n+k/2}\left(\sum_{|n'|\sim 2^n}\int_{|\tau|\sim 2^k}|\hat g(\tau,n')|^2\,d\tau\right)^{1/2}\right)\\ &\quad+\mathbb{E}\left(\sum_{n=0}^\infty 2^{\sigma n}\left(\sum_{|n'|\sim 2^n}\int_{|\tau|\le 1}|\hat g(\tau,n')|^2\,d\tau\right)^{1/2}\right)\\ &\quad\le C(\theta)\sum_{n=0}^\infty\|\Delta_n\phi\|_{L^\infty([0,T]\times\Omega;L_2^{0,\sigma})}.\end{aligned}\tag{2.1}$$

We first estimate the second term in (2.1).

$$\begin{aligned}&\mathbb{E}\left(\sum_{n=0}^\infty 2^{\sigma n}\left(\sum_{|n'|\sim 2^n}\int_{|\tau|\le 1}\left|\sum_{\ell\in\mathbb{N}}\widehat{\theta I_{n',\ell}}(\tau)\right|^2 d\tau\right)^{1/2}\right)\\ &\le\sum_{n=0}^\infty 2^{\sigma n}\left(\sum_{|n'|\sim 2^n}\mathbb{E}\int_{|\tau|\le 1}\left|\sum_{\ell\in\mathbb{N}}\int_{\mathbb{R}}\theta(t)\int_{-\infty}^t\theta(s)e^{isn'^3}\varphi_{n',\ell}(s)\,d\beta_\ell(s)e^{-i\tau t}dt\right|^2 d\tau\right)^{1/2}\\ &\le\sum_{n=0}^\infty 2^{\sigma n}\left(\sum_{|n'|\sim 2^n}\int_{|\tau|\le 1}\mathbb{E}\left|\sum_{\ell\in\mathbb{N}}\int_{\mathbb{R}}\theta(s)\varphi_{n',\ell}(s)e^{isn'^3}\int_s^{+\infty}\theta(t)e^{-i\tau t}\,dt\,d\beta_\ell(s)\right|^2 d\tau\right)^{1/2},\end{aligned}$$

and using the independence of the $(\beta_\ell)_{\ell\in\mathbb{N}}$, the above term is bounded by

$$\begin{aligned}
&\sum_{n=0}^{\infty} 2^{\sigma n}\left(\sum_{|n'|\sim 2^n}\int_{|\tau|\le 1}\sum_{\ell\in\mathbb{N}}\int_{\mathbb{R}}\mathbb{E}(\theta^2(s)|\varphi_{n',\ell}(s)|^2)\left|\int_s^{+\infty}\theta(t)e^{-i\tau t}\,dt\right|^2 ds\,d\tau\right)^{1/2}\\
&\quad\le 2|\theta|^2_{L^1(\mathbb{R})}|\theta|^2_{L^2(\mathbb{R})}\sum_{n=0}^{+\infty}2^{\sigma n}\sup_{s\in\mathbb{R}}\left|\mathbb{E}\left(\sum_{\ell\in\mathbb{N}}\sum_{|n'|\sim 2^n}|\varphi_{n',\ell}(s)|^2\right)\right|^{1/2}\\
&\quad\le C(\theta)\sum_{n=0}^{+\infty}\sup_{s\in\mathbb{R}}\left|\mathbb{E}\left(\|\Delta_n\phi(s)\|^2_{L_2^{0,\sigma}}\right)\right|^{1/2}\\
&\quad\le C(\theta)\sum_{n=0}^{+\infty}\|\Delta_n\phi(\cdot)\|_{L^\infty([0,T]\times\Omega;L_2^{0,\sigma})},
\end{aligned}$$

and this proves the estimate on the second term in (2.1).

In what follows, we assume that $|\tau|\ge 1/2$; by the stochastic Fubini theorem and using an integration by parts, we easily get, for $n\in\mathbb{Z}$, $\ell\in\mathbb{N}$, and $|\tau|\ge 1/2$,

$$\widehat{\theta I_{n,\ell}}(\tau)=A_{n,\ell}(\tau)+B_{n,\ell}(\tau),$$

with

$$A_{n,\ell}(\tau)=\int_{\mathbb{R}}\theta^2(s)e^{isn^3}\varphi_{n,\ell}(s)\frac{e^{-i\tau s}}{i\tau}\,d\beta_\ell(s)$$

and

$$B_{n,\ell}(\tau)=\int_{\mathbb{R}}\theta(s)e^{isn^3}\varphi_{n,\ell}(s)\int_s^{+\infty}\theta'(t)\frac{e^{-i\tau t}}{i\tau}\,dt\,d\beta_\ell(s).$$

Hence, two terms will be involved in the estimate of the second term in (2.1), which are

$$I=\mathbb{E}\left(\sum_{n=0}^{\infty}\sup_{k\in\mathbb{N}}2^{\sigma n+k/2}\left(\sum_{|n'|\sim 2^n}\int_{|\tau|\sim 2^k}\left|\sum_{\ell=0}^{\infty}A_{n',\ell}(\tau)\right|^2 d\tau\right)^{1/2}\right)$$

and

$$II=\mathbb{E}\left(\sum_{n=0}^{\infty}\sup_{k\in\mathbb{N}}2^{\sigma n+k/2}\left(\sum_{|n'|\sim 2^n}\int_{|\tau|\sim 2^k}\left|\sum_{\ell=0}^{\infty}B_{n',\ell}(\tau)\right|^2 d\tau\right)^{1/2}\right).$$

We may assume that $\sigma=0$, replacing $\varphi_{n',\ell}$ by $2^{\sigma n}\varphi_{n',\ell}$ in the estimate we want to prove. We first estimate the second term above. With this aim in view, we first write,

for $k \geq 0$ and $n \geq 0$,

$$\begin{aligned}
&\mathbb{E}\left(2^k \sum_{|n'|\sim 2^n} \int_{|\tau|\sim 2^k} \left|\sum_{\ell=0}^{\infty} B_{n',\ell}(\tau)\right|^2 d\tau\right) \\
&\quad = 2^k \int_{|\tau|\sim 2^k} \sum_{|n'|\sim 2^n} \mathbb{E}\left|\sum_{\ell=0}^{\infty} \int_{\mathbb{R}} \theta(s) e^{isn'^3} \varphi_{n',\ell}(s) \int_s^{+\infty} \theta'(t) \frac{e^{-i\tau t}}{i\tau} dt \, d\beta_\ell(s)\right|^2 d\tau \\
&\quad = 2^k \int_{|\tau|\sim 2^k} \sum_{\ell=0}^{\infty} \sum_{|n'|\sim 2^n} \int_{\mathbb{R}} |\theta(s)|^2 \mathbb{E}(|\varphi_{n',\ell}(s)|^2) \left|\int_s^{+\infty} \theta'(t) \frac{e^{-i\tau t}}{i\tau} dt\right|^2 ds \, d\tau,
\end{aligned}$$

where we have used again the independence of the family $(\beta_\ell)_{\ell \geq 0}$. Now, for $|\tau|$ in $[2^{k-1}, 2^{k+1}]$, we have

$$\left|\int_s^{+\infty} \theta'(t) \frac{e^{-i\tau t}}{i\tau} dt\right| \leq \left|\frac{\theta'(s)}{\tau^2}\right| + \left|\int_s^{+\infty} \theta''(t) \frac{e^{-i\tau t}}{\tau^2} dt\right| \leq \frac{C(\theta)}{2^{2k}};$$

hence we get, for k, $n \geq 0$,

$$\begin{aligned}
&\mathbb{E}\left(2^k \sum_{|n'|\sim 2^n} \int_{|\tau|\sim 2^k} \left|\sum_{\ell=0}^{\infty} B_{n',\ell}(\tau)\right|^2 d\tau\right) \\
&\quad \leq C(\theta) 2^{-3k} \int_{|\tau|\sim 2^k} \left|\mathbb{E}\left(\|\Delta_n \phi(s)\|^2_{L_2^{0,0}}\right)\right|_{L_s^\infty} d\tau \\
&\quad \leq C(\theta) 2^{-2k} \|\Delta_n \phi(\cdot)\|^2_{L^\infty([0,T]\times\Omega; L_2^{0,0})},
\end{aligned}$$

and using the Cauchy–Schwarz inequality, we get

$$\begin{aligned}
II &= \sum_{n=0}^{+\infty} \sup_{k\in\mathbb{N}} \mathbb{E}\left(2^{k/2} \left(\sum_{|n'|\sim 2^n} \int_{|\tau|\sim 2^k} \left|\sum_{\ell=0}^{\infty} B_{n',\ell}(\tau)\right|^2 d\tau\right)^{1/2}\right) \\
&\leq \sum_{n=0}^{\infty} \sum_{k\in\mathbb{N}} \left(\mathbb{E}\left(2^k \sum_{|n'|\sim 2^n} \int_{|\tau|\sim 2^k} \left|\sum_{\ell=0}^{\infty} B_{n',\ell}(\tau)\right|^2 d\tau\right)\right)^{1/2} \qquad (2.2) \\
&\leq C(\theta) \sum_{n=0}^{\infty} \sum_{k\in\mathbb{N}} 2^{-k} \|\Delta_n \phi(\cdot)\|_{L^\infty([0,T]\times\Omega; L_2^{0,0})} \\
&\leq C(\theta) \sum_{n=0}^{\infty} \|\Delta_n \phi(\cdot)\|_{L^\infty([0,T]\times\Omega; L_2^{0,0})}.
\end{aligned}$$

Our aim is now to estimate I. We set

$$A_{n',\ell}(\tau,s)=\int_{-\infty}^{s}\theta^2(t)e^{itn'^3}\varphi_{n',\ell}(t)\frac{e^{-i\tau t}}{i\tau}d\beta_\ell(t)$$

so that

$$A_{n',\ell}(\tau)=\int_{-\infty}^{+\infty}dA_{n',\ell}(\tau,s).$$

Moreover, using the Itô formula, we have

$$\begin{aligned}\left|\sum_{\ell=0}^{+\infty}A_{n',\ell}(\tau)\right|^2&=\int_{\mathbb{R}}d\left|\sum_{\ell\in\mathbb{N}}A_{n',\ell}(\tau,t)\right|^2\\&=2\,\mathrm{Re}\left(\sum_{\ell,m=0}^{\infty}\int_{\mathbb{R}}\int_{-\infty}^{t}\theta^2(s)e^{isn'^3}\varphi_{n',\ell}(s)\frac{e^{-i\tau s}}{i\tau}\,d\beta_\ell(s)\right.\\&\qquad\left.\times\,\theta^2(t)e^{-itn'^3}\varphi_{n',m}(t)\frac{e^{i\tau t}}{-i\tau}d\beta_m(t)\right)+\sum_{\ell\in\mathbb{N}}\int_{\mathbb{R}}\theta^4(t)\frac{\varphi^2_{n',\ell}(t)}{\tau^2}dt\\&=I^1_{n'}(\tau)+I^2_{n'}(\tau).\end{aligned}$$

Hence, again, two terms are involved in the estimate of I. The estimate of the second term is immediate. Indeed, we have

$$\begin{aligned}&\mathbb{E}\left(\sum_{n=0}^{\infty}\sup_{k\in\mathbb{N}}2^{k/2}\left(\sum_{|n'|\sim2^n}\int_{|\tau|\sim2^k}I^2_{n'}(\tau)d\tau\right)^{1/2}\right)\\&\quad=\mathbb{E}\left(\sum_{n=0}^{\infty}\sup_{k\in\mathbb{N}}2^{k/2}\left(\sum_{|n'|\sim2^n}\int_{|\tau|\sim2^k}\sum_{\ell\in\mathbb{N}}\int_{\mathbb{R}}\theta^4(t)\frac{\varphi^2_{n',\ell}(\tau)}{\tau^2}dt\,d\tau\right)^{1/2}\right)\\&\quad\le C(\theta)\sum_{n=0}^{\infty}\sup_{k\in\mathbb{N}}2^{k/2}\left(\int_{|\tau|\sim2^k}\frac{1}{\tau^2}\left|\sum_{|n'|\sim2^n}\sum_{\ell\in\mathbb{N}}\varphi^2_{n',\ell}(\cdot)\right|_{L^\infty([0,T]\times\Omega)}d\tau\right)^{1/2}\\&\quad\le C(\theta)\sum_{n=0}^{\infty}\|\Delta_n\phi(\cdot)\|_{L^\infty([0,T]\times\Omega;L^{0,0}_2)}.\end{aligned}\tag{2.3}$$

In order to estimate the contribution of the stochastic integral, i.e., of $I^1_{n'}(\tau)$ in the bound of I, we start with the estimate

(2.4)

$$\begin{aligned}
&\mathbb{E}\left(\left|\sum_{|n'|\sim 2^n}\int_{|\tau|\sim 2^k} I^1_{n'}(\tau)d\tau\right|^2\right)\\
&\quad=\mathbb{E}\left(\left|\sum_{|n'|\sim 2^n}\int_{|\tau|\sim 2^k} 2\,\mathrm{Re}\left(\sum_{\ell,m=0}^{\infty}\int_{\mathbb{R}}\int_{-\infty}^{t}\theta^2(s)e^{isn'^3}\varphi_{n',\ell}(s)\frac{e^{-i\tau s}}{i\tau}\,d\beta_\ell(s)\right.\right.\right.\\
&\qquad\qquad\left.\left.\left.\times\,\theta^2(t)e^{-itn'^3}\varphi_{n',m}(t)\frac{e^{i\tau t}}{-i\tau}\,d\beta_m(t)\right)d\tau\right|^2\right)\\
&\quad=\mathbb{E}\left(\left|\sum_{m=0}^{\infty}\int_{\mathbb{R}}\sum_{|n'|\sim 2^n}\sum_{\ell=0}^{\infty}\int_{-\infty}^{t}2\,\mathrm{Re}\left(\theta^2(s)e^{-i(t-s)n'^3}\int_{|\tau|\sim 2^k}\frac{e^{i\tau(t-s)}}{\tau^2}d\tau\right.\right.\right.\\
&\qquad\qquad\left.\left.\left.\times\,\theta^2(t)\right)\varphi_{n',\ell}(s)\,d\beta_\ell(s)\varphi_{n',m}(t)\,d\beta_m(t)\right|^2\right)\\
&\quad=\sum_{m=0}^{\infty}\int_{\mathbb{R}}\mathbb{E}\left|\sum_{|n'|\sim 2^n}\sum_{\ell=0}^{\infty}\int_{-\infty}^{t}2\,\mathrm{Re}\left(\theta^2(s)\theta^2(t)e^{-i(t-s)n'^3}\int_{|\tau|\sim 2^k}\frac{e^{i\tau(t-s)}}{\tau^2}d\tau\right)\right.\\
&\qquad\qquad\left.\times\,\varphi_{n',\ell}(s)d\beta_\ell(s)\varphi_{n',m}(t)\right|^2 dt,
\end{aligned}$$

where we have used again the independence of the family $(\beta_m)_{m\in\mathbb{N}}$. Using now the Cauchy–Schwarz inequality in n', the above term is bounded by

(2.5)

$$\begin{aligned}
&\sum_{m=0}^{\infty}\int_{\mathbb{R}}\mathbb{E}\left(\sum_{|n'|\sim 2^n}\left(\sum_{\ell=0}^{\infty}\int_{-\infty}^{t}2\,\mathrm{Re}\left(\theta^2(s)\theta^2(t)e^{-i(t-s)n'^3}\right.\right.\right.\\
&\qquad\qquad\left.\left.\left.\times\int_{|\tau|\sim 2^k}\frac{e^{i\tau(t-s)}}{\tau^2}d\tau\right)\varphi_{n',\ell}(s)d\beta_\ell(s)\right)^2\sum_{|n'|\sim 2^n}\varphi^2_{n',m}(t)\right)dt\\
&\le\left(\sup_{[0,t]\times\Omega}\sum_{|n'|\sim 2^n}\sum_{m=0}^{\infty}|\varphi_{n',m}(t)|^2\right)\\
&\quad\times\int_{\mathbb{R}}\mathbb{E}\sum_{|n'|\sim 2^n}\left(\sum_{\ell=0}^{\infty}\int_{-\infty}^{t}2\,\mathrm{Re}\left(\theta^2(s)\theta^2(t)e^{-i(t-s)n'^3}\int_{|\tau|\sim 2^k}\frac{e^{i\tau(t-s)}}{\tau^2}d\tau\right)\right.\\
&\qquad\qquad\left.\times\,\varphi_{n',\ell}(s)\,d\beta_\ell(s)\right)^2 dt.
\end{aligned}$$

Concerning the first term in the right-hand side above, we have

$$\begin{aligned}
&\sup_{[0,T]\times\Omega} \sum_{|n'|\sim 2^n} \sum_{m=0}^{\infty} |\varphi_{n',m}(t)|^2 \\
&\quad \le \sup_{[0,T]\times\Omega} \sum_{m=0}^{\infty} \sum_{|n'|\sim 2^n} |\phi(t) e_m(n')|^2 \\
&\quad \le \sup_{[0,T]\times\Omega} \sum_{m=0}^{\infty} |\Delta_n \phi(t) e_m|^2_{L^2(\mathbb{T})} = \|\Delta_n \phi(\cdot)\|^2_{L^\infty([0,T]\times\Omega; L_2^{0,0})},
\end{aligned}$$

while the remaining term in (2.5) is bounded above by

$$\begin{aligned}
&\int_{\mathbb{R}} \sum_{|n'|\sim 2^n} \sum_{\ell=0}^{\infty} \int_{-\infty}^{t} \theta^4(t)\theta^4(s) \left| \int_{|\tau|\sim 2^k} \frac{e^{i\tau(t-s)}}{\tau^2} d\tau \right|^2 \mathbb{E}(|\varphi_{n',\ell}(s)|^2) ds\, dt \\
&\quad \le \|\Delta_n \phi(\cdot)\|^2_{L^\infty([0,T]\times\Omega; L_2^{0,0})} \int_{\mathbb{R}} \int_{-\infty}^{t} \theta^4(t)\theta^4(s) \left| \int_{|\tau|\sim 2^k} \frac{e^{i\tau(t-s)}}{\tau^2} d\tau \right|^2 ds\, dt.
\end{aligned}$$

We then notice that, by interpolation between the cases $\alpha = 0$ and $\alpha = 1$, for any $\alpha \in [0,1]$ there is a positive constant C_α such that

$$\left| \int_{|\tau|\sim 2^k} \frac{e^{i\tau(t-s)}}{\tau^2} d\tau \right| \le \frac{C_\alpha}{|t-s|^\alpha} 2^{-(1+\alpha)k}.$$

Applying this with $\alpha = 1/4$, we get that the second term in (2.5) is bounded above by

$$\begin{aligned}
&C 2^{-\frac{5}{2}k} \|\Delta_n \phi(\cdot)\|^2_{L^\infty([0,T]\times\Omega; L_2^{0,0})} \int_{\mathbb{R}} \int_{-\infty}^{t} \frac{\theta^4(t)\theta^4(s)}{\sqrt{t-s}} ds\, dt \\
&\quad \le C(\theta) 2^{-\frac{5}{2}k} \|\Delta_n \phi(\cdot)\|^2_{L^\infty([0,T]\times\Omega; L_2^{0,0})}.
\end{aligned}$$

Collecting all these estimates from (2.4), we get

$$\mathbb{E}\left(\left| \sum_{|n'|\sim 2^n} \int_{|\tau|\sim 2^k} I^1_{n'}(\tau) d\tau \right|^2 \right) \le C(\theta) 2^{-\frac{5}{2}k} \|\Delta_n \phi(\cdot)\|^4_{L^\infty([0,T]\times\Omega; L_2^{0,0})}$$

and we deduce from this latest inequality that

$$
\begin{aligned}
&\mathbb{E}\left(\sum_{n=0}^{\infty}\sup_{k\in\mathbb{N}}2^{k/2}\left(\sum_{|n'|\sim 2^n}\int_{|\tau|\sim 2^k} I^1_{n'}(\tau)d\tau\right)^{1/2}\right)\\
&\quad\le \mathbb{E}\left(\sum_{n=0}^{\infty}\sum_{k=0}^{\infty}2^{k/2}\left(\sum_{|n'|\sim 2^n}\int_{|\tau|\sim 2^k} I^1_{n'}(\tau)d\tau\right)^{1/2}\right)\\
&\quad\le \sum_{n=0}^{\infty}\sum_{k=0}^{\infty}2^{k/2}\left[\mathbb{E}\left(\left(\sum_{|n'|\sim 2^n}\int_{|\tau|\sim 2^k} I^1_{n'}(\tau)d\tau\right)^2\right)\right]^{1/4}\\
&\quad\le C(\theta)\sum_{n=0}^{\infty}\left(\sum_{k=0}^{\infty}2^{-\frac{k}{8}}\right)\|\Delta_n\phi(\cdot)\|_{L^\infty([0,T]\times\Omega;L_2^{0,0})},
\end{aligned}
$$

where we have used Hölder's inequality at the third line; this, together with (2.3), completes the proof of the estimate of I.

In this way, the first inequality in Proposition 2.1 is proved after an application of Lemma 1.6, with $\sigma' < \sigma$. The second inequality follows from the obvious fact that

$$
\begin{aligned}
&\sum_{n=0}^{\infty}\|\Delta_n\phi(\cdot)\|^2_{L^\infty([0,T]\times\Omega;L_2^{0,\sigma})}\\
&\quad=\sum_{n=0}^{\infty}2^{\sigma n}\|\Delta_n\phi(\cdot)\|^2_{L^\infty([0,T]\times\Omega;L_2^{0,0})}\\
&\quad\le\left(\sum_{n=0}^{\infty}2^{-2(s'-\sigma)n}\right)^{1/2}\left(\sum_{n=0}^{\infty}2^{s'n}\|\Delta_n\phi(\cdot)\|^2_{L^\infty([0,T]\times\Omega;L_2^{0,0})}\right)^{1/2}\\
&\quad\le C(s',\sigma)\|\phi(\cdot)\|_{L^\infty([0,T]\times\Omega;L_2^{0,s'})}.\qquad\square
\end{aligned}
$$

3. Bilinear estimates. We now turn to prove some bilinear estimates which will allow us to handle the nonlinear term in (1.5). The next one is the crucial estimate.

PROPOSITION 3.1. *Let $-\frac{1}{2}\le s\le 0$ and $f\in X^{s,1/2}_{1,1}$, $g\in X^{s,1/2}_{1,\infty}$; then $\partial_x(fg)\in X^{s,-1/2}_{1,1}$ and there is a constant $C>0$ such that*

$$
|\partial_x(fg)|_{X^{s,-1/2}_{1,1}}\le C|f|_{X^{s,1/2}_{1,1}}|g|_{X^{s,1/2}_{1,\infty}}.
$$

Proof. Let f and g be as above; using a duality argument, it is sufficient, as usually, to prove that for some constant $C>0$, and for any function h in $X^{-s,1/2}_{\infty,\infty}$—where $X^{-s,1/2}_{\infty,\infty}$ is defined in an obvious way by modifying the definition of $X^{-s,1/2}_{1,1}$—we have

$$
|\langle\partial_x(fg),h\rangle|\le C|f|_{X^{s,1/2}_{1,1}}|g|_{X^{s,1/2}_{1,\infty}}|h|_{X^{-s,1/2}_{\infty,\infty}}.
$$

Using the Plancherel theorem, one has

$$
|\langle\partial_x(fg),h\rangle|=\left|\sum_{n'\neq 0}\sum_{\substack{n'_1\neq 0\\ n'_1\neq n'}}\int_{\tau\in\mathbb{R}}\int_{\tau_1\in\mathbb{R}} n'\bar{\hat{h}}(\tau,n')\hat{g}(\tau_1,n'_1)\hat{f}(\tau-\tau_1,n'-n'_1)d\tau_1\,d\tau\right|.
$$

We will denote $\sigma = \sigma(\tau, n') = \tau - {n'}^3$, $\sigma_1 = \sigma(\tau_1, n_1')$, $\sigma_2 = \sigma(\tau - \tau_1, n' - n_1')$. We also set $\hat{G}(\tau, n') = {n'}^s \langle\sigma\rangle^{1/2} \hat{g}(\tau, n')$, $\hat{F}(\tau, n') = {n'}^s \langle\sigma\rangle^{1/2} \hat{f}(\tau, n')$, and $\hat{H}(\tau, n') = {n'}^{-s} \langle\sigma\rangle^{1/2} \bar{\hat{h}}(\tau, n')$, so that F, G, and H lie, respectively, in $X^{0,0}_{1,1}$, $X^{0,0}_{1,\infty}$, and $X^{0,0}_{\infty,\infty}$. It suffices to prove that

(3.1)
$$\sum_{n' \neq 0} \sum_{\substack{n_1' \neq 0 \\ n_1' \neq n'}} \int_{\tau \in \mathbb{R}} \int_{\tau_1 \in \mathbb{R}} \frac{|n'|^{1+s} |n_1'|^{-s} |n' - n_1'|^{-s} |\hat{H}_{\tau, n'}| \, |\hat{G}_{\tau_1, n_1'}| \, |\hat{F}_{\tau - \tau_1, n' - n_1'}|}{\langle\sigma\rangle^{1/2} \langle\sigma_1\rangle^{1/2} \langle\sigma_2\rangle^{1/2}} d\tau_1 \, d\tau \leq C |H|_{X^{0,0}_{\infty,\infty}} |G|_{X^{0,0}_{1,\infty}} |F|_{X^{0,0}_{1,1}},$$

where we use $\hat{H}_{\tau, n'}$ for $\hat{H}(\tau, n')$ and so on. We divide the region $(n', n_1', \tau, \tau_1) \in (\mathbb{Z} \setminus \{0\})^2 \times \mathbb{R}^2$ arising in the left-hand side of (3.1) into three subregions,

$$\begin{aligned} &\text{(Region I)} && \langle\sigma_1\rangle = \max\{\langle\sigma\rangle, \langle\sigma_1\rangle, \langle\sigma_2\rangle\}, \\ &\text{(Region II)} && \langle\sigma\rangle = \max\{\langle\sigma\rangle, \langle\sigma_1\rangle, \langle\sigma_2\rangle\}, \\ &\text{(Region III)} && \langle\sigma_2\rangle = \max\{\langle\sigma\rangle, \langle\sigma_1\rangle, \langle\sigma_2\rangle\}, \end{aligned}$$

and we estimate separately the contributions of each of these regions to the left-hand side of (3.1).

Region I. From the identity

$$3|n'| \, |n_1'| \, |n' - n_1'| = |\tau - {n'}^3 - (\tau_1 - {n_1'}^3) - ((\tau - \tau_1) - (n' - n_1')^3)|$$

we get as usual that in Region I,

$$\frac{1}{2} |n'|^2 \leq |n'| \, |n_1'| \, |n' - n_1'| \leq \langle\sigma_1\rangle,$$

so that for any $s \in [-\frac{1}{2}, 0]$,

$$|n'|^{1+s} |n_1'|^{-s} |n' - n_1'|^{-s} \leq C \langle\sigma_1\rangle^{1/2}.$$

Hence, it is sufficient to prove that the contribution of Region I to

$$I = \sum_{n' \neq 0} \sum_{\substack{n_1' \neq 0, \\ n_1' \neq n'}} \int_{\tau \in \mathbb{R}} \int_{\tau_1 \in \mathbb{R}} \frac{|\hat{H}_{\tau, n'}| \, |\hat{G}_{\tau_1, n_1'}| \, |\hat{F}_{\tau - \tau_1, n' - n_1'}|}{\langle\sigma\rangle^{1/2} \langle\sigma_2\rangle^{1/2}} d\tau_1 \, d\tau$$

is bounded above by $C|H|_{X^{0,0}_{\infty,\infty}} |G|_{X^{0,0}_{1,\infty}} |F|_{X^{0,0}_{1,1}}$. Again, we will divide Region I into several subregions.

Region I-a. We consider here the subregion for which $\langle\sigma\rangle \geq \frac{1}{4} {n'}^2$.

We then estimate the contribution of this region to I; it is bounded above by its contribution to

$$\sum_{n, n_1 \in \mathbb{N}} \sum_{k, k_1 \in \mathbb{N}} \sum_{\substack{|n'| \sim 2^n \\ |n_1'| \sim 2^{n_1} \\ n' \neq n_1'}} \int_{|\tau| \sim 2^k} \int_{|\tau_1| \sim 2^{k_1}} \frac{|\hat{H}_{\tau, n'}| \, |\hat{G}_{\tau_1, n_1'}| \, |\hat{F}_{\tau - \tau_1, n' - n_1'}|}{\langle\sigma\rangle^{1/2} \langle\sigma_2\rangle^{1/2}} d\tau_1 \, d\tau, \tag{3.2}$$

with the convention that for $k = 0$, $|\tau| \sim 2^k$ means $|\tau| \leq 2$. This latest term is bounded above, using the Cauchy–Schwarz inequality, by

$$\sum_{n,n_1\in\mathbb{N}} \sum_{k,k_1\in\mathbb{N}} \left(\sum_{|n_1'|\sim 2^{n_1}} \int_{|\tau_1|\sim 2^{k_1}} |\hat{G}_{\tau_1,n_1'}|^2 d\tau_1 \right)^{1/2} \times \left(\sum_{|n_1'|\sim 2^{n_1}} \int_{|\tau_1|\sim 2^{k_1}} \left(\sum_{|n'|\sim 2^n} \int_{|\tau|\sim 2^k} \frac{|\hat{H}_{\tau,n'}|\,|\hat{F}_{\tau-\tau_1,n'-n_1'}|}{\langle\sigma\rangle^{1/2}\langle\sigma_2\rangle^{1/2}} d\tau \right)^2 d\tau_1 \right)^{1/2}. \tag{3.3}$$

Now, we use the fact that in Region I-a, we have, for $\varepsilon > 0$ small, which will be chosen more precisely later,

$$\frac{1}{\langle\sigma(\tau,n')\rangle^{1/2}} \leq C|n'|^{-\varepsilon} \frac{1}{\langle\sigma(\tau,n')\rangle^{1/2-\varepsilon/2}},$$

and using the Cauchy–Schwarz inequality in (τ, n') in (3.3), it is bounded above by

(3.4)

$$\begin{aligned} & C \sum_{n,n_1\in\mathbb{N}} \sum_{k,k_1\in\mathbb{N}} \left(\sum_{|n_1'|\sim 2^{n_1}} \int_{|\tau_1|\sim 2^{k_1}} |\hat{G}_{\tau_1,n_1'}|^2 d\tau_1 \right)^{1/2} \\ & \qquad \times \left[\sum_{|n_1'|\sim 2^{n_1}} \int_{|\tau_1|\sim 2^{k_1}} \left(\sum_{|n'|\sim 2^n} \int_{|\tau|\sim 2^k} |n'|^{-2\varepsilon} \langle\sigma_{\tau,n'}\rangle^{-\varepsilon} |\hat{H}_{\tau,n'}|^2 |\hat{F}_{\tau-\tau_1,n'-n_1'}|^2 d\tau \right) \right. \\ & \qquad\qquad \left. \times \left(\sum_{|n'|\sim 2^n} \int_{|\tau\sim 2^k} \frac{d\tau}{\langle\sigma_{\tau,n'}\rangle^{1-2\varepsilon}\langle\sigma_{\tau-\tau_1,n'-n_1'}\rangle} \right) d\tau_1 \right]^{1/2} \\ & \leq C \sum_{n_1\in\mathbb{N}} \left[\sup_{k_1\in\mathbb{N}} \left(\sum_{|n_1'|\sim 2^{n_1}} \int_{|\tau_1|\sim 2^{k_1}} |\hat{G}_{\tau_1,n_1'}|^2 d\tau_1 \right)^{1/2} \right. \\ & \qquad \times \sup_{k_1,n,k\in\mathbb{N}} \sup_{|n_1'|\sim 2^{n_1}} \sup_{|\tau_1|\sim 2^{k_1}} \left(\sum_{|n'|\sim 2^n} \int_{|\tau|\sim 2^k} \frac{d\tau}{\langle\sigma_{\tau,n'}\rangle^{1-2\varepsilon}\langle\sigma_{\tau-\tau_1,n'-n_1'}\rangle} \right)^{1/2} \\ & \qquad \times \sum_{k_1,n,k\in\mathbb{N}} \left(\sum_{|n_1'|\sim 2^{n_1}} \int_{|\tau_1|\sim 2^{k_1}} \sum_{|n'|\sim 2^n} \int_{|\tau|\sim 2^k} |n'|^{-2\varepsilon} \langle\sigma_{\tau,n'}\rangle^{-\varepsilon} |\hat{H}_{\tau,n'}|^2 \right. \\ & \qquad\qquad \left. \left. \times\, |\hat{F}_{\tau-\tau_1,n'-n_1'}|^2 d\tau d\tau_1 \right)^{1/2} \right]. \end{aligned}$$

But now, the fact that

$$\int_{-\infty}^{+\infty} \frac{d\theta}{(1+|\theta|)^{1-2\varepsilon}(1+|\theta-a|)} \leq \frac{C}{(1+|a|)^{1-4\varepsilon}}$$

for $a \in \mathbb{R}$ and the proof of Lemma 5.1 in [15] show that there is a constant $C > 0$ such that

$$\begin{aligned}
&\sup_{n_1\in\mathbb{Z}^*}\sup_{\tau_1\in\mathbb{R}}\sup_{n,k\in\mathbb{N}}\left(\sum_{|n'|\sim 2^n}\int_{|\tau|\sim 2^k}\frac{d\tau}{\langle\sigma_{\tau,n'}\rangle^{1-2\varepsilon}\langle\sigma_{\tau-\tau_1,n'-n_1}\rangle}\right)\\
&\quad\le \sup_{n_1\in\mathbb{Z}^*}\sup_{\tau_1\in\mathbb{R}}\left(\sum_{\substack{n\in\mathbb{Z}\setminus\{0\}\\ n\neq n_1}}\int_{\tau\in\mathbb{R}}\frac{d\tau}{\langle\sigma_{\tau,n}\rangle^{1-2\varepsilon}\langle\sigma_{\tau-\tau_1,n-n_1}\rangle}\right)\\
&\quad\le C
\end{aligned}$$

for any $\varepsilon > 0$ such that $1 - 4\varepsilon \ge 3/4$, i.e., for any $\varepsilon \le 1/16$. On the other hand the last line in (3.4) is bounded above by

$$\begin{aligned}
&\sum_{\substack{n,n_1\in\mathbb{N}\\ k,k_1\in\mathbb{N}}}\left[\left(\sum_{|n'|\sim 2^n}\int_{|\tau|\sim 2^k}|n'|^{-\varepsilon}\langle\sigma\rangle^{-\varepsilon/2}|\hat H_{\tau,n'}|^2d\tau\right)^{1/2}\right.\\
&\qquad\left.\times\sup_{|n'|\sim 2^n}\sup_{|\tau|\sim 2^k}\left(\sum_{|n_1'|\sim 2^{n_1}}\int_{|\tau_1|\sim 2^{k_1}}|n'|^{-\varepsilon}\langle\sigma\rangle^{-\varepsilon/2}|\hat F_{\tau-\tau_1,n'-n_1'}|^2d\tau_1\right)^{1/2}\right]\\
&\le \sup_{n,k\in\mathbb{N}}\sum_{n_1,k_1\in\mathbb{N}}\sup_{\substack{|n'|\sim 2^n\\ |\tau|\sim 2^k}}\left(\sum_{|n_1'|\sim 2^{n_1}}\int_{|\tau_1|\sim 2^{k_1}}|n'|^{-\varepsilon}\langle\sigma\rangle^{-\varepsilon/2}|\hat F_{\tau-\tau_1,n'-n_1'}|^2d\tau_1\right)^{1/2}\\
&\quad\times\sum_{n,k\in\mathbb{N}}\left(\sum_{|n'|\sim 2^n}\int_{|\tau|\sim 2^k}|n'|^{-\varepsilon}\langle\sigma\rangle^{-\varepsilon/2}|\hat H_{\tau,n'}|^2d\tau\right)^{1/2}\\
&\le C_\varepsilon|H|_{X^{0,0}_{\infty,\infty}}\\
&\quad\times\sup_{n,k\in\mathbb{N}}\sum_{n_1,k_1\in\mathbb{N}}\left(\sup_{\substack{|n'|\sim 2^n\\ |\tau|\sim 2^k}}\left(\sum_{|n_1'|\sim 2^{n_1}}\int_{|t_1|\sim 2^{k_1}}|n'|^{-\varepsilon}\langle\sigma\rangle^{-\varepsilon/2}|\hat F_{\tau-\tau_1,n'-n_1'}|^2d\tau_1\right)^{1/2}\right).
\end{aligned}$$

One may then notice that if $|n_1'| \ge 4|n'|$, then $|n' - n_1'| \sim 2^{n_1}$, and if $|\tau_1| \ge 4|\tau|$, then $|\tau - \tau_1| \sim 2^{k_1}$, so that for any $n, k \in \mathbb{N}$,

$$\begin{aligned}
&\sum_{n_1,k_1\in\mathbb{N}}\sup_{\substack{|n'|\sim 2^n\\ |\tau|\sim 2^k}}\left(\sum_{\substack{|n_1'|\sim 2^{n_1}\\ |n_1'|\ge 4|n'|}}\int_{\substack{|\tau_1|\sim 2^{k_1}\\ |\tau_1|\ge 4|\tau|}}|n'|^{-\varepsilon}\langle\sigma\rangle^{-\varepsilon/2}|\hat F_{\tau-\tau_1,n'-n_1'}|^2d\tau_1\right)^{1/2}\\
&\quad\le C\sum_{n_1,k_1\in\mathbb{N}}\left(\sum_{|n_1'|\sim 2^{n_1}}\int_{|\tau_1|\sim 2^{k_1}}|\hat F_{\tau_1,n_1'}|^2d\tau_1\right)^{1/2}\\
&\quad\le C|F|_{X^{0,0}_{1,1}},
\end{aligned}$$

while if $|n_1'| \le 4|n'|$ (still with $|\tau_1| \ge 4|\tau|$), then $|n'|^{-\varepsilon} \le C|n_1'|^{-\varepsilon}$ and for all $n, k, \in \mathbb{N}$,

$$\begin{aligned}
&\sum_{n_1,k_1\in\mathbb{N}} \sup_{\substack{|n'|\sim 2^n\\ |\tau|\sim 2^k}} \left(\sum_{\substack{|n_1'|\sim 2^{n_1}\\ |n_1'|\le 4|n'|}} \int_{\substack{|\tau_1|\sim 2^{k_1}\\ |\tau_1|\ge 4|\tau|}} |n'|^{-\varepsilon}\langle\sigma\rangle^{-\varepsilon/2} |\hat{F}_{\tau-\tau_1,n'-n_1'}|^2 d\tau_1 \right)^{1/2} \\
&\quad \le C \sum_{n_1\in\mathbb{N}} \left(\sup_{|n_1'|\sim 2^{n_1}} |n_1'|^{-\varepsilon} \right) \sum_{k_1\in\mathbb{N}} \sup_{|n'|\sim 2^n} \left(\sum_{|n_1'|\sim 2^{n_1}} \int_{|\tau_1|\sim 2^{k_1}} |\hat{F}_{\tau_1,n'-n_1'}|^2 d\tau_1 \right)^{1/2} \\
&\quad \le C_\varepsilon \sum_{k_1\in\mathbb{N}} \left(\sum_{\ell\in\mathbb{Z}} \int_{|\tau_1|\sim 2^{k_1}} |\hat{F}_{\tau,\ell}|^2 d\tau \right)^{1/2} \\
&\quad \le C_\varepsilon \sum_{k_1\in\mathbb{N}} \sum_{\ell\in\mathbb{N}} \left(\sum_{|\ell'|\sim 2^{\ell}} \int_{|\tau_1|\sim 2^{k_1}} |\hat{F}_{\tau,\ell'}|^2 d\tau \right)^{1/2} \\
&\quad \le C_\varepsilon |F|_{X^{0,0}_{1,1}}.
\end{aligned}$$

The cases for which $|\tau_1| \le 4|\tau|$ are treated in the same way as the latest case above, using the fact that in this case, $\langle\sigma\rangle^{-\varepsilon} \le C\langle\tau_1 - n'^3\rangle^{-\varepsilon}$, so that the sum over k_1 converges.

It follows from these estimates that (3.4) is bounded above by

$$C|G|_{X^{0,0}_{\infty,\infty}} |H|_{X^{0,0}_{\infty,\infty}} |F|_{X^{0,0}_{1,1}} \le C|G|_{X^{0,0}_{1,\infty}} |H|_{X^{0,0}_{\infty,\infty}} |F|_{X^{0,0}_{1,1}},$$

and this achieves the estimate of the contribution of Region I-a.

Region I-b. Assume here that $\langle\sigma_2\rangle \ge \frac{1}{4}n'^2$.

We may then proceed as in Region I-a by noticing that here we have, for $\varepsilon > 0$ small,

$$\frac{1}{\langle\sigma\rangle^{1/2}\langle\sigma_2\rangle^{1/2}} \le C_\varepsilon |n'|^{-\varepsilon}\langle\sigma\rangle^{-\varepsilon/2} \frac{1}{\langle\sigma\rangle^{1/2-\varepsilon/2}\langle\sigma_2\rangle^{1/2-\varepsilon/2}}$$

and that

$$\sup_{\substack{n,k\in\mathbb{N}\\ n_1'\in\mathbb{Z}\setminus\{0\}\\ \tau_1\in\mathbb{R}}} \left(\sum_{|n'|\sim 2^n} \int_{|\tau|\sim 2^k} \frac{d\tau}{\langle\sigma(\tau,n')\rangle^{1-\varepsilon}\langle\sigma(\tau-\tau_1,n'-n_1')\rangle^{1-\varepsilon}} \right) < +\infty$$

for any $\varepsilon \le 1/8$.

Region I-c. We consider now the region where $\langle\sigma\rangle \le \frac{1}{4}n'^2$ and $\langle\sigma_2\rangle \le \frac{1}{4}n'^2$.

The contribution of this region to I will be the most difficult to estimate. Again, we use in (3.2) the Cauchy–Schwarz inequality in (τ, n') to bound the contribution of

Region I-c to (3.2) by its contribution to

$$\sum_{n,n_1\in\mathbb{N}}\sum_{k,k_1\in\mathbb{N}}\Bigg(\sum_{|n_1'|\sim 2^{n_1}}\int_{|\tau_1|\sim 2^{k_1}}|\hat{G}_{\tau_1,n_1'}|^2d\tau_1\Bigg)^{1/2}$$
$$\times\Bigg[\sum_{|n_1'|\sim 2^{n_1}}\int_{|\tau_1|\sim 2^{k_1}}\Bigg(\sum_{|n'|\sim 2^{n}}\int_{|\tau|\sim 2^{k}}|\hat{H}_{\tau,n'}|^2|\hat{F}_{\tau-\tau_1,n'-n_1'}|^2d\tau\Bigg)$$
$$\times\Bigg(\sum_{|n'|\sim 2^{n}}\int_{|\tau|\sim 2^{k}}\frac{d\tau}{\langle\sigma\rangle\langle\sigma_2\rangle}\Bigg)\Bigg]^{1/2}$$
$$\le\sum_{n_1\in\mathbb{N}}\Bigg[\sup_{k_1\in\mathbb{N}}\Bigg(\sum_{|n_1'|\sim 2^{n_1}}\int_{|\tau_1|\sim 2^{k_1}}|\hat{G}_{\tau_1,n_1'}|^2d\tau_1\Bigg)^{1/2}$$
$$\times\sum_{n\in\mathbb{N}}\sum_{k_1\in\mathbb{N}}\sum_{k\in\mathbb{N}}\Bigg\{\sup_{|n_1'|\sim 2^{n_1}}\sup_{|\tau_1|\sim 2^{k_1}}\Bigg(\sum_{|n'|\sim 2^{n}}\int_{|\tau|\sim 2^{k}}\frac{d\tau}{\langle\sigma\rangle\langle\sigma_2\rangle}\Bigg)^{1/2}$$
$$\times\Bigg(\sum_{|n_1'|\sim 2^{n_1}}\int_{|\tau_1|\sim 2^{k_1}}\sum_{|n'|\sim 2^{n}}\int_{|\tau|\sim 2^{k}}|\hat{H}_{\tau,n'}|^2|\hat{F}_{\tau-\tau_1,n'-n_1'}|^2d\tau\,d\tau_1\Bigg)^{1/2}\Bigg\}\Bigg]$$
$$\le\sum_{n_1\in\mathbb{N}}\Bigg[\sup_{k_1\in\mathbb{N}}\Bigg(\sum_{|n_1'|\sim 2^{n_1}}\int_{|\tau_1|\sim 2^{k_1}}|\hat{G}_{\tau_1,n_1'}|^2d\tau_1\Bigg)^{1/2}$$
$$\times\sum_{n\in\mathbb{N}}\sum_{k_1\in\mathbb{N}}\Bigg\{\sum_{k\in\mathbb{N}}\sup_{|n_1'|\sim 2^{n_1}}\sup_{|\tau_1|\sim 2^{k_1}}\Bigg(\sum_{|n'|\sim 2^{n}}\int_{|\tau|\sim 2^{k}}\frac{d\tau}{\langle\sigma\rangle\langle\sigma_2\rangle}\Bigg)^{1/2}$$
$$\times\sup_{k\in\mathbb{N}}\Bigg(\sum_{|n_1'|\sim 2^{n_1}}\int_{|\tau_1|\sim 2^{k_1}}\sum_{|n'|\sim 2^{n}}\int_{|\tau|\sim 2^{k}}|\hat{H}_{\tau,n'}|^2|\hat{F}_{\tau-\tau_1,n'-n_1'}|^2d\tau\,d\tau_1\Bigg)^{1/2}\Bigg\}\Bigg],$$

and using the Cauchy–Schwarz inequality in n, this is bounded above by

(3.5)
$$\sum_{n_1\in\mathbb{N}}\Bigg\{\sup_{k_1\in\mathbb{N}}\Bigg(\sum_{|n_1'|\sim 2^{n_1}}\int_{|\tau_1|\sim 2^{k_1}}|\hat{G}_{\tau_1,n_1'}|^2d\tau_1\Bigg)^{1/2}$$
$$\times\sum_{k_1\in\mathbb{N}}\Bigg[\Bigg(\sum_{n\in\mathbb{N}}\Bigg(\sum_{k\in\mathbb{N}}B(n_1,k_1,n,k)\Bigg)^2\Bigg)^{1/2}$$
$$\times\Bigg(\sum_{n\in\mathbb{N}}\sup_{k\in\mathbb{N}}\sum_{|n_1'|\sim 2^{n_1}}\int_{|\tau_1|\sim 2^{k_1}}\sum_{|n'|\sim 2^{n}}\int_{|\tau|\sim 2^{k}}|\hat{H}_{\tau,n'}|^2|\hat{F}_{\tau-\tau_1,n'-n_1'}|^2d\tau\,d\tau_1\Bigg)^{1/2}\Bigg]\Bigg\},$$

with

$$B(n_1,k_1,n,k)=\sup_{\substack{|n_1'|\sim 2^{n_1}\\ |\tau_1|\sim 2^{k_1}}}\left(\sum_{|n'|\sim 2^n}\int_{|\tau|\sim 2^k}\frac{d\tau}{\langle\sigma\rangle\langle\sigma_2\rangle}\right)^{1/2}. \tag{3.6}$$

We will then make use of the following lemma.

LEMMA 3.2. *Let N be an integer, k_0 a function of $(n_1,k_1,n)\in\mathbb{N}^3$ with values in $\mathbb{N}$, and n_0 a function of $(n_1,k_1)\in\mathbb{N}^2$ with values in $\mathbb{N}$. Denote by $A(N,n_1,k_1)$ the region in $\mathbb{N}^2$ given by*

$$\begin{aligned}A(N,n_1,k_1)=\{(n,k)\in\mathbb{N}^2,\ k_0(n_1,k_1,n)\le k\le k_0(n_1,k_1,n)+N,\\ n_0(n_1,k_1)\le n\le n_0(n_1,k_1)+N\}.\end{aligned}$$

Then there is a constant $C(N)$ depending only on N such that

$$\sup_{n_1,k_1\in\mathbb{N}}\sum_{(n,k)\in A(N,n_1,k_1)}B(n_1,k_1,n,k)\le C(N),$$

where $B(n_1,k_1,n,k)$ is defined by (3.6).

Proof of Lemma 3.2. It follows easily from Lemma 5.1 in [15], since

$$\begin{aligned}&\sup_{k_1,n_1\in\mathbb{N}}\sum_{\substack{(n,k)\\ \in A(N,n_1,k_1)}}B(n_1,k_1,n,k)\\ &\quad\le N^2\sup_{k_1,n_1\in\mathbb{N}}\sup_{n,k\in\mathbb{N}}B(n_1,k_1,n,k)\\ &\quad\le N^2\sup_{k_1,n_1\in\mathbb{N}}\left(\sum_{\substack{n\in\mathbb{Z}\setminus\{0\}\\ n\ne n_1}}\int_{\mathbb{R}}\frac{d\tau}{\langle\sigma(\tau,n)\rangle\langle\sigma(\tau-\tau_1,n-n_1)\rangle}\right)^{1/2}\\ &\quad<+\infty\end{aligned}$$

by Lemma 5.1 in [15]. □

Now, in order to apply Lemma 3.2, we need to show that Region I-c is embedded in a region of the form

$$\{(n,k,n_1,k_1)\in\mathbb{N}^4,\ (n,k)\in A(N,n_1,k_1)\}$$

for some N and for some functions $n_0(n_1,k_1)$ and $k_0(n_1,k_1,n)$.

Note that we have, in Region I-c,

$$|\tau-n'^3|\le\langle\sigma(\tau,n)\rangle\le\frac{1}{4}n'^2\le\frac{1}{4}|n'|^3;$$

hence

$$\frac{3}{4}|n'|^3\le|\tau|\le\frac{5}{4}|n'|^3$$

and the property $3n-4\le k\le 3n+4$ follows easily. Hence, to prove the preceding result, we only have to find $n_0(n_1,k_1)$ and N such that for any (n,k,n_1,k_1) in Region I-c,

$$n_0(n_1,k_1) \le n \le n_0(n_1,k_1)+N.$$

In order to prove this fact, we again use a partition of Region I-c into three subregions.

- *Region* I-c-1: $2^{-12}|n_1'| \le |n'| \le 2^{12}|n_1'|$. In this region, we obviously have the result with $n_0(n_1,k_1) = n_1 - 4$.
- *Region* I-c-2: $|n'| \le 2^{-12}|n_1'|$. We recall that

$$|\sigma - \sigma_1 - \sigma_2| = 3|n'|\,|n_1'|\,|n'-n_1'|,$$

from which it follows that (since $\langle\sigma_1\rangle$ is dominant)

$$|n_1'|\,|n'|\,|n'-n_1'| \le \langle\sigma_1\rangle \le 3|n_1'|\,|n'|\,|n'-n_1'| + \langle\sigma\rangle + \langle\sigma_2\rangle;$$

using the fact that $|n'| \le \frac{1}{2}|n_1'|$ and that $\langle\sigma\rangle \le \frac{1}{4}|n'|^2$ and $\langle\sigma_2\rangle \le \frac{1}{4}|n'|^2$, from the preceding inequality we easily get

$$\frac{1}{2}|n_1'|^2|n'| \le \langle\sigma_1\rangle \le 5|n_1'|^2|n'|,$$

and the property follows easily with

$$n_0(n_1,k_1) = \frac{\ln|2^{k_1} - 2^{3n_1}|}{\ln 2} - \frac{\ln 5}{\ln 2} - 2n_1.$$

- *Region* I-c-3: $|n'| \ge 2^{12}|n_1'|$. We infer here, from the inequality

$$|n'|\,|n_1'|\,|n'-n_1'| \le \langle\sigma_1\rangle \le 3|n'|\,|n_1'|\,|n'-n_1'| + \langle\sigma\rangle + \langle\sigma_2\rangle,$$

that

$$\frac{1}{2}|n'|^2|n_1'| \le \langle\sigma_1\rangle \le 5|n'|^2|n_1'|,$$

and we conclude as in the preceding case.

Now, going back to (3.5), we may use Lemma 3.2 to show that the contribution of Region I-c to

$$\sup_{n_1,k_1\in\mathbb{N}} \left(\sum_{n\in\mathbb{N}} \left(\sum_{k\in\mathbb{N}} B(n_1,k_1,n,k) \right)^2 \right)^{1/2}$$
$$\le \sup_{n_1,k_1\in\mathbb{N}} \sum_{n,k\in\mathbb{N}} B(n_1,k_1,n,k)$$

is bounded above by an absolute constant.

Hence, each of the contributions of Regions I-c-1, I-c-2, and I-c-3 to (3.5) is bounded above by

(3.7)

$$
\begin{aligned}
&C\sum_{n_1\in\mathbb{N}}\Bigg\{\sup_{k_1\in\mathbb{N}}\left(\sum_{|n_1'|\sim 2^{n_1}}\int_{|\tau_1|\sim 2^{k_1}}|\hat{G}_{\tau_1,n_1}|^2d\tau_1\right)^{1/2}\\
&\qquad\times\sum_{k_1\in\mathbb{N}}\Bigg[\sum_{n=n_0(n_1,k_1)}^{n_0+N}\sup_{k_0(n)\le k\le k_0(n)+N}\sum_{|n_1'|\sim 2^{n_1}}\int_{|\tau_1|\sim 2^{k_1}}\sum_{|n'|\sim 2^{n}}\int_{|\tau|\sim 2^{k}}|\hat{H}_{\tau,n'}|^2\\
&\qquad\qquad\times|\hat{F}_{\tau-\tau_1,n'-n_1'}|^2d\tau\,d\tau_1\Bigg]^{1/2}\Bigg\}\\
&\le CN\sum_{n_1\in\mathbb{N}}\sup_{k_1\in\mathbb{N}}\left(\sum_{|n_1'|\sim 2^{n_1}}\int_{|\tau_1|\sim 2^{k_1}}|\hat{G}_{\tau_1,n_1'}|^2d\tau_1\right)^{1/2}\\
&\qquad\times\sup_{\substack{n_1\in\mathbb{N}\\ k_1\in\mathbb{N}}}\sup_{\substack{n_0\le n\le n_0+N\\ k_0\le k\le k_0+N}}\left(\sum_{|n'|\sim 2^{n}}\int_{|\tau|\sim 2^{k}}|\hat{H}_{\tau,n'}|^2d\tau\right)^{1/2}\\
&\qquad\times\sup_{n_1\in\mathbb{N}}\sum_{k_1\in\mathbb{N}}\sup_{\substack{n_0\le n\le n_0+N\\ k_0\le k\le k_0+N}}\sup_{\substack{|n'|\sim 2^{n}\\ |\tau|\sim 2^{k}}}\left(\sum_{|n_1'|\sim 2^{n_1}}\int_{|\tau_1|\sim 2^{k_1}}|\hat{F}_{\tau-\tau_1,n'-n_1'}|^2d\tau_1\right)^{1/2}\\
&\le CN|G|_{X^{0,0}_{1,\infty}}|H|_{X^{0,0}_{\infty,\infty}}\\
&\qquad\times\sup_{n_1\in\mathbb{N}}\sum_{k_1\in\mathbb{N}}\sup_{\substack{n_0\le n\le n_0+N\\ k_0\le k\le k_0+N}}\sup_{\substack{|n'|\sim 2^{n}\\ |\tau|\sim 2^{k}}}\left(\sum_{|n_1'|\sim 2^{n_1}}\int_{|\tau_1|\sim 2^{k_1}}|\hat{F}_{\tau-\tau_1,n'-n_1'}|^2d\tau_1\right)^{1/2}.
\end{aligned}
$$

It remains to bound the last term in the right-hand side above by $C|F|_{X^{0,0}_{1,1}}$. However, this is not completely obvious, and we again have to consider separately each of the Regions I-c-1, I-c-2, and I-c-3.

Region I-c-2. Recall that we have here $|n'|\le 2^{-12}|n_1'|$.

Then the last term in the right-hand side of the above inequality is clearly bounded by

$$
2\sup_{n_1\in\mathbb{N}}\sum_{k_1\in\mathbb{N}}\sum_{\substack{n=n_0(n_1,k_1)\\ n\le n_1-10}}^{n_0+N}\sum_{k=k_0}^{k_0+N}\left(\sum_{|n_1'|\sim 2^{n_1}}\int_{|\tau_1|\sim 2^{k_1}}|\hat{F}_{\tau-\tau_1,n'-n_1'}|^2d\tau_1\right)^{1/2}.
$$

- The contribution to this term of the k and k_1 for which $k\le k_1-4$ or $k\ge k_1+4$ is clearly bounded above by

$$
\begin{aligned}
&2N^2\sup_{n_1\in\mathbb{N}}\sum_{k\in\mathbb{N}}\left(\sum_{|n_1'|\sim 2^{n_1}}\int_{|\tau|\sim 2^{k}}|\hat{F}_{\tau,n_1'}|^2d\tau\right)^{1/2}\\
&\qquad\le 2N^2|F|_{X^{0,0}_{1,1}}.
\end{aligned}
$$

• It remains to consider the sum in k_1 and k, the contribution of the terms for which $k_1 - 4 \le k \le k_1 + 4$. Since for such terms, $\tau - \tau_1$ may stay bounded, we need to show that there are only a finite number of possibilities for k_1. We recall that in Region I-c, $k \le 3n + 4$, while in Region I-c-2, $n \le n_1 - 10$; it follows easily that if in addition $k_1 - 4 \le k \le k_1 + 4$, then $k_1 \le 3n_1 - 4$. Hence, $n_0(n_1, k_1) = \ln \frac{|2^{k_1} - 2^{3n_1}|}{\ln 2} - 2n_1 = n_1 - 1$ and the region is actually empty.

Region I-c-3: $|n'| \ge 2^{12}|n_1'|$. Again, the last term in (3.7) is easily bounded above by

$$2 \sup_{n_1 \in \mathbb{N}} \sum_{k_1 \in \mathbb{N}} \sum_{\substack{n = n_0(n_1,k_1) \\ n \ge n_1 + 3}}^{n_0+N} \sum_{k=k_0}^{k_0+N} \sup_{|\tau| \sim 2^k} \left(\sum_{|n'| \sim 2^n} \int_{|\tau_1| \sim 2^{k_1}} |\hat{F}_{\tau - \tau_1, n'}|^2 d\tau_1 \right)^{1/2}.$$

• In the same way as before, the contribution in this sum of the terms for which $k \le k_1 - 4$ or $k \ge k_1 + 4$ is bounded by

$$2N^2 \sup_{n \in \mathbb{N}} \sum_{k \in \mathbb{N}} \left(\sum_{|n'| \sim 2^n} \int_{|\tau| \sim 2^k} |\hat{F}_{\tau, n'}|^2 d\tau \right)^{1/2} \le 2N^2 |F|_{X^{0,0}_{1,1}}.$$

• In the region where $k_1 - 4 \le k \le k_1 + 4$, we easily get $k_1 \ge 3n_1 + 4$, and from the expression of $n_0(n_1, k_1)$ in Region I-c-3, we get $n_0(n_1, k_1) = \frac{1}{2}(k_1 - n_1)$. Hence, $n \ge \frac{1}{2}(3n - 8) - \frac{1}{2}n_1$, from which it follows that $n_1 \ge n - 8$, and again the region is empty, since $n \ge n_1 + 10$.

Region I-c-1: $2^{-12}|n_1'| \le |n'| \le 2^{12}|n_1'|$. This is the most difficult part; clearly, we can take in this region $n_0(n_1, k_1) = n_1$. Again, we will divide the region into three subregions depending on the size of k and k_1 compared to each other.

• $k \le k_1 - 4$: the contribution of this region to the last term in (3.7) is then bounded above by

$$2 \sup_{n_1 \in \mathbb{N}} \sum_{k_1 \in \mathbb{N}} \sum_{n=n_1}^{n_1+N} \sup_{|n'| \sim 2^n} \left(\sum_{|n_1'| \sim 2^{n_1}} \int_{|\tau_1| \sim 2^{k_1}} |\hat{F}_{\tau_1, n' - n_1'}|^2 d\tau_1 \right)^{1/2}.$$

Now, since, for each n_1, k_1, n, and n' such that $|n'| \sim 2^n$, one has

$$\left(\sum_{|n_1'| \sim 2^{n_1}} \int_{|\tau_1| \sim 2^{k_1}} |\hat{F}_{\tau_1, n' - n_1'}|^2 d\tau_1 \right)^{1/2} \le \sum_{\ell \in \mathbb{N}} \left(\sum_{|\ell'| \sim 2^\ell} \int_{|\tau_1| \sim 2^{k_1}} |\hat{F}_{\tau_1, \ell'}|^2 d\tau_1 \right)^{1/2},$$

the preceding term is easily bounded above by

$$2 \sup_{n_1 \in \mathbb{N}} \sum_{n=n_1}^{n_1+N} \sum_{k_1 \in \mathbb{N}} \sum_{\ell \in \mathbb{N}} \left(\sum_{|\ell'| \sim 2^\ell} \int_{|\tau_1| \sim 2^{k_1}} |\hat{F}_{\tau_1, \ell'}|^2 d\tau_1 \right)^{1/2} \le 2N |F|_{X^{0,0}_{1,1}}.$$

• $k_1 - 4 \le k \le k_1 + 4$: Using again the arguments immediately above, the contribution of this region to the last term in (3.7) may be bounded above by

$$\sup_{n_1\in\mathbb{N}} \sum_{k_1\in\mathbb{N}} \sup_{\substack{n_1\le n\le n_1+N\\ 3n\le k\le 3n+N\\ k_1-4\le k\le k_1+4}} \sup_{|\tau|\sim 2^k} \sum_{\ell\in\mathbb{N}} \left(\sum_{|\ell'|\sim 2^\ell} \int_{|\tau_1|\sim 2^{k_1}} |\hat{F}_{\tau-\tau_1,\ell'}|^2 d\tau_1 \right)^{1/2}$$

$$\le \sup_{n_1\in\mathbb{N}} \sum_{k_1\in\mathbb{N}} \sup_{\substack{3n_1\le k\le 3n_1+4N\\ k_1-4\le k\le k_1+4}} \sup_{|\tau|\sim 2^k} \sum_{\ell\in\mathbb{N}} \left(\sum_{|\ell'|\sim 2^\ell} \int_{|\tau_1|\sim 2^{k_1}} |\hat{F}_{\tau-\tau_1,\ell'}|^2 d\tau_1 \right)^{1/2}.$$

Here again, $\tau - \tau_1$ may stay bounded even for large k and k_1; however, for a fixed n_1, the number of k_1 for which the right-hand side gives a nonzero contribution is bounded by the total number of k_1 for which there exists at least one k such that $3n_1 \le k \le 3n_1 + 4N$ and $k_1 - 4 \le k \le k_1 + 4$. This number is bounded by $4N + 8$. In this way, the term above is bounded by

$$(4N+8) \sup_{n_1\in\mathbb{N}} \sup_{k,k_1\in\mathbb{N}} \sup_{|\tau|\sim 2^k} \sum_{\ell\in\mathbb{N}} \left(\sum_{|\ell'|\sim 2^\ell} \int_{|\tau_1|\sim 2^{k_1}} |\hat{F}_{\tau-\tau_1,\ell'}|^2 d\tau_1 \right)^{1/2}$$

$$\le (4N+8) \sum_{\ell\in\mathbb{N}} \sum_{j\in\mathbb{N}} \left(\sum_{|\ell'|\sim 2^\ell} \int_{|\tau|\sim 2^j} |\hat{F}_{\tau,\ell'}|^2 d\tau \right)^{1/2}$$

$$\le (4N+8) |F|_{X^{0,0}_{1,1}}.$$

• $k \ge k_1 + 4$: This region is a little bit more delicate than the preceding ones to handle. In the same way as before, we may bound above the contribution of the present region to the last term in the right-hand side of (3.7) by

$$\sup_{n_1\in\mathbb{N}} \sum_{k_1\in\mathbb{N}} \sup_{\substack{3n_1\le k\le 3n_1+2N\\ k\ge k_1+4}} \sup_{|\tau|\sim 2^k} \sum_{\ell\in\mathbb{N}} \left(\sum_{|\ell'|\sim 2^\ell} \int_{|\tau_1|\sim 2^{k_1}} |\hat{F}_{\tau-\tau_1,\ell'}|^2 d\tau_1 \right)^{1/2}$$

$$\le 2 \sup_{n_1\in\mathbb{N}} \sum_{k_1\in\mathbb{N}} \sup_{\substack{3n_1\le k\le 3n_1+2N\\ k\ge k_1+4}} \sum_{\ell\in\mathbb{N}} \left(\sum_{|\ell'|\sim 2^\ell} \int_{|\tau|\sim 2^k} |\hat{F}_{\tau,\ell'}|^2 d\tau \right)^{1/2}. \tag{3.8}$$

Again, we have to show that the number of possible k_1 (or n or n_1) in this region is finite. We recall that here, n and n' are of the same order; moreover, since $|\tau - {n'}^3| \le \frac{1}{4}|n'|^2$ and $|\tau - \tau_1 - (n' - n_1')^3| \le \frac{1}{4}|n'|^2$, it follows that $|\tau|$ is of the order of $|n'|^3$; then $|\tau_1|$, which is negligible compared with $|\tau|$, is negligible compared with $|n'|^3$. Hence $\tau_1 - {n_1'}^3 \sim -{n_1'}^3$ for n_1 sufficiently large. Now, we have the relation

$$\tau_1 - {n_1'}^3 - \tau + {n'}^3 + \tau - \tau_1 - (n' - n_1')^3 = 3n'n_1'(n' - n_1'). \tag{3.9}$$

• Consider first the case where n' and n_1' have opposite signs. Then, taking into account the preceding considerations, one may note that the left-hand side in (3.9) is of the order of $-{n_1'}^3$ (for $|n_1'|$ large), while the right-hand side has the sign of ${n_1'}^3$. Hence (3.9) cannot remain true for large $|n_1'|$, which implies that the number of n_1' in this region is finite.

• Now, if n' and n'_1 have the same sign, then comparing the signs of both sides in (3.9) shows that if $|n'_1|$ is large, then necessarily $n' - n'_1$ has a sign opposite to that of n'_1. But then, for $|n'|$ (or equivalently for $|n'_1|$) large, the facts that $\tau \sim n'^3$, $\tau - \tau_1 \sim (n' - n'_1)^3$, and τ_1 is negligible compared with τ lead again to incompatible signs.

This shows that, in any case, the number of possible n_1 (or n, or k_1) in this region is finite. Hence, (3.8) is bounded above by

$$C \sup_{n_1} \sup_{k_1} \sup_{3n_1 \le k \le 3n_1+2N} \sum_{\ell} \left(\sum_{|\ell'|\sim 2^{\ell}} \int_{|\tau|\sim 2^k} |\hat{F}_{\tau,\ell'}|^2 d\tau \right)^{1/2} \le C|F|_{X^{0,0}_{1,1}}.$$

This ends the proof of the required estimate in Region I, that is, when $\langle \sigma_1 \rangle$ dominates.

Region II. Here, we use the fact that

$$\frac{1}{2}|n'|^2 \le |n' n'_1 (n' - n'_1)| \le \langle \sigma \rangle$$

so that for any $s \in [-\frac{1}{2}, 0]$,

$$|n'|^{1+s} |n'_1|^{-s} |n' - n'_1|^{-s} \le C \langle \sigma \rangle^{1/2}.$$

Exchanging then the roles of n' and n'_1—and hence the roles of $\hat{G}$ and $\hat{H}$—we are led back to proving that the contribution of Region I to I is bounded above by $C|H|_{X^{0,0}_{1,\infty}} |G|_{X^{0,0}_{\infty,\infty}} |F|_{X^{0,0}_{1,1}}$.

For Regions I-a and I-b, this was already done, since the contribution of Regions I-a and I-b to I was actually bounded above by $C|H|_{X^{0,0}_{\infty,\infty}} |G|_{X^{0,0}_{\infty,\infty}} |F|_{X^{0,0}_{1,1}}$.

It remains only to consider the case of Region I-c. Again, the same computations as before lead to bounding the contribution of Region I-c to I as in (3.7), except that the sum over n_1 will be supported by $\hat{H}$ or $\hat{F}$, so that this contribution is bounded by (see (3.7))

(3.10)

$$\begin{aligned} &CN \sup_{n_1 \in \mathbb{N}} \sup_{k_1 \in \mathbb{N}} \left(\sum_{|n'_1|\sim 2^{n_1}} \int_{|\tau_1|\sim 2^{k_1}} |\hat{G}_{\tau_1, n'_1}|^2 d\tau_1 \right)^{1/2} \\ &\times \sum_{n_1 \in \mathbb{N}} \left\{ \sup_{k_1 \in \mathbb{N}} \sup_{n_0 \le n \le n_0+N} \sup_{k_0 \le k \le k_0+N} \left(\sum_{|n'|\sim 2^n} \int_{|\tau|\sim 2^k} |\hat{H}_{\tau,n'}|^2 d\tau \right)^{1/2} \right. \\ &\qquad \left. \times \sum_{k_1 \in \mathbb{N}} \sup_{n_0 \le n \le n_0+n} \sup_{k_0 \le k \le k_0+N} \sup_{\substack{|n'|\sim 2^n \\ |\tau|\sim 2^k}} \left(\sum_{|n'_1|\sim 2^{n_1}} \int_{|\tau_1|\sim 2^{k_1}} |\hat{F}_{\tau-\tau_1, n'-n'_1}|^2 d\tau_1 \right)^{1/2} \right\}. \end{aligned}$$

Hence, we have to bound above the last two lines in (3.10) by $C|H|_{X^{0,0}_{1,\infty}} |F|_{X^{0,0}_{1,1}}$.

Considering the way we have estimated the contribution of Regions I-c-2 and I-c-3 to (3.7), it is clear that the sum over k_1 in these regions can be supported by

$|\hat{F}_{\tau-\tau_1,n'-n_1'}|^2$; in Region I-c-1, we have $2^{-12}|n_1'| \le |n'| \le 2^{12}|n_1'|$ so that $n_0(n_1,k_1) = n_1$ and

$$\sum_{n_1\in\mathbb{N}} \sup_{k_1\in\mathbb{N}} \sup_{n_0\le n\le n_0+N} \sup_{k_0\le k\le k_0+N} \left(\sum_{|n'|\sim 2^n} \int_{|\tau|\sim 2^k} |\hat{H}_{\tau,n'}|^2 d\tau \right)^{1/2}$$
$$\le C \sum_{n_1\in\mathbb{N}} \sup_{k_1\in\mathbb{N}} \left(\sum_{|n'|\sim 2^{n_1}} \int_{|\tau_1|\sim 2^{k_1}} |\hat{H}_{\tau_1,n_1'}|^2 d\tau_1 \right)^{1/2}$$
$$\le C|H|_{X^{0,0}_{1,\infty}}.$$

We may conclude as before.

Region III. Again, exchanging the roles of $\hat{G}$ and $\hat{F}$, we are led back to proving that the contribution of Region I to I is bounded above by $C|G|_{X^{0,0}_{1,1}}|H|_{X^{0,0}_{\infty,\infty}}|F|_{X^{0,0}_{1,\infty}}$, but this is easily done by using the same analysis as for Region I. Hence, the proof of Proposition 3.1 is complete. □

We now prove that when local in time spaces are considered, that is, when $X^{s,-1/2}_{1,1}$ is replaced by $X^{s,-1/2,T}_{1,1}$, a small power of T can be recovered in the right-hand side of the estimate in Proposition 3.1. This will be useful in the contraction procedure, since as is now classical, no small power of T is gained, but on the contrary a $\ln T$ factor is lost in the estimate of the integral convolution with the linear semigroup when dealing with spaces of regularity $1/2$ in time.

The argument of the proof of the next proposition relies, as usual, on the fact that we have wasted a small power of $\langle\sigma\rangle$ or $\langle\sigma_2\rangle$ in Lemma 3.2. Actually, looking carefully to the proof shows that Lemma 3.2 is still true with $B(n_1,k_1,n,k)$ replaced by

$$\tilde{B}(n_1,k_1,n,k) = \sup_{\substack{|n_1'|\sim 2^{n_1}\\ |\tau_1|\sim 2^{k_1}}} \left(\sum_{|n'|\sim 2^n} \int_{|\tau|\sim 2^k} \frac{d\tau}{\langle\sigma\rangle^{1-\varepsilon}\langle\sigma_2\rangle^{1-\varepsilon}} \right)^{1/2}$$

for any $\varepsilon < 1/4$.

PROPOSITION 3.3. *Let $-1/2 \le s \le 0$ and $f \in X^{s,1/2,T}_{1,1}$, $g \in X^{s,1/2,T}_{1,\infty}$; then for any $\alpha < 1/16$, there is a constant C_α such that*

$$|\partial_x(fg)|_{X^{s,-1/2,T}_{1,1}} \le C_\alpha T^\alpha |f|_{X^{s,1/2,T}_{1,1}} |g|_{X^{s,1/2,T}_{1,\infty}}.$$

Proof. Let $f \in X^{s,1/2}_{1,1}$, $g \in X^{s,1/2}_{1,\infty}$, $s \ge 1/2$, both with support in $[-2T,2T]$. Using the arguments immediately above shows that we have actually proved, during the course of the proof of Proposition 3.1, that

$$|\partial_x(fg)|_{X^{s,-1/2}_{1,1}} \le C\left(|f|_{X^{s,1/2}_{1,1}} |g|_{X^{s,\delta}_{1,\infty}} + |f|_{X^{s,\delta}_{1,1}} |g|_{X^{s,1/2}_{1,\infty}} \right)$$

for any $\delta > 3/8$. Let $s = 0$ (the arguments are exactly the same if $s < 0$) and let δ be such that $3/8 < \delta < 1/2$. By an obvious interpolation inequality, one gets

$$|g|_{X^{0,\delta}_{1,\infty}} \le C|g|^{1-2\delta}_{X^{0,0}_{1,\infty}} |g|^{2\delta}_{X^{0,1/2}_{1,\infty}}.$$

On the other hand, using the notation introduced at the beginning of section 2, we have

$$\begin{aligned}
|g|_{X^{0,0}_{1,\infty}} &= \sum_{n=0}^{\infty} \sup_{k\in\mathbb{N}} \left(\sum_{n'\in\mathbb{N}\setminus\{0\}} \int_{|\tau|\sim 2^k} |\widehat{\Delta_n g}(\tau,n')|^2 d\tau \right)^{1/2} \\
&\le \sum_{n=0}^{\infty} |\Delta_n g|_{L^2_{x,t}([-2T,2T]\times\mathbb{T})} \\
&\le CT^{1/4} \sum_{n=0}^{\infty} |\Delta_n g|_{L^4_{x,t}([-2T,2T]\times\mathbb{T})} \\
&\le CT^{1/4} \sum_{n=0}^{\infty} \left(\sum_{n'\in\mathbb{N}\setminus\{0\}} \int_{\tau\in\mathbb{R}} \langle\sigma\rangle^{2/3} |\widehat{\Delta_n g}(\tau,n')|^2 d\tau \right)^{1/2},
\end{aligned}$$

where we have used in the last line above the Strichartz estimate proved in [4]. It follows readily that

$$|g|_{X^{0,0}_{1,\infty}} \le CT^{1/4} |g|_{X^{0,1/3}_{1,2}} \le CT^{1/4} |g|_{X^{0,1/2}_{1,\infty}},$$

and from the above interpolation inequality,

$$|g|_{X^{0,\delta}_{1,\infty}} \le CT^{(1-2\delta)/4} |g|_{X^{0,1/2}_{1,\infty}}.$$

In the same way, we estimate f as follows: taking a small positive ε, one has

$$|f|_{X^{0,\delta}_{1,1}} \le C|f|^{(1-2\delta)/(1+2\varepsilon)}_{X^{0,-\varepsilon}_{1,1}} |f|^{2(\varepsilon+\delta)/(1+2\varepsilon)}_{X^{0,1/2}_{1,1}}$$

and

$$|f|_{X^{0,-\varepsilon}_{1,1}} \le C|f|_{X^{0,0}_{1,2}} \le CT^{1/4} |f|_{X^{0,1/3}_{1,2}}$$

by again using the estimate in [4] for $\Delta_n f$; it follows that

$$|f|_{X^{0,-\varepsilon}_{1,1}} \le CT^{1/4} |f|_{X^{0,1/2}_{1,1}}.$$

Finally,

$$|\partial_x(fg)|_{X^{0,-1/2}_{1,1}} \le C_\alpha T^\alpha |f|_{X^{0,1/2}_{1,1}} |g|_{X^{0,1/2}_{1,\infty}},$$

where α is chosen such that $\alpha < (1-2\delta)/4$, with $\delta > 3/8$, so that at the very end, $\alpha < 1/16$, and since f and g have supports in $[-2T, 2T]$, the proof of Proposition 3.3 follows. □

We now prove an estimate of the same type as those in Propositions 3.1 and 3.3, but in Y_s spaces. We recall that the use of these spaces is needed to handle the integral estimate in Duhamel's formula (see Proposition 4.1).

PROPOSITION 3.4. *Let $-1/2 \le s \le 0$, $f \in X^{s,1/2}_{1,1}$, $g \in X^{s,1/2}_{1,\infty}$; then $\partial_x(fg) \in Y_s$. Moreover, for any $\alpha < 1/16$, there is a constant $C_\alpha > 0$ such that*

$$|\partial_x(fg)|_{Y_{s,T}} \le C_\alpha T^\alpha |f|_{X^{s,1/2,T}_{1,1}} |g|_{X^{s,1/2,T}_{1,\infty}}.$$

Proof. We only sketch the proof, since it is a slight modification of the proof of Proposition 3.1, using, e.g., the arguments in [20]. Let $f \in X^{s,1/2}_{1,1}$ and $g \in X^{s,1/2}_{1,\infty}$. We prove only the estimate

$$|\partial_x(fg)|_{Y_s} \leq C|f|_{X^{s,1/2}_{1,1}}|g|_{X^{s,1/2}_{1,\infty}};$$

the T^α factor can indeed be recovered exactly as in the proof of Proposition 3.3.

By a duality argument, the estimate will be proved if we show that there is a constant $C > 0$ such that for any function w (of the space variable x) lying in the Besov space $B^{-s}_{2,\infty}$, one has

$$\left| \sum_{n' \neq 0} |n'| \int_{\mathbb{R}} \frac{\widehat{fg}(\tau, n')}{\langle \sigma(\tau, n') \rangle} d\tau \hat{w}(n') \right| \leq C|f|_{X^{s,1/2}_{1,1}}|g|_{X^{s,1/2}_{1,\infty}}|w|_{B^{-s}_{2,\infty}}.$$

Setting as above $\hat{F}(\tau, n') = n'^s \langle \sigma(\tau, n') \rangle^{1/2} \hat{f}(\tau, n')$, $\hat{G} = n'^s \langle \sigma \rangle^{1/2} \hat{g}$, and $\hat{W} = n'^{-s} \hat{w}$, it suffices to prove that

(3.11)
$$\sum_{n' \neq 0} \sum_{\substack{n'_1 \neq 0 \\ n'_1 \neq n'}} \int_{\tau \in \mathbb{R}} \int_{\tau_1 \in \mathbb{R}} \frac{|n'|^{1+s} |n'_1|^{-s} |n' - n'_1|^{-s} |\hat{F}_{\tau - \tau_1, n' - n'_1}| \, |\hat{G}_{\tau_1, n'_1}| \, |\hat{W}_{n'}|}{\langle \sigma(\tau, n') \rangle \langle \sigma(\tau - \tau_1, n' - n'_1) \rangle^{1/2} \langle \sigma(\tau_1, n'_1) \rangle^{1/2}} d\tau \, d\tau_1$$
$$\leq C|F|_{X^{0,0}_{1,1}}|G|_{X^{0,0}_{1,\infty}}|W|_{B^0_{2,\infty}}.$$

Again, we will consider separately the three regions defined at the beginning of the proof of Proposition 3.1.

Region I: $\langle \sigma_1 \rangle = \max(\langle \sigma \rangle, \langle \sigma_1 \rangle, \langle \sigma_2 \rangle)$. As already noted, we have in this region

$$|n'|^{1+s} |n'_1|^{-s} |n' - n'_1|^{-s} \leq C \langle \sigma(\tau_1, n'_1) \rangle^{1/2}.$$

Hence, taking $\varepsilon > 0$ small, we have

$$\frac{|n'|^{1+s} |n'_1|^{-s} |n' - n'_1|^{-s} |\hat{F}_{\tau - \tau_1, n' - n'_1}| \, |\hat{G}_{\tau_1, n'_1}| \, |\hat{W}_{n'}|}{\langle \sigma(\tau, n') \rangle \langle \sigma(\tau - \tau_1, n' - n'_1) \rangle^{1/2} \langle \sigma(\tau_1, n'_1) \rangle^{1/2}}$$
$$\leq C \frac{|\hat{W}_{n'}|}{\langle \sigma(\tau, n') \rangle^{1/2+\varepsilon}} \frac{|\hat{F}_{\tau - \tau_1, n' - n'_1}| \, |\hat{G}_{\tau_1, n'_1}|}{\langle \sigma(\tau, n') \rangle^{1/2-\varepsilon} \langle \sigma(\tau - \tau_1, n' - n'_1) \rangle^{1/2}},$$

and we conclude as in the proof of Proposition 3.1, using the fact that $\mathcal{F}^{-1}(\frac{\hat{W}}{\langle \sigma \rangle^{1/2+\varepsilon}}) \in X^{0,0}_{\infty,\infty}$, with

$$\left| \mathcal{F}^{-1} \left(\frac{\hat{W}}{\langle \sigma \rangle^{1/2+\varepsilon}} \right) \right|_{X^{0,0}_{\infty,\infty}} \leq C_\varepsilon |W|_{B^0_{2,\infty}}$$

and using again the fact that Lemma 3.2 is still true with a smaller power of $\sigma(\tau, n')$.

Region III, that is, $\langle \sigma_2 \rangle = \max(\langle \sigma \rangle, \langle \sigma_1 \rangle, \langle \sigma_2 \rangle)$, is treated in the same way.

Region II: $\langle \sigma \rangle = \max(\langle \sigma \rangle, \langle \sigma_1 \rangle, \langle \sigma_2 \rangle)$. Here, we have

$$|n'|^{1+s} |n'_1|^{-s} |n' - n'_1|^{-s} \leq C \langle \sigma(\tau, n') \rangle^{1/2}$$

and it follows that

$$\frac{1}{\langle\sigma(\tau,n')\rangle} \le \frac{C}{\langle\sigma(\tau,n')\rangle + |n'|^{2+2s}|n_1'|^{-2s}|n'-n_1'|^{-2s}}.$$

Hence, going back to the way we have proved Proposition 3.1, it suffices to show that for a fixed n_1',

$$\mathcal{F}^{-1}\left(\frac{|n'|^{1+s}|n_1'|^{-s}|n'-n_1'|^{-s}\hat{W}(n')}{\langle\sigma(\tau,n')\rangle + |n'|^{2+2s}|n_1'|^{-2s}|n'-n_1'|^{-2s}}\right) \in X^{0,0}_{\infty,\infty},$$

with

$$\left|\mathcal{F}^{-1}\left(\frac{|n'|^{1+s}|n_1'|^{-s}|n'-n_1'|^{-s}\hat{W}(n')}{\langle\sigma(\tau,n')\rangle + |n'|^{2+2s}|n_1'|^{-2s}|n'-n_1'|^{-2s}}\right)\right|_{X^{0,0}_{\infty,\infty}} \le C|W|_{B^0_{2,\infty}}$$

and a constant C that does not depend on n_1'.

But this follows from the next easy computation, once we have noticed that $\int_{\mathbb{R}} \frac{d\tau}{(\langle\tau\rangle+a^2)^2} \le \frac{C}{a^2}$:

$$\begin{aligned}
&\left|\mathcal{F}^{-1}\left(\frac{|n'|^{1+s}|n_1'|^{-s}|n'-n_1'|^{-s}\hat{W}(n')}{\langle\sigma(\tau,n')\rangle + |n'|^{2+2s}|n_1'|^{-2s}|n'-n_1'|^{-2s}}\right)\right|^2_{X^{0,0}_{\infty,\infty}} \\
&\quad = \sup_{n,k} \sum_{|n'|\sim 2^n} \int_{|\tau|\sim 2^k} \frac{|n'|^{2+2s}|n_1'|^{-2s}|n'-n_1'|^{-2s}|\hat{W}(n')|^2}{(\langle\sigma(\tau,n')\rangle + |n'|^{2+2s}|n_1'|^{-2s}|n'-n_1'|^{-2s})^2} \\
&\quad \le \sup_{n} \left(\sum_{|n'|\sim 2^n} |n'|^{2+2s}|n_1'|^{-2s}|n'-n_1'|^{-2s}|\hat{W}(n')|^2 \right. \\
&\qquad\qquad \left. \times \int_{\mathbb{R}} \frac{d\tau}{(\langle\sigma(\tau,n')\rangle + |n'|^{2+2s}|n_1'|^{-2s}|n'-n_1'|^{-2s})^2} \right) \\
&\quad \le C|W|^2_{B^0_{2,\infty}}.
\end{aligned}$$

This ends the proof of Proposition 3.4. □

As a last, but easy, bilinear estimate, we briefly show that we can handle terms like $\partial_x(g^2)$ in $X^{s,-1/2}_{1,1}$ if g is only in $X^{s+\varepsilon,1/2}_{\infty,\infty}$ (the ε loss of regularity seems to be necessary here). Our motivation to treat such terms arises from the fact that the stochastic convolution which was studied in Proposition 2.1 belongs to such spaces (or even to $X^{s+\varepsilon,1/2}_{1,\infty}$) if sufficient regularity is assumed on the operator ϕ, but never belongs to $X^{s,1/2}_{1,1}$, due to the lack of regularity of the Brownian motion.

PROPOSITION 3.5. *Let $-1/2 \le s \le 0$ and $\varepsilon > 0$; then there is a constant $C > 0$ such that for any $g \in X^{s+\varepsilon,1/2}_{\infty,\infty}$,*

$$|\partial_x(g^2)|_{X^{s,-1/2}_{1,1}} \le C|g|^2_{X^{s+\varepsilon,1/2}_{\infty,\infty}}.$$

If, moreover, g is supported in $[-2T,2T]$ and $\partial_x(g^2)$ is considered in $X^{s,-1/2,T}_{1,1}$, then a factor T^α can be recovered in the right-hand side above for any $\alpha < 1/16$.

Finally, the same estimate holds if in the left-hand side, $X^{s,-1/2}_{1,1}$ (resp., $X^{s,-1/2,T}_{1,1}$) is replaced by Y_s (resp., $Y_{s,T}$).

Proof. Here again, we only sketch the proof, since the arguments are the same as in the easiest cases of the proof of Proposition 3.1, that is, when some small power of $\langle\sigma\rangle$ or $\langle\sigma_1\rangle$ can be lost.

Indeed, taking $f, g \in X^{s+\varepsilon,1/2}_{\infty,\infty}$, $h \in X^{-s,1/2}_{\infty,\infty}$, and setting as before $\hat{F} = n'^{s+\varepsilon}\langle\sigma\rangle^{1/2}\hat{f}$, $\hat{G} = n'^{s+\varepsilon}\langle\sigma\rangle^{1/2}\hat{g}$, and $\hat{H} = n'^{-s}\langle\sigma\rangle^{1/2}\hat{h}$, we need to show that

$$\sum_{n'\neq 0}\sum_{\substack{n'_1\neq 0\\ n'_1\neq n'}}\int_{\tau\in\mathbb{R}}\int_{\tau_1\in\mathbb{R}}\frac{|n'|^{1+s}|n'_1|^{-s-\varepsilon}|n'-n'_1|^{-s-\varepsilon}}{\langle\sigma\rangle^{1/2}\langle\sigma_1\rangle^{1/2}\langle\sigma_2\rangle^{1/2}}|\hat{H}_{\tau,n'}|\,|\hat{G}_{\tau_1,n'_1}|\,|\hat{F}_{\tau-\tau_1,n'-n'_1}|d\tau d\tau_1$$
$$\leq C|H|_{X^{0,0}_{\infty,\infty}}|F|_{X^{0,0}_{\infty,\infty}}|G|_{X^{0,0}_{\infty,\infty}}.$$

Consider, e.g., Region I, where $\langle\sigma_1\rangle$ dominates, and where we have the inequality

$$|n'|^{1+s}|n'_1|^{-s}|n'-n'_1|^{-s} \leq C\langle\sigma_1\rangle^{1/2},$$

so that we are led to estimate

$$\sum_{n'\neq 0}\sum_{\substack{n'_1\neq 0\\ n'_1\neq n'}}\int_{\tau\in\mathbb{R}}\int_{\tau_1\in\mathbb{R}}\frac{|n'_1|^{-\varepsilon}|n'-n'_1|^{-\varepsilon}}{\langle\sigma\rangle^{1/2}\langle\sigma_2\rangle^{1/2}}|\hat{H}_{\tau,n'}|\,|\hat{G}_{\tau_1,n'_1}|\,|\hat{F}_{\tau-\tau_1,n'-n'_1}|d\tau d\tau_1.$$

This latest term is then handled by the same arguments as those used in Region I-a in the proof of Proposition 3.1, keeping in addition a small power of $\langle\sigma_2\rangle$ to be able to sum over k, and hence to replace the norm $|F|_{X^{0,0}_{1,1}}$ by $|F|_{X^{0,0}_{\infty,\infty}}$ (the sum over n being handled by using $|n'-n'_1|^{-\varepsilon}$).

All the other regions are treated in the same way, and the arguments for the other statements of Proposition 3.5 are exactly the same as those of Propositions 3.3 and 3.4. ∎

4. Proofs of Theorems 1.2 and 1.5. As was pointed out in the introduction, it mainly remains to show that we may gain one degree of regularity in time when passing from $\partial_x(gf)$ to $\int_0^t U(t-s)\partial_x(fg)(s)ds$. The result is stated in the next proposition.

PROPOSITION 4.1. *There is a constant $C>0$ such that if $f \in X^{s,-1/2}_{1,1} \cap Y_s$, $s\in\mathbb{R}$, then $t \mapsto \int_0^t U(t-s)f(s)ds \in X^{s,1/2,T}_{1,1}$ and*

$$\left|\int_0^{\cdot} U(\cdot-s)f(s)ds\right|_{X^{s,1/2,T}_{1,1}} \leq C\left(|f|_{X^{s,-1/2}_{1,1}} + |f|_{Y_s}\right)$$

for any $T\leq 1$.

Moreover, for any $f\in Y_s$, the map $t\mapsto\int_0^t U(t-s)f(s)ds$ is continuous with values in $B^s_{2,1}(\mathbb{T})$ and there is a constant $C>0$ such that

$$\sup_{t\in[-T,T]}\left|\int_0^t U(t-s)f(s)ds\right|_{B^s_{2,1}} \leq C|f|_{Y_s}.$$

Proof. The arguments of the proof are similar to those in [14]. We consider a cut-off function ψ with $\psi\equiv 1$ on $[0,1]$ and $\operatorname{supp}\psi\subset[-1,2]$; it is sufficient to prove

that

$$\left|\psi \int_0^{\cdot} U(\cdot - s) f(s) ds\right|_{X_{1,1}^{s,1/2}} \leq C \left(|f|_{X_{1,1}^{s,-1/2}} + |f|_{Y_s} \right).$$

We first write

$$\begin{aligned}
&\psi(t) \int_0^t U(t-s) f(s) ds \\
&\quad = \psi(t) \sum_{n' \in \mathbb{Z}} \int_{|\tau_1 - n'^3| \leq 1} e^{ixn'} \hat{f}(\tau_1, n') \frac{e^{it(\tau_1 - n'^3)} - 1}{\tau_1 - n'^3} e^{itn'^3} d\tau_1 \\
&\qquad + \psi(t) \sum_{n' \in \mathbb{Z}} \int_{|\tau_1 - n'^3| \geq 1} e^{ixn'} \hat{f}(\tau_1, n') \frac{e^{it\tau_1} - e^{itn'^3}}{\tau_1 - n'^3} d\tau_1 \\
&\quad = g_1(t,x) + g_2(t,x).
\end{aligned}$$

To estimate g_1, we expand the exponential as

$$\frac{e^{it(\tau_1 - n'^3)} - 1}{\tau_1 - n'^3} = \sum_{k=1}^{\infty} \frac{i^k t^k (\tau_1 - n'^3)^k}{k!}$$

so that

$$g_1(t,x) = \sum_{k=1}^{\infty} \frac{i^k t^k}{k!} \psi(t) \sum_{n' \in \mathbb{Z}} \int_{|\tau_1 - n'^3| \leq 1} e^{ixn' + itn'^3} \hat{f}(\tau_1, n') (\tau_1 - n'^3)^k d\tau_1.$$

Let $\varphi_k(t) = t^k \psi(t)$; then

$$\hat{g}_1(\tau, n') = \sum_{k=1}^{\infty} \frac{i^k}{k!} \int_{|\tau_1 - n'^3| \leq 1} \hat{\varphi}_k(\tau - n'^3) \hat{f}(\tau_1, n') (\tau_1 - n'^3)^k d\tau_1$$

and it follows that

$$\begin{aligned}
&|g_1|_{X_{1,1}^{s,1/2}} \\
&= \sum_{n \in \mathbb{N}} 2^{sn} \sum_{\ell=0}^{\infty} \left(\sum_{|n'| \sim 2^n} \int_{|\tau| \sim 2^\ell} |\langle \tau - n'^3 \rangle^{1/2} \hat{g}_1(\tau, n')|^2 d\tau \right)^{1/2} \\
&= \sum_{n \in \mathbb{N}} 2^{sn} \sum_{\ell=0}^{\infty} \Bigg(\sum_{|n'| \sim 2^n} \int_{|\tau| \sim 2^\ell} \langle \tau - n'^3 \rangle \\
&\qquad \times \left[\int_{|\tau_1 - n'^3| \leq 1} \sum_{k=1}^{\infty} \frac{i^k}{k!} \hat{\varphi}_k(\tau - n'^3) \hat{f}(\tau_1, n') (\tau_1 - n'^3)^k d\tau_1 \right]^2 d\tau \Bigg)^{1/2}.
\end{aligned}$$

Now,

$$\sum_{|n'|\sim 2^n}\int_{|\tau|\sim 2^\ell}\langle \tau-n'^3\rangle\left[\int_{|\tau_1-n'^3|\le 1}\sum_{k=1}^{\infty}\frac{i^k}{k!}\hat\varphi_k(\tau-n'^3)\hat f(\tau_1,n')(\tau_1-n'^3)^k d\tau_1\right]^2 d\tau$$
$$\le \sum_{|n'|\sim 2^n}\int_{|\tau|\sim 2^\ell}\langle \tau-n'^3\rangle\left(\sum_{k=1}^{\infty}\frac{|\hat\varphi_k(\tau-n'^3)|}{k!}\right)^2 d\tau\left(\int_{|\tau_1-n'^3|\le 1}|\hat f(\tau_1,n')|^2 d\tau_1\right)$$
$$\le \sup_{|n'|\sim 2^n}\int_{|\tau|\sim 2^\ell}\langle \tau-n'^3\rangle\left(\sum_{k=1}^{\infty}\frac{|\hat\varphi_k(\tau-n'^3)|}{k!}\right)^2 d\tau\left(\sum_{|n'|\sim 2^n}\int_{|\tau_1-n'^3|\le 1}|\hat f(\tau_1,n')|^2 d\tau_1\right).$$

We deduce that

$$|g_1|_{X^{s,1/2}_{1,1}} \le \sup_{n\in\mathbb{N}}\sum_{\ell=0}^{\infty}\left(\sup_{|n'|\sim 2^n}\int_{|\tau|\sim 2^\ell}\langle \tau-n'^3\rangle\left(\sum_{k=1}^{\infty}\frac{|\hat\varphi_k(\tau-n'^3)|}{k!}\right)^2 d\tau\right)^{1/2}$$
$$\times\sum_{n\in\mathbb{N}}2^{sn}\left(\sum_{|n'|\sim 2^n}\int_{|\tau_1-n'^3|\le 1}|\hat f(\tau_1,n')|^2 d\tau_1\right)^{1/2}.$$

Now, we have, for $\varepsilon>0$,

$$\sup_{n\in\mathbb{N}}\sum_{\ell=0}^{\infty}\sup_{|n'|\sim 2^n}\int_{|\tau|\sim 2^\ell}\langle \tau-n'^3\rangle\left(\sum_{k=1}^{\infty}\frac{|\hat\varphi_k(\tau-n'^3)|}{k!}\right)^2 d\tau$$
$$\le \sup_{n\in\mathbb{N}}\sum_{\ell\in\mathbb{N}}\sup_{|n'|\sim 2^n}\left(\sup_{|\tau|\sim 2^\ell}\langle \tau-n'^3\rangle^{-\varepsilon}\right)\int_{|\tau|\sim 2^\ell}\langle \tau-n'^3\rangle^{1+\varepsilon}\left(\sum_{k=1}^{\infty}\frac{|\hat\varphi_k(\tau-n'^3)|}{k!}\right)^2 d\tau$$
$$\le C\left(\sup_{n\in\mathbb{N}}\sum_{\ell\in\mathbb{N}}\langle 2^\ell-2^{3n}\rangle^{-\varepsilon}\right)\sup_{n'\in\mathbb{N}}\int_{\mathbb{R}}\langle \tau-n'^3\rangle^{1+\varepsilon}\left(\sum_{k=1}^{\infty}\frac{|\hat\varphi_k(\tau-n'^3)|}{k!}\right)^2 d\tau$$
$$\le C\left|\sum_{k=1}^{\infty}\frac{\varphi_k}{k!}\right|_{H^{1/2+\varepsilon/2}}.$$

Hence

$$|g_1|_{X^{s,1/2}_{1,1}} \le C\left|\sum_{k=1}^{\infty}\frac{\varphi_k}{k!}\right|_{H^{1/2+\varepsilon/2}}|f|_{X^{s,0}_{1,1}}.$$

In order to estimate the norm of g_2, we write

$$g_2(t,x)=g_{2,1}(t,x)+g_{2,2}(t,x),$$

with

$$g_{2,1}(t,x)=\psi(t)\sum_{n'\in\mathbb{Z}}\int_{|\tau_1-n'^3|\ge 1}e^{ixn'}\hat f(\tau_1,n')\frac{e^{it\tau_1}}{\tau_1-n'^3}d\tau_1$$

and

$$g_{2,2}(t,x) = -\psi(t) \sum_{n' \in \mathbb{Z}} \int_{|\tau_1 - n'^3| \geq 1} e^{ixn'} \hat{f}(\tau_1, n') \frac{e^{itn'^3}}{\tau_1 - n'^3} d\tau_1.$$

We have

$$\hat{g}_{2,1}(\tau, n') = \int_{|\tau_1 - n'^3| \geq 1} \hat{\psi}(\tau - \tau_1) \frac{\hat{f}(\tau_1, n')}{\tau_1 - n'^3} d\tau_1,$$

and

$$\begin{aligned}
&\sum_{|n'| \sim 2^n} \int_{|\tau| \sim 2^k} \langle \tau - n'^3 \rangle |\hat{g}_{2,1}(\tau, n')|^2 d\tau \\
&\quad \leq C \sum_{|n'| \sim 2^n} \int_{|\tau| \sim 2^k} \left[\int_{|\tau_1 - n'^3| \geq 1} \langle \tau_1 - n'^3 \rangle^{1/2} |\hat{\psi}(\tau - \tau_1)| \left| \frac{\hat{f}(\tau_1, n')}{\tau_1 - n'^3} \right| d\tau_1 \right]^2 d\tau \\
&\qquad + C \sum_{|n'| \sim 2^n} \int_{|\tau| \sim 2^k} \left[\int_{|\tau_1 - n'^3| \geq 1} \langle \tau - \tau_1 \rangle^{1/2} |\hat{\psi}(\tau - \tau_1)| \left| \frac{\hat{f}(\tau_1, n')}{\tau_1 - n'^3} \right| d\tau_1 \right]^2 d\tau \\
&\quad \leq I + II.
\end{aligned}$$

For the term I, we have

$$I \leq C \sum_{|n'| \sim 2^n} \int_{|\tau| \sim 2^k} \left[\int_{\mathbb{R}} |\hat{\psi}(\tau_1)| \frac{|\hat{f}(\tau - \tau_1, n')|}{\langle \tau - \tau_1 - n'^3 \rangle^{1/2}} d\tau_1 \right]^2 d\tau.$$

Let $\hat{h} \in L^2_{\tau, n'}$; then

$$\begin{aligned}
&\left| \sum_{|n'| \sim 2^n} \int_{|\tau| \sim 2^k} \hat{h}(\tau, n') \int_{\mathbb{R}} |\hat{\psi}(\tau_1)| \frac{|\hat{f}(\tau - \tau_1, n')|}{\langle \tau - \tau_1 - n'^3 \rangle^{1/2}} d\tau_1 d\tau \right| \\
&\qquad \leq \int_{\mathbb{R}} |\hat{\psi}(\tau_1)| \left(\sum_{|n'| \sim 2^n} \int_{|\tau| \sim 2^k} |\hat{h}(\tau, n')|^2 d\tau \right)^{1/2} \\
&\qquad\qquad \times \left(\sum_{|n'| \sim 2^n} \int_{|\tau| \sim 2^k} \frac{|\hat{f}(\tau - \tau_1, n')|^2}{\langle \tau - \tau_1 - n'^3 \rangle} d\tau \right)^{1/2} d\tau_1.
\end{aligned}$$

We deduce from the preceding estimate that

$$I \leq C \left(\int_{\mathbb{R}} |\psi(\tau_1)| \left(\sum_{|n'| \sim 2^n} \int_{|\tilde{\tau} + \tau_1| \sim 2^k} \frac{|\hat{f}(\tilde{\tau}, n')|^2}{\langle \tilde{\tau} - n'^3 \rangle} d\tilde{\tau} \right)^{1/2} d\tau_1 \right)^2.$$

In the same way, we can prove that

$$II \leq C \left(\int_{\mathbb{R}} \langle \tau_1 \rangle^{1/2} |\psi(\tau_1)| \left(\sum_{|n'| \sim 2^n} \int_{|\tilde{\tau} + \tau_1| \sim 2^k} \frac{|\hat{f}(\tilde{\tau}, n')|^2}{\langle \tilde{\tau} - n'^3 \rangle^2} d\tilde{\tau} \right)^{1/2} d\tau_1 \right)^2.$$

Hence we have

$$I+II\leq C\left(\int_{\mathbb{R}}\langle\tau_1\rangle^{1/2}|\psi(\tau_1)|\left(\sum_{|n'|\sim 2^n}\int_{|\tilde\tau+\tau_1|\sim 2^k}\frac{|\hat f(\tilde\tau,n')|^2}{\langle\tilde\tau-n'^3\rangle}d\tilde\tau\right)^{1/2}d\tau_1\right)^2$$

and we deduce that

$$\sum_{k=0}^{\infty}\left(\sum_{|n'|\sim 2^n}\int_{|\tau|\sim 2^k}\langle\tau-n'^3\rangle|\hat g_{2,1}(\tau,n')|^2d\tau\right)^{1/2}$$

$$\leq C\sum_{k=0}^{\infty}\int_{\mathbb{R}}\langle\tau_1\rangle^{1/2}|\hat\psi(\tau_1)|\left(\sum_{|n'|\sim 2^n}\int_{|\tilde\tau+\tau_1|\sim 2^k}\frac{|\hat f(\tilde\tau,n')|^2}{\langle\tilde\tau-n'^3\rangle}d\tilde\tau\right)^{1/2}d\tau_1$$

$$\leq C\sum_{k_1=0}^{\infty}\int_{|\tau_1|\sim 2^{k_1}}\langle\tau_1\rangle^{1/2}|\hat\psi(\tau_1)|$$

$$\times\left(\sum_{k<k_1-4}+\sum_{k_1-4\leq k\leq k_1+4}+\sum_{k>k_1+4}\right)\left(\sum_{|n'|\sim 2^n}\int_{|\tilde\tau+\tau_1|\sim 2^k}\frac{|\hat f(\tilde\tau,n')|^2}{\langle\tilde\tau-n'^3\rangle}d\tilde\tau\right)^{1/2}d\tau_1.$$

Since

$$\left(\sum_{k<k_1-4}+\sum_{k_1-4\leq k\leq k_1+4}+\sum_{k>k_1+4}\right)\left(\sum_{|n'|\sim 2^n}\int_{|\tilde\tau+\tau_1|\sim 2^k}\frac{|\hat f(\tilde\tau,n')|^2}{\langle\tilde\tau-n'^3\rangle}d\tilde\tau\right)^{1/2}$$

$$\leq C(k_1-4)\left(\sum_{|n'|\sim 2^n}\sum_{j=-1}^{+1}\int_{|\tilde\tau|\sim 2^{k_1+j}}\frac{|\hat f(\tilde\tau,n')|^2}{\langle\tilde\tau-n'^3\rangle}d\tilde\tau\right)^{1/2}$$

$$+8\left(\sum_{|n'|\sim 2^n}\int_{|\tilde\tau|\sim 2^{k_1+5}}\frac{|\hat f(\tilde\tau,n')|^2}{\langle\tilde\tau-n'^3\rangle}d\tilde\tau\right)^{1/2}$$

$$+C\sum_{k\in\mathbb{N}}\left(\sum_{|n'|\sim 2^n}\int_{|\tilde\tau|\sim 2^k}\frac{|\hat f(\tilde\tau,n')|^2}{\langle\tilde\tau-n'^3\rangle}d\tilde\tau\right)^{1/2},$$

we may easily bound the preceding term by

$$C\sum_{k_1=0}^{\infty}\int_{|\tau_1|\sim 2^{k_1}}(1+k_1)2^{k_1/2}|\hat\psi(\tau_1)|d\tau_1\sum_{k=0}^{\infty}\left(\sum_{|n'|\sim 2^n}\int_{|\tau|\sim 2^k}\frac{|\hat f(\tilde\tau,n')|^2}{\langle\tilde\tau-n'^3\rangle}d\tilde\tau\right)^{1/2}$$

$$\leq C_\varepsilon|\langle\tau\rangle^{1/2+\varepsilon}\hat\psi|_{L^1(\mathbb{R})}\sum_{k=0}^{\infty}\left(\sum_{|n'|\sim 2^n}\int_{|\tau|\sim 2^k}\frac{|\hat f(\tilde\tau,n')|^2}{\langle\tilde\tau-n'^3\rangle}d\tilde\tau\right)^{1/2},$$

and thus

$$|g_{2,1}|_{X^{s,1/2}_{1,1}}\leq C_\varepsilon|\langle\tau\rangle^{1/2+\varepsilon}\hat\psi|_{L^1(\mathbb{R})}|f|_{X^{s,-1/2}_{1,1}}.$$

At last,

$$\hat{g}_{2,2}(\tau,n')=\hat{\psi}(\tau-n'^3)\int_{|\tau_1-n'^3|\geq 1}\frac{\hat{f}(\tau_1,n')}{\tau_1-n'^3}d\tau_1,$$

and hence

$$\begin{aligned}
&\sum_{n\in\mathbb{N}}2^{sn}\sum_{k\in\mathbb{N}}\left(\int_{|\tau|\sim 2^k}\sum_{|n'|\sim 2^n}\langle\tau-n'^3\rangle|\hat{g}_{2,2}(\tau,n')|^2d\tau\right)^{1/2}\\
&\quad\leq\sum_{n\in\mathbb{N}}2^{sn}\sum_{k\in\mathbb{N}}\left(\sup_{|n'|\sim 2^n}\int_{|\tau|\sim 2^k}\langle\tau-n'^3\rangle|\hat{\psi}(\tau-n'^3)|^2d\tau\right)^{1/2}\\
&\qquad\qquad\times\left(\sum_{|n'|\sim 2^n}\left(\int_{\mathbb{R}}\frac{|\hat{f}(\tau_1,n')|}{\langle\tau_1-n'^3\rangle}d\tau_1\right)^2\right)^{1/2}\\
&\quad\leq C\sum_{n\in\mathbb{N}}2^{sn}\left(\sum_{k\in\mathbb{N}}\langle 2^k-2^{3n}\rangle^{-\varepsilon}\right)\\
&\qquad\qquad\times\left(\sup_{|n'|\sim 2^n}\int_{\mathbb{R}}\langle\tau-n'^3\rangle^{1+\varepsilon}|\hat{\psi}(\tau-n'^3)|^2d\tau\right)^{1/2}\\
&\qquad\qquad\times\left(\sum_{|n'|\sim 2^n}\left(\int_{\mathbb{R}}\frac{|\hat{f}(\tau_1,n')|}{\langle\tau_1-n'^3\rangle}d\tau_1\right)^2\right)^{1/2}\\
&\quad\leq C|\psi|_{H^{1/2+\varepsilon/2}}|f|_{Y_s}.
\end{aligned}$$

This ends the proof of the first estimate in Proposition 4.1. The proof of continuity with values in $B^s_{2,1}$ and the second estimate follow in an obvious way from a slight modification of the proof of Lemma 2.2 in [10]. □

The next lemma shows that the free term in (1.5) belongs to $X^{\sigma,1/2,T}_{1,1}$ if u_0 is in $B^\sigma_{2,1}(\mathbb{T})$.

LEMMA 4.2. *Let $u_0\in B^\sigma_{2,1}(\mathbb{T})$ and $T\leq 1$. Then $U(t)u_0\in X^{\sigma,1/2,T}_{1,1}$ and there is a constant $C>0$ such that*

$$|U(t)u_0|_{X^{\sigma,1/2,T}_{1,1}}\leq C|u_0|_{B^\sigma_{2,1}}.$$

Proof. Let ψ be a cut-off function with $\psi\equiv 1$ on $[0,1]$ and let us prove that $|\psi U(t)u_0|_{X^{\sigma,1/2}_{1,1}}\leq C|u_0|_{B^\sigma_{2,1}}$.

We use the fact that $X^{\sigma,1/2+\varepsilon}_{1,\infty}\subset X^{\sigma,1/2}_{1,1}$ for any $\varepsilon>0$, and that

$$\widehat{\psi U(t)u_0}(\tau,n')=\hat{u}_0(n')\psi(\tau-n'^3)$$

to get the following bound:

$$
\begin{aligned}
&|\psi U(t)u_0|_{X^{\sigma,1/2}_{1,1}} \le C_\varepsilon |\psi U(t)u_0|_{X^{\sigma,1/2+\varepsilon}_{1,\infty}} \\
&\le C_\varepsilon \sum_{n=0}^{\infty} 2^{\sigma n} \sup_{k\ge 0} \left(\sum_{|n'|\sim 2^n} \int_{|\tau|\sim 2^k} \langle \tau - n'^3\rangle^{1+2\varepsilon} |\hat{u}_0(n')|^2 |\hat{\psi}(\tau - n'^3)|^2 d\tau \right)^{1/2} \\
&\le C_\varepsilon \sum_{n=0}^{\infty} 2^{\sigma n} \sup_{k\ge 0} \left(\sum_{|n'|\sim 2^n} \sum_{j=0}^{\infty} \int_{\substack{|\tau|\sim 2^k \\ |\tau - n'^3|\sim 2^j}} \langle \tau - n'^3\rangle^{1+2\varepsilon} |\hat{u}_0(n')|^2 |\hat{\psi}(\tau - n'^3)|^2 d\tau \right)^{1/2} \\
&\le C_\varepsilon \sum_{n=0}^{\infty} 2^{\sigma n} \sum_{j=0}^{\infty} \left(\sum_{|n'|\sim 2^n} \int_{|\tau - n'^3|\sim 2^j} \langle \tau - n'^3\rangle^{1+2\varepsilon} |\hat{u}_0(n')|^2 |\hat{\psi}(\tau - n'^3)|^2 d\tau \right)^{1/2} \\
&\le C_\varepsilon \sum_{n=0}^{\infty} 2^{\sigma n} \left(\sum_{|n'|\sim 2^n} |\hat{u}_0(n')|^2 \right)^{1/2} \sum_{j=0}^{\infty} \left(\int_{|\tau|\sim 2^j} \langle \tau\rangle^{1+2\varepsilon} |\hat{\psi}(\tau)|^2 d\tau \right)^{1/2} \\
&\le C_\varepsilon |u_0|_{B^\sigma_{2,1}} |\psi|_{B^{1/2+\varepsilon}_{2,1}}. \qquad \square
\end{aligned}
$$

Proof of Theorem 1.2. We now have all the estimates in hand, and we proceed exactly as in [3]; we work pathwise on (1.5), using a fixed point argument in the space $X^{\sigma,1/2,T}_{1,1}$ with $-1/2 \le \sigma < s$, s being defined by the assumption on ϕ, and $T \le 1$ sufficiently small.

Let u_0 be $\mathcal{F}_0$-measurable with $u_0 \in B^\sigma_{2,1}(\mathbb{T})$ almost surely, σ as above, and assume first that $\hat{u}_0(0) = 0$ almost surely. We set

$$z(t) = U(t)u_0; \tag{4.1}$$

then by Lemma 4.2, $z \in X^{\sigma,1/2,T}_{1,1}$ for any $T \le 1$ almost surely, and

$$|z|_{X^{\sigma,1/2,T}_{1,1}} \le C|u_0|_{B^\sigma_{2,1}} \text{ almost surely.} \tag{4.2}$$

Let $w(t)$ be defined by (1.4). By Proposition 2.1, $w \in X^{\sigma',1/2,T}_{1,\infty} \subset X^{\sigma',1/2,T}_{\infty,\infty}$ almost surely for any σ' with $\sigma < \sigma' < s$. We fix such a σ' and consider $\omega \in \Omega$ such that $u_0 \in B^\sigma_{2,1}(\mathbb{T})$ and $w \in X^{\sigma',1/2,T}_{1,\infty}$ for any $T \le 1$ almost surely.

In terms of $v(t) = u(t) - z(t) - w(t)$, (1.5) is written as

$$v(t) = \mathcal{T}v(t) := -\frac{1}{2}\int_0^t U(t-s)\partial_x(v^2 + w^2 + z^2 + 2vw + 2vz + 2wz)(s)ds. \tag{4.3}$$

Taking $0 < \alpha < 1/16$ in Propositions 3.3, 3.4, and 3.5, and applying Proposition 4.1, we easily get the existence of a constant $C_\alpha > 0$ such that

$$|\mathcal{T}v|_{X^{\sigma,1/2,T}_{1,1}} \le C_\alpha T^\alpha \Big(|v|^2_{X^{\sigma,1/2,T}_{1,1}} + |w|^2_{X^{\sigma,1/2,T}_{1,\infty}} + |u_0|^2_{B^\sigma_{2,1}} \Big).$$

In the same way, if v_1, $v_2 \in X^{\sigma,1/2,T}_{1,1}$, then

$$
\begin{aligned}
|\mathcal{T}v_1 - \mathcal{T}v_2|_{X^{\sigma,1/2,T}_{1,1}} \le C_\alpha T^\alpha \Big(&|v_1|_{X^{\sigma,1/2,T}_{1,1}} + |v_2|_{X^{\sigma,1/2,T}_{1,1}} \\
&+ |w|_{X^{\sigma,1/2,T}_{1,\infty}} + |u_0|_{B^\sigma_{2,1}} \Big) |v_1 - v_2|_{X^{\sigma,1/2,T}_{1,1}}.
\end{aligned}
$$

Hence, first setting

$$R_\omega^t = |w|_{X_{1,\infty}^{\sigma,1/2,t}} + |u_0|_{B_{2,1}^\sigma}$$

and then defining the stopping time T_ω by

$$T_\omega = \inf\{t > 0,\ 2C_\alpha t^\alpha R_\omega^t \geq 1/2\}$$

it is easily checked that $\mathcal{T}$ maps the ball of radius $R_\omega^{T_\omega}$ in $X_{1,1}^{\sigma,1/2,T_\omega}$ into itself, and that

$$|\mathcal{T}v_1 - \mathcal{T}v_2|_{X_{1,1}^{\sigma,1/2,T_\omega}} \leq \frac{3}{4}|v_1 - v_2|_{X_{1,1}^{\sigma,1/2,T_\omega}}.$$

Hence $\mathcal{T}$ has a unique fixed point, which is the unique solution of (4.3) in $X_{1,1}^{\sigma,1/2,T_\omega}$.

It follows from classical arguments and the second part of Proposition 4.1 that z and v are in $C([0,T_\omega];B_{2,1}^\sigma(\mathbb{T}))$ almost surely. On the other hand, since $\phi \in L_2^{0,s}$ and $U(t)$ is a unitary group in $H^s(\mathbb{T})$, we have $w \in C([0,T_\omega];H^s(\mathbb{T})) \subset C([0,T_\omega];B_{2,1}^\sigma(\mathbb{T}))$ by Theorem 6.10 in [8]. Hence, the solution $u = v + z + w$ of (1.5) is almost surely continuous with values in $B_{2,1}^\sigma(\mathbb{T})$.

One classically gets rid of the condition $\hat{u}_0(0) = 0$ almost surely by considering $v(t,x) = u(t,x+\alpha_0 t) - \alpha_0$ with $\alpha_0 = \int_{\mathbb{T}} u_0(x)dx$; indeed, v then satisfies the KdV equation (1.2) and the condition $\hat{v}_0(0) = 0$.

This ends the proof of Theorem 1.2. □

We now explain how we can get rid of the condition that the spatial mean of the noise is zero almost surely at any time.

PROPOSITION 4.3. *The conclusion of Theorem* 1.2 *is still true without the assumption that* $\operatorname{Im}\phi \subset \operatorname{span}\{e_j,\ j \geq 1\}$.

Proof. Let P be the orthogonal projector on $\operatorname{span}\{e_0\}$ in $L^2(\mathbb{T})$, i.e., $(Pu)(x) = (u,e_0)e_0$ for $u \in L^2(\mathbb{T})$, where $(\cdot,\cdot)$ denotes the inner product in $L^2(\mathbb{T})$. Then, clearly, $\tilde{\phi} = (I-P)\phi$ satisfies $\operatorname{Im}\tilde{\phi} \subset \operatorname{span}\{e_j,\ j \geq 1\}$; on the other hand, $W = P\phi\tilde{W} + \tilde{\phi}\tilde{W}$, and $\tilde{\beta}(t) = P\phi\tilde{W}(t) = \sum_{k\in\mathbb{N}}(\phi e_k, e_0)\beta_k(t)e_0$ is a real-valued Brownian motion since $\sum_{k\in\mathbb{N}}(\phi e_k, e_0)^2 = |\phi^* e_0|^2_{L^2(\mathbb{T})} < +\infty$.

Let $v = u - \tilde{\beta}$; then if u satisfies the KdV equation (1.2), v satisfies

$$\begin{cases} dv + (\partial_x^3 v + (v+\tilde{\beta})\partial_x v)dt = \tilde{\phi}d\tilde{W}, \\ v(0) = u_0, \end{cases}$$

and setting $\tilde{v}(t,x) = v(t, x + \int_0^t \tilde{\beta}(s)ds)$, we get the equation for $\tilde{v}$

$$\begin{cases} d\tilde{v} + (\partial_x^3\tilde{v} + \tilde{v}\partial_x\tilde{v})dt = d\hat{W}, \\ \tilde{v}(0) = u_0, \end{cases} \tag{4.4}$$

with $\hat{W}(t,x) = \sum_{k\in\mathbb{N}}(\tilde{\phi}e_k)(x - \int_0^t \tilde{\beta}(s)ds)\beta_k(t)$, and it is clear that we can apply all the arguments of the proof of Theorem 1.2 to (4.4), leading to the existence and uniqueness of $\tilde{v}$ from which we deduce the existence and uniqueness of u. Indeed, note that in Proposition 2.1, ϕ was allowed to depend on t and ω provided that it was in $L^\infty((0,T)\times\Omega;L_2^{0,s})$, which is obviously the case here. □

Proof of Theorem 1.5. The arguments are exactly the same as in [3]: let $T > 0$ be fixed; under the assumptions of Theorem 1.5, considering a sequence ϕ_n in $L_2^{0,4}$

such that $\phi_n \to \phi$ in $L_2^{0,0}$ and a sequence $u_{0,n}$ in $L^2(\Omega; H^3(\mathbb{T}))$ such that $u_{0,n} \to u_0$ in $L^2(\Omega; L^2(\mathbb{T}))$, one can easily prove (see [2]) the existence of a unique solution u_n in $C([0,T]; H^3(\mathbb{T}))$ of

$$u_n(t) = U(t)u_{0,n} - \frac{1}{2}\int_0^t U(t-s)\partial_x(u_n^2(s))ds + \int_0^t U(t-s)\phi_n d\tilde{W}(s).$$

Using Itô's formula on $|u_n|^2_{L^2(\mathbb{T})}$ and a martingale inequality, one gets as in [3]

$$\mathbb{E}\left(\sup_{t\in[0,T]} |u_n(t)|^2_{L^2(\mathbb{T})}\right) \le \mathbb{E}(|u_{0,n}|^2_{L^2(\mathbb{T})}) + C(T)\|\phi_n\|^2_{L_2^{0,0}};$$

hence, up to a subsequence, u_n converges in $L^2(\Omega; L^\infty(0,T; L^2(\mathbb{T})))$ weak star to some process $\tilde{u}$. Then if $\mathcal{T}_n$ is defined in the same way as $\mathcal{T}$ in the proof of Theorem 1.2, replacing u_0 and ϕ, respectively, by $u_{0,n}$ and ϕ_n, one shows that, given $\sigma < 0$, $\mathcal{T}_n$ is a uniform contraction in the ball of radius $R_\omega^{T_\omega}$ in $X_{1,1}^{\sigma,1/2,T_\omega}$; moreover the unique fixed point of $\mathcal{T}_n$ is equal to u_n, which, as a result, converges to u (the solution given by Theorem 1.2) in $X_{1,1}^{\sigma,1/2,T_\omega}$ for any $\sigma < 0$. It follows that $u = \tilde{u}$ almost surely on $[0, T_\omega]$, and that

$$|u(T_\omega)|_{B^\sigma_{2,1}(\mathbb{T})} \le C_\sigma |u(T_\omega)|_{L^2(\mathbb{T})} \le |\tilde{u}|_{L^\infty(0,T;L^2(\mathbb{T}))} \text{ almost surely.}$$

so that u may be extended to $[0,T]$ almost surely, giving the result. □

Acknowledgment. The authors would like to thank Professor Masayoshi Takeda for mentioning the regularity in Besov spaces of Brownian motion to them.

REFERENCES

[1] J. Bona and R. Smith, *The initial value problem for the Korteweg-de Vries equation*, Philos. Trans. Roy. Soc. London Ser. A, 278 (1975), pp. 555–601.
[2] A. de Bouard and A. Debussche, *On the stochastic Korteweg-de Vries equation*, J. Funct. Anal., 154 (1998), pp. 215–251.
[3] A. de Bouard, A. Debussche, and Y. Tsutsumi, *White noise driven Korteweg-de Vries equation*, J. Funct. Anal., 169 (1999), pp. 532–558.
[4] J. Bourgain, *Fourier restriction phenomena for certain lattice subsets and applications to nonlinear evolution equations, part* II, Geom. Funct. Anal., 3 (1993), pp. 209–262.
[5] H. Y. Chang, Ch. Lien, S. Sukarto, S. Raychaudhury, J. Hill, E. K. Tsikis, and K. E. Lonngren, *Propagation of ion-acoustic solitons in a non-quiescent plasma*, Plasma Phys. Control. Fusion, 28 (1986), pp. 675–681.
[6] Z. Ciesielski, *Orlicz spaces, spline systems, and Brownian motion*, Constr. Approx., 9 (1993), pp. 191–208.
[7] J. Colliander, M. Keel, G. Staffilani, H. Takaoka, and T. Tao, *Sharp global well-posedness for KdV and modified KdV on* $\mathbb{R}$ *and* $\mathbb{T}$, J. Amer. Math. Soc., 16 (2003), pp. 705–749.
[8] G. Da Prato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, Encyclopedia Math. Appl. 44, Cambridge University Press, Cambridge, UK 1992.
[9] J. Ginibre, *Le problème de Cauchy pour des EDP semi-linéaires périodiques en variables d'espace (d'après Bourgain)*, Séminaire Bourbaki 796, Astérisque, 237 (1996), pp. 163–187.
[10] J. Ginibre, Y. Tsutsumi, and G. Velo, *The Cauchy problem for the Zakharov system*, J. Funct. Anal., 151 (1997), pp. 384–436.
[11] O. Goubet, *Asymptotic smoothing effect for weakly damped forced Korteweg-de Vries equations*, Discrete Contin. Dynam. Systems, 6 (2000), pp. 625–644.
[12] R. Herman, *The stochastic, damped Korteweg-de Vries equation*, J. Phys. A., 23 (1990), pp. 1063–1084.

[13] C. E. Kenig, G. Ponce, and L. Vega, *Well-posedness of the initial value problem for the Korteweg-de Vries equation*, J. Amer. Math. Soc., 4 (1991), pp. 323–347.
[14] C. E. Kenig, G. Ponce, and L. Vega, *The Cauchy problem for the Korteweg-de Vries equation in Sobolev spaces of negative indices*, Duke Math. J., 71 (1993), pp. 1–21.
[15] C. E. Kenig, G. Ponce, and L. Vega, *A bilinear estimate with application to the KdV equation*, J. Amer. Math. Soc., 9 (1996), pp. 573–604.
[16] V. V. Konotop and L. Vasquez, *Nonlinear Random Waves*, World Scientific, Singapore, 1994.
[17] B. Roynette, *Mouvement brownien et espaces de Besov*, Stochastics Stochastics Rep., 43 (1993), pp. 221–260.
[18] J. C. Saut and R. Temam, *Remarks on the Korteweg-de Vries equation*, Israel J. Math., 24 (1976), pp. 78–87.
[19] M. Scalerandi, A. Romano, and C. A. Condat, *Korteweg-de Vries solitons under additive stochastic perturbations*, Phys. Rev. E, 58 (1998), pp. 4166–4173.
[20] N. Tzvetkov, *Remarque sur la régularité locale de l'équation de Kadomtsev-Petviashvili*-II, C. R. Acad. Sci. Paris Sér. I Math., 326 (1998), pp. 709–712.
[21] M. Wadati, *Stochastic Korteweg-de Vries equation*, J. Phys. Soc. Japan, 52 (1983), pp. 2642–2648.
[22] M. Wadati and Y. Akutsu, *Stochastic Korteweg-de Vries equation with and without damping*, J. Phys. Soc. Japan, 53 (1984), pp. 3342–3350.

SIAM J. MATH. ANAL.
Vol. 36, No. 3, pp. 856–881

HOMOGENIZATION OF TRANSPORT EQUATIONS: WEAK MEAN FIELD APPROXIMATION*

THIERRY GOUDON† AND FRÉDÉRIC POUPAUD‡

Abstract. We are interested, with respect to the small parameter ϵ, in the behavior of solutions ρ^ϵ of the conservative advection-diffusion equation $\partial_t\rho^\epsilon + \nabla_x\cdot(\rho^\epsilon u^\epsilon) = \eta\Delta_x\rho^\epsilon$, driven by a large velocity field, $|u^\epsilon| = \mathcal{O}(1/\epsilon)$, which oscillates periodically with respect to time and space variables. The novelty of our approach compared to that of previous works is that we deal with the periodic case in its full generality. In particular, the cell equation which allows us to compute effective coefficients is parabolic and not elliptic. We also derive estimates on the homogenized solution via entropy methods.

Key words. homogenization, advection-diffusion equation, entropy dissipation

AMS subject classifications. 35B27, 74Q15, 76M50

DOI. 10.1137/S0036141003415032

1. Introduction.

We consider the evolution of a scalar physical field $\rho(t,x)$ submitted to convection by a velocity field $u(t,x)$ and molecular diffusion with a diffusion coefficient $\eta > 0$. We are led to the classical equation

$$\partial_t\rho + \mathrm{div}_x(\rho u) = \eta\Delta_x\rho. \tag{1.1}$$

For instance, ρ can be the mass density or the temperature in a fluid. We can also interpret (1.1) as the Fokker–Planck equation for the probability density $\rho(t,x)$ of particles whose trajectories obey the differential Langevin equation

$$\mathrm{d}X(t) = u(t,X(t))\,\mathrm{d}t + \sqrt{2\eta}\,\mathrm{d}W(t), \tag{1.2}$$

W being a Brownian motion. We are interested in the effects of microscopic dynamics at large scale: the given velocity field u is an oscillating quantity with a very fast period. The fluctuations of u can be either deterministic or random. Without being more precise for the time being, we suppose that the variation of u depends on a characteristic scale ϵ, small compared to the large (macroscopic) scale of observation of the scalar field ρ. Therefore we are interested in deriving the bulk properties of ρ from the microscopic behavior of u, a homogenization problem.

While the problem (1.1) is linear, determination of the evolution of the average quantity $\langle\rho\rangle$ requires the knowledge of the correlation $\langle\rho u\rangle$. Indeed, averaging (1.1) yields

$$\partial_t\langle\rho\rangle = \mathrm{div}_x\langle\rho u\rangle + \eta\Delta_x\langle\rho\rangle.$$

The "turbulent moment closure problem" consists of looking for a relation between this correlation function and the average quantity $\langle\rho\rangle$. Classically, a simple linear

*Received by the editors July 22, 2003; accepted for publication (in revised form) February 27, 2004; published electronically October 14, 2004.
http://www.siam.org/journals/sima/36-3/41503.html

†Laboratoire Paul Painlevé, U.M.R. 8524, C.N.R.S.–Université des Sciences et Technologies de Lille, Cité Scientifique, F-59655 Villeneuve d'Ascq Cedex, France (thierry.goudon@math.univ-lille1.fr).

‡Laboratoire J. A. Dieudonné, U.M.R. 6621, C.N.R.S.–Université Nice-Sophia Antipolis, Parc Valrose, F-06108 Nice Cedex 2, France (poupaud@math.unice.fr).

relation is postulated,

$$\langle \rho u \rangle = \langle u \rangle \ \langle \rho \rangle + \alpha \langle \rho \rangle + \beta \nabla_x \langle \rho \rangle,$$

involving effective tensors α, β. It yields the average (effective) equation

$$\partial_t \langle \rho \rangle = \operatorname{div}_x \big((\langle u \rangle + \alpha) \langle \rho \rangle \big) + \operatorname{div}_x \Big((\eta + \beta) \nabla_x \langle \rho \rangle \Big)$$

with a modification of both convection and diffusion terms by the homogenization process. The aim of this work is to determine the effective coefficients α, β and to prove rigorously (in the mathematical sense) the convergence of average concentrations $\langle \rho \rangle$ towards solutions of the above effective equation.

The question of homogenizing transport equations with highly oscillating coefficients has motivated many works. Such a problem is related to the propagation of oscillations in fluid dynamics equations; see Di Perna and Majda [12]. Clearly limit processes depend strongly on the scaling used in the equation; some examples are provided in McLaughlin, Papanicolaou, and Pironneau [29]. We also mention the very complete and deep presentation of these problems in the recent review paper of Kramer and Majda [24]. We also refer to the classical book of Bensoussan, Lions, and Papanicolaou [6], or more recently Jikov, Kozlov, and Oleinik [22], for a presentation of homogenization problems and of the classical mathematical methods used to solve it.

The modeling of the fast oscillating (turbulent) velocity field u can be done in two different ways. The first consists of considering deterministic fields which are periodic with respect to some fast variables. The second corresponds to random fields. In this paper we focus on the former case (concerning the latter, general situations are treated by Kesten and Papanicolaou [23]; the interested reader will find details and more references, for instance, in [24]). Let us give some of the known results in this field.

The periodic case can induce resonant phenomena which make the problem considerably more difficult when the molecular diffusivity η vanishes. When $\eta = 0$, the problem of determining the limit of the solutions of the transport equation when the periodicity vanishes is a challenging open question. It has been attacked only by considering simplified geometry of the characteristics associated to u. The first studies are concerned essentially with shear flows: oscillations of the velocity field hold in a direction transverse to the variable of derivation of the scalar field. The approach initiated by Tartar [31, 32] has permitted one to bring out memory effects induced by the homogenization procedure as in the papers of Mascarenhas [28], Amirat, Hamdache, and Ziani [2, 3], Hamdache [19, 20] and others. These effects can also be interpreted as an increase in the order of the equation. The second simplified geometry situation which has been studied is the case of divergence free field in 2 space dimensions. The field is then a curl of a potential and the characteristics are level sets of this potential. This situation has been investigated in the works by Brenier [7], Hou and Xin [21], and E [13]. It cannot be generalized in higher dimension. Actually a common mistake is to assume a Fredholm alternative for transport equations with periodic conditions which does not hold in general.

When $\eta \neq 0$, the positive diffusivity changes completely the asymptotic regime. The case which has been the most studied is when the velocity field is divergence free. The divergence-free case is much easier from a mathematical point of view. Indeed we have in this situation L^2 and H^1 estimates which are uniform with respect to u

thanks to the relation

$$\frac{\mathrm{d}}{\mathrm{d}t}\int |\rho(t)|^2\,\mathrm{d}x + 2\eta\int |\nabla\rho(t)|^2\,\mathrm{d}x = 0.$$

The compactness properties induced by this relation help a lot to mathematically justify the homogenized limit (see, for instance, [6]). When dealing with divergence-free velocity fields, the original diffusion is enhanced by the homogenization process. On the contrary, compressible velocity fields can give rise to a diffusion coefficient which can be depleted in the limit. From a physical point of view this role of compressibility has been pointed out recently by Avellaneda and Vergassola [5]. There do not seem to be mathematical results in this case, and the present work is an attempt to fill this gap.

In this work the homogenization of advection-diffusion equations ($\eta \neq 0$) with periodic oscillation situations is investigated. The most general situation is considered. The velocity field depends on fast variables in time and space but depends also on the slow variables. The equation is written in conservative form. The scaling assumption leads to large fluctuations of the velocity field (order $1/\epsilon$, with ϵ the scale of space fluctuations of the velocity). It corresponds to the invariant scaling of the Navier–Stokes equation. Therefore this scaling is particularly relevant for the study of turbulent flows. In such a case, we cannot obtain compactness properties on the solutions ρ, and the only available estimate is in L^1. We use a method of oscillating test functions (in the spirit of Tartar [11, 31] or Evans [14, 15]).

Under the assumption that a certain mean value of the velocity field (the ballistic velocity) vanishes, we prove the convergence of the solutions to those of a drift-diffusion equation. According to [5], the effective diffusion coefficient is positive, but it can be depleted compared to the original one. The concentration is factorized by a bulk concentration times a given fast oscillating function. We also obtain uniform estimates on the bulk component, which allows us to obtain the uniqueness of the limit.

We also give partial answers to the problem when the ballistic velocity does not vanish. Actually when the velocity does not depend on the macroscopic variables we give a complete description of the homogenized limit. To the best of our knowledge this result is new even from a physical viewpoint.

The two main mathematical difficulties in this work are the following:

- The first one is, when considering oscillating in time velocities, that the cell equation is parabolic and not elliptic as usual. However, considering periodic boundary conditions, the corresponding operators are proved to satisfy the Fredholm alternative.
- The second one relies on the lack of regularity of the sequence of solutions. The only available estimate is in L^1. This is not sufficient to guarantee the uniqueness of the limit. To overcome this difficulty we obtain uniform estimates on the bulk component of the solutions by using an entropy method. It allows us to obtain enough regularity on the homogenized limit to justify the uniqueness of this limit.

2. The periodic homogenization problem. In this section we consider a passive scalar field,

$$\rho^\epsilon : (t,x) \in [0,\infty)\times\mathbb{R}^N \longmapsto \rho^\epsilon(t,x) \in \mathbb{R}^+.$$

It represents a physical quantity whose evolution is driven by a fluid flow. The velocity field of the fluid is

$$u^\epsilon : (t,x) \in [0,\infty) \times \mathbb{R}^N \longmapsto u^\epsilon(t,x) \in \mathbb{R}^N,$$

which is parametrized by $\epsilon > 0$. The parameter ϵ represents the length scale of oscillations of the velocity field u^ϵ. The scalar quantity is advected by the fluid and is also possibly subject to a diffusion process with a diffusion coefficient $\eta \geq 0$. Therefore, the scalar field ρ^ϵ is a solution of the following advection-diffusion equation:

$$\partial_t \rho^\epsilon(t,x) + \operatorname{div}_x\big(\rho^\epsilon u^\epsilon\big)(t,x) = \eta\ \Delta_x \rho^\epsilon(t,x) \quad \forall (t,x) \in (0,\infty) \times \mathbb{R}^N, \tag{2.1}$$

$$\rho^\epsilon(0,x) = \rho_I^\epsilon(x) \quad \forall x \in \mathbb{R}^N. \tag{2.2}$$

The scalar field $\rho_I^\epsilon : \mathbb{R}^N \to \mathbb{R}^+$ is a given initial data.

We are interested in situations where the velocity field can be a model for turbulence. In order to obtain in the limit $\epsilon \to 0$ an effective turbulent diffusion coefficient due to the velocity field, the strength of $|u^\epsilon|$ should be large, of order $1/\epsilon$. On the other hand, the velocity field gives rise to an effective drift term $\overline{u}$, which is the mean over fluctuations of u^ϵ. Therefore, the average value of u^ϵ should remain of order 1. Let us make these considerations more precise. We get u^ϵ of the form

$$u^\epsilon(t,x) = \frac{1}{\epsilon}\left(u^0\left(t,x;\frac{t}{\epsilon^2},\frac{x}{\epsilon}\right) + \epsilon\, u^1\left(t,x;\frac{t}{\epsilon^2},\frac{x}{\epsilon}\right)\right). \tag{2.3}$$

The velocity fields

$$u^{0,1} : (t,x;\tau,y) \in [0,\infty) \times \mathbb{R}^N \times \mathbb{R}^{N+1} \longmapsto u^{0,1}(t,x;\tau,y) \in \mathbb{R}^N$$

are periodic with respect to the fast variables $\tau,\ y$:

$$\begin{gathered}\forall (t,x;\tau,y) \in [0,\infty) \times \mathbb{R}^N \times \mathbb{R}^{N+1},\ \forall n = (n_0,n') \in \mathbb{Z} \times \mathbb{Z}^N,\\ u^{0,1}(t,x;\tau+n_0,y+n') = u^{0,1}(t,x;\tau,y).\end{gathered} \tag{2.4}$$

Finally, setting $Y = (0,1)^{N+1}$, the situation is very different if u^0 has a vanishing or not vanishing mean value

$$\forall (t,x) \in [0,\infty) \times \mathbb{R}^N, \quad \int_Y u^0(t,x;\tau,y)\,\mathrm{d}\mu(\tau,y) \tag{2.5}$$

for a measure $\mathrm{d}\mu(\tau,y)$, which will be made more precise later on. This mean value is the so-called ballistic velocity. Hence, in the vanishing ballistic velocity case, the quantity $\overline{u}(t,x) = \int_Y u^1(t,x;\tau,y)\,\mathrm{d}\mu(\tau,y)$ corresponds to the mean value of the velocity field u^ϵ.

Remark 1. The scaling we use in the fast variables, $u \to u^\epsilon = \frac{1}{\epsilon}u(\frac{t}{\epsilon^2},\frac{x}{\epsilon})$, is the invariant scaling of the incompressible Navier–Stokes equation. (If u is a solution of the Navier–Stokes equation, u^ϵ is also a solution of the same equation.) It shows the relevance of this scaling for studying turbulence in fluids.

Remark 2. Actually, we are working with dimensionless equations, obtained from the original variables (t',x') by setting $t = t'/T$, $x = x'/L$, where T and L are characteristic values of time and length, respectively, under which the evolution of ρ is studied. We point out that the choice of T and L is free. Let us denote by ℓ the typical length of microscopic variations of the velocity field and by τ_0 its typical time scale

of variation. The typical value for the mean velocity field is denoted by U_m and for the fluctuations by U_{fl}. We define the parameter ϵ as $\epsilon = \frac{U_m}{U_{fl}}$. It is natural to choose the scale T and L such that $U_m = \frac{L}{T}$. If we want that the fluctuations of the velocity field give rise to a diffusion process, U_{fl} should be scaled as a parabolic scaling. Then L and T are completely determined by the relation $U_{fl} = \frac{L^2}{T}\ \frac{1}{\ell}$. Consequently, we have $\epsilon = \frac{\ell}{L}$.

Up to now there has been no physical assumption. The regime we are studying in this work corresponds to the assumption $U_{fl} \gg U_m$ or, in other words, $\epsilon \ll 1$. It is also natural to assume that the path of a particle with velocity U_{fl} during the microscopic time period τ_0 is of order ℓ. It leads to

$$U_{fl}\,\tau_0 = \frac{L^2}{\ell}\ \frac{\tau_0}{T} = \ell, \quad \frac{\tau_0}{T} = \frac{\ell^2}{L^2} = \epsilon^2.$$

We have also to assume that the molecular diffusivity acts at the macroscopic scale L and T. If η' stands for the diffusion coefficient in the physical variables, it means that η' has the same order as $\frac{L^2}{T} = U_{fl}\ell$, while $\eta = \eta' \frac{T}{L^2}$. This physical situation corresponds to the so-called weak mean field approximation.

Remark 3. The case of a purely periodic and divergence-free velocity field with null average,

$$u^\epsilon(x) = \frac{1}{\epsilon}\ u(x/\epsilon), \qquad \int_Y u(y)\,\mathrm{d}(y) = 0, \qquad \mathrm{div}_y(u) = 0,$$

can actually be recast into a more classical question of homogenization of parabolic equation. Indeed, in such a case, we can associate to u a skew-symmetric matrix B by solving $\Delta_y B = \nabla_y u$. Using Fourier expansion yields

$$B_{ij}(y) = \sum_{n\neq 0} e^{2i\pi n\cdot y}\ \frac{i}{|n|^2}\Big(n_i\widehat{u_j}(n) - n_j\widehat{u_i}(n)\Big).$$

We have $\mathrm{Div}_y(B) = u(y)$ so that

$$\frac{1}{\epsilon}\ u(x/\epsilon)\cdot\nabla_x\rho + \eta\Delta\rho = \nabla_x\cdot((\eta + B(x/\epsilon))\nabla_x\rho).$$

This remark is used, for instance, in [4]. Difficulties arise when $\int_Y u\,\mathrm{d}(y) \neq 0$. This kind of problem arises, for instance, in neutron transport theory; we refer to Capdeboscq [8, 9] for a treatment of the spectral problem.

3. Formal asymptotics. As usual (see [6]), we try to guess the result with a formal double-scale ansatz,

$$\rho^\epsilon(t,x) = R^0(t,x;t/\epsilon^2,x/\epsilon) + \epsilon R^1(t,x;t/\epsilon^2,x/\epsilon) + \epsilon^2 R^2(t,x;t/\epsilon^2,x/\epsilon) + \cdots,$$

where the $R^i(t,x;\tau,y)$'s are Y-periodic. The action of the operator

$$\mathcal{T}_\epsilon = \partial_t(\cdot) + \mathrm{div}_x\big(u^\epsilon\ \cdot\big) - \eta\Delta_x(\cdot)$$

on functions of the form $r^\epsilon(x) = R(t,x;t/\epsilon^2,x/\epsilon)$ reads as

$$\mathcal{T}_\epsilon(r^\epsilon)\left(t,x;\frac{t}{\epsilon^2},\frac{x}{\epsilon}\right) = \epsilon^{-2}\mathcal{T}_0(R)\left(t,x;\frac{t}{\epsilon^2},\frac{x}{\epsilon}\right) + \epsilon^{-1}\mathcal{T}_1(R)\left(t,x;\frac{t}{\epsilon^2},\frac{x}{\epsilon}\right) + \mathcal{T}_2(R)\left(t,x;\frac{t}{\epsilon^2},\frac{x}{\epsilon}\right)$$

with

$$\begin{cases} \mathcal{T}_0(R)(t,x;\tau,y) = \Big(\partial_\tau R + \mathrm{div}_y\big(u^0 R\big) - \eta\Delta_y R\Big)(t,x;\tau,y), \\ \mathcal{T}_1(R)(t,x;\tau,y) = \Big(\mathrm{div}_y\big(u^1 R\big) + \mathrm{div}_x\big(u^0 R\big) - 2\eta\nabla_x\cdot\nabla_y R\Big)(t,x;\tau,y), \\ \mathcal{T}_2(R)(t,x;\tau,y) = \Big(\partial_t R + \mathrm{div}_x\big(u^1 R\big) - \eta\Delta_x R\Big)(t,x;\tau,y). \end{cases}$$

Plugging the double-scale ansatz into (2.1) and identifying the terms with the same power of ϵ yield

$$\begin{aligned} &\epsilon^{-2}\text{term:} && \mathcal{T}_0 R^0 = 0, \\ &\epsilon^{-1}\text{term:} && \mathcal{T}_0 R^1 = -\mathcal{T}_1 R^0, \\ &\epsilon^{-0}\text{term:} && \mathcal{T}_0 R^2 = -\mathcal{T}_2 R^0 - \mathcal{T}_1 R^1. \end{aligned}$$

These relations can be generically written as $\mathcal{T}_0 R^p = S^p$, where S^p depends only on $R^0,\dots,R^{p-1}$. We also remark that the operator $\mathcal{T}_0$ is a differential operator in the variables $\tau,\ y$, parametrized by t and x. Thus, we are dealing with cell equations with variable $(\tau,y)\in Y$ and periodic boundary conditions. At least formally, we aim at solving recursively these cell equations.

However, by integrating over Y, we realize that $\int_Y S\,\mathrm{d}(\tau,y) = 0$ is a necessary condition for the cell problem $\mathcal{T}_0 R = S$ to have a solution. We also note that the kernel of the adjoint operator $\mathcal{T}_0^\star = -\partial_\tau - u^0\cdot\nabla_y - \eta\Delta_y$ contains the constants. Actually it can be shown that $\mathcal{T}_0$ is a Fredholm operator of index 0. Hence, the condition $\int_Y S\,\mathrm{d}(\tau,y) = 0$ is also sufficient, and we will solve the cell equations by means of the Fredholm alternative. At this point a difficulty should be pointed out. When $\eta = 0$ the cell problem is no more of Fredholm type. The ergodic property that the null space of $\mathcal{T}_0^\star$ is spanned by the constants (with respect to y) is not sufficient to guarantee that the problem $\mathcal{T}_0 R = S$ has a solution under the condition $\int_Y S\,\mathrm{d}(\tau,y) = 0$. These facts will be detailed in the following subsection concerning rigorous proofs.

Now let us assume the following facts:

(A) For $(t,x)\in[0,\infty)\times\mathbb{R}^N$ fixed, 0 is a simple eigenvalue of $\mathcal{T}_0$ and the nullspace is spanned by a normalized function

$$\mathrm{Ker}(\mathcal{T}_0) = \mathrm{Span}\{\Theta\}, \qquad \int_Y \Theta\ \mathrm{d}(\tau,y) = 1.$$

(B) The cell problem $\mathcal{T}_0 R = S$ has a unique solution (up to elements of $\mathrm{Ker}(\mathcal{T}_0)$) under the necessary and sufficient condition $\int_Y S\,\mathrm{d}(\tau,y) = 0$.

From the ϵ^{-2} equation and (A), we infer that the leading term $R^0\in\mathrm{Ker}(\mathcal{T}_0)$ reads

$$R^0(t,x;\tau,y) = \rho(t,x)\ \Theta(t,x;\tau,y).$$

Next, the ϵ^{-1} equation becomes

$$\begin{aligned} \mathcal{T}_0 R^1 = -\mathcal{T}_1(R^0) = &-\Big(\mathrm{div}_x(u^0\ \Theta) + \mathrm{div}_y(u^1\ \Theta) - 2\eta\nabla_x\cdot\nabla_y\Theta\Big)\rho \\ &-\Big(\Theta\ u^0 - 2\eta\nabla_y\Theta\Big)\cdot\nabla_x\rho. \end{aligned}$$

Let us define $\chi(t,x;\tau,y) = (\chi_1,\dots,\chi_N)\in\mathbb{R}^N$ as the solution (with zero mean) of the auxiliary cell problem

$$\mathcal{T}_0\chi_j = \Theta\ {u^0}_j - 2\eta\partial_{y_j}\Theta \tag{3.1}$$

and $\kappa(t,x;\tau,y)$ the solution of

$$\mathcal{T}_0\kappa = -\mathrm{div}_x(u^0\ \Theta) - \mathrm{div}_y(u^1\ \Theta) + 2\eta\nabla_x\cdot\nabla_y\Theta.$$

Remark that, in view of (B), these problems can be solved under the condition

$$\int_Y u^0\ \Theta\,\mathrm{d}(\tau,y) = 0. \tag{3.2}$$

Therefore, the measure mentioned in (2.5) is nothing but $\mathrm{d}\mu(\tau,y) = \Theta\,\mathrm{d}(\tau,y)$. This condition is referred to as the "vanishing ballistic velocity condition." From now on we assume this relation holds true. Thus, the ϵ^{-1} equation is solved by

$$R^1(t,x;\tau,y) = -\chi(t,x;\tau,y)\cdot\nabla_x\rho(t,x) + \kappa(t,x;\tau,y)\rho(t,x) + S^1(t,x)\Theta(t,x;\tau,y),$$

where S^1 is an arbitrary function of (t,x). We will see that the choice of S^1 is irrelevant to obtaining the equation on R^0. As usual in two-scale asymptotics, the determination of S^1 is only necessary to obtain the equation for R^1. Now, let us look at the ϵ^0 term, which provides the equation satisfied by ρ. The solvability condition yields

$$\int_Y \mathcal{T}_2R^0\,\mathrm{d}(\tau,y) + \int_Y \mathcal{T}_1R^1\,\mathrm{d}(\tau,y) = 0.$$

Besides, we have

$$\begin{aligned}
\int_Y \mathcal{T}_2R^0\,\mathrm{d}(\tau,y) &= \partial_t\rho(t,x) + \mathrm{div}_x\left(\int_Y u^1\ \Theta(t,x;\tau,y)\,\mathrm{d}(\tau,y)\ \rho(t,x)\right) - \Delta_x\rho(t,x),\\
\int_Y \mathcal{T}_1R^1\,\mathrm{d}(\tau,y) &= \mathrm{div}_x\left(\int_Y u^0R^1\,\mathrm{d}(\tau,y)\right)\\
&= -\,\mathrm{div}_x\left(\int_Y u^0\otimes\chi(t,x;\tau,y)\,\mathrm{d}(\tau,y)\ \cdot\nabla_x\rho(t,x)\right)\\
&\quad + \mathrm{div}_x\left(\int_Y u^0\kappa(t,x;\tau,y)\,\mathrm{d}(\tau,y)\,\rho(t,x)\right).
\end{aligned}$$

Let us define the turbulent diffusion matrix D by

$$D(t,x) = \int_Y u^0(t,x;\tau,y)\otimes\chi(t,x;\tau,y)\,\mathrm{d}(\tau,y) \tag{3.3}$$

and the effective drift term by

$$v(t,x) = \int_Y \big(u^1\ \Theta(t,x;\tau,y) + u^0\ \kappa(t,x;\tau,y)\big)\,\mathrm{d}(\tau,y). \tag{3.4}$$

We are finally led to the following effective equation for $\rho(t,x) = \lim_{\epsilon\to0}\rho^\epsilon(t,x)$:

$$\partial_t\rho(t,x) + \mathrm{div}_x\big(\rho\ v\big)(t,x) - \mathrm{div}_x\left((D+\eta I_N)\cdot\nabla_x\rho\right)(t,x) = 0.$$

The matrix I_N is the $N\times N$ identity matrix. Therefore, the homogenization procedure induces an effective diffusion which is the sum of the original diffusion ηI with the matrix D.

Remark 4. We might wonder if the effects of the turbulent velocity field are to increase the molecular diffusivity and if the (symmetric part of the) matrix D is nonnegative. This is not true in general. Actually, considering potential flow, we will see that D is negative. This fact has been already noticed by Avellaneda and Vergassola [5]. However, we will check that $\eta I_N + D$ remains nonnegative.

Also the above asymptotic is based on the vanishing ballistic velocity assumption (3.2). The problem of what happens when (3.2) does not hold is addressed in subsection 4.5.

4. Rigorous results.

4.1. Functional preliminaries. First, let us introduce some functional spaces.

Let $\Omega \subset \mathbb{R}^D$, and let $p,\ q \in \mathbb{N}$. The set $C^p_\#(\Omega \times \mathbb{R}^{N+1})$ is the set of p-times continuously differentiable functions on $\Omega \times \mathbb{R}^{N+1}$ which are Y-periodic with respect to the last variable. Similarly, $C^p_{b,\#}(\Omega \times \mathbb{R}^{N+1})$ and $C^p_{c,\#}(\Omega \times \mathbb{R}^{N+1})$ are the subspaces of functions having bounded derivatives up to order p, and being supported in $K \times \mathbb{R}^{N+1}$ for some compact subset K of Ω, respectively. We also define anisotropic spaces $C^{[p;q,\alpha]}_\#([0,\infty) \times \mathbb{R}^N \times \mathbb{R}^{N+1})$ to be the spaces of continuous functions $u = u(t,x,\tau,y)$ such that

- $(\partial_t)^r(\partial_x)^s u$ exist and are continuous for $2r + s \le p$, where $(\partial_x)^s$ denotes any partial derivative of order s with respect to $x \in \mathbb{R}^N$, and $(\partial_t)^r$ stands for the rth derivative with respect to $t \ge 0$;
- $(\partial_\tau)^r(\partial_y)^s u$ exist and are continuous for $2r + s \le q$, where $(\partial_y)^s$ denotes any partial derivative of order s with respect to $y \in \mathbb{R}^N$, and $(\partial_\tau)^r$ stands for the rth derivative with respect to $\tau \in \mathbb{R}$;
- $(\partial_\tau)^r(\partial_y)^s u$ for $2r+s = q$ are Hölder continuous with exponent α with respect to y and with exponent $\alpha/2$ with respect to τ.

Finally, we also need $C^{[p;q,\alpha]}_{c,\#}([0,\infty)\times\Omega\times\mathbb{R}^N)$, the subspaces of functions with compact support with respect to the first two variables.

The formal limit obtained in the previous section can be rigorously justified by using the double-scale techniques as developed by Nguetseng [30] and Allaire [1]. This technique is quite equivalent to the method of oscillating test functions of Tartar [11, 31] and Evans [14, 15] and can be seen as a systematic way (but restricted to periodic situations) to choose the "good" test functions. In the present problem there is no a priori L^2-estimates, except if the u^i's are divergence free. Then the natural framework is an L^1-setting. We will adapt the results of [1, 30] to this setting. Now let us introduce some definitions and basic results about the family of parametrized measures.

DEFINITION 4.1. *Let I be an interval of $\mathbb{R}$. A family $\{\mu(t);\ t \in I\}$ of Radon measures on $\mathbb{R}^N$ is said to be vaguely continuous if and only if*

$$\forall \varphi \in C^0_c(\mathbb{R}^N), \qquad t \longmapsto \int_{\mathbb{R}^N} \varphi(x)\ \mu(t,x)\,\mathrm{d}x \quad \text{is a continuous function on } I.$$

DEFINITION 4.2. *A sequence $\{\mu_n(t);\ t \in I,\ n \in \mathbb{N}\}$ is said to be equibounded and vaguely equicontinuous on I if and only if*

(i) *there exists $M > 0$ such that*

$$\sup_{t\in I,\, n\in\mathbb{N}} |\mu_n|(t,\mathbb{R}^N) \le M;$$

(ii) *for any $\varphi \in C^0_c(\mathbb{R}^N)$, the sequence of functions $\left(t \longmapsto \int_{\mathbb{R}^N} \varphi(x)\ \mu_n(t,x)\,\mathrm{d}x\right)_{n\in\mathbb{N}}$ is equicontinuous on I.*

We recall the following classical compactness result.

PROPOSITION 4.3. *Let I be an interval of $\mathbb{R}$. Let $\big(\mu_n(t)\big)_{n\in\mathbb{N}}$ be a sequence of Radon measures on $\mathbb{R}^N$, equibounded and vaguely equicontinuous on I.*

Then there exist a measure $\mu(t)$ vaguely continuous on I and a subsequence $\big(\mu_{n_k}(t)\big)_{k\in\mathbb{N}}$ such that

$$\forall\varphi\in C_c^0(I\times\mathbb{R}^N),\quad \int_{\mathbb{R}^N}\varphi(t,x)\ \mu_{n_k}(t,x)\,\mathrm{d}x\xrightarrow[k\to\infty]{}\int_{\mathbb{R}^N}\varphi(t,x)\ \mu(t,x)\,\mathrm{d}x,$$

uniformly with respect to $t\in I$. We say that sequence $\big(\mu_{n_k}(t)\big)_{k\in\mathbb{N}}$ converges vaguely to $\mu(t)$ locally, uniformly on I.

Then we also have the existence of a double-scale limit, in the spirit of Allaire [1] and Nguetseng [30].

PROPOSITION 4.4. *Let $\big(\epsilon_n\big)_{n\in\mathbb{N}}$ be a sequence of positive numbers converging to 0. Let $\big(\mu_n(t)\big)_{n\in\mathbb{N}}$ be a sequence of measures on $\mathbb{R}^N$, equibounded on an interval $I\subset\mathbb{R}$. Then there exist a subsequence $\big(\mu_{n_k}(t)\big)_{k\in\mathbb{N}}$ and a measure M on $I\times\mathbb{R}\times\mathbb{R}^N\times Y$ such that for any $\varphi\in C^0_{c,\#}(I\times\mathbb{R}^N\times\mathbb{R}^{N+1})$, we have*

$$\int_I\int_{\mathbb{R}^N}\varphi(t,x;t/\epsilon_{n_k}^2,x/\epsilon_{n_k})\ \mu_{n_k}(t,x)\,\mathrm{d}x\,\mathrm{d}t \xrightarrow[k\to\infty]{}\int_I\int_{\mathbb{R}^N}\int_Y\varphi(t,x;\tau,y)\ M(t,x;\tau,y)\,\mathrm{d}(\tau,y)\,\mathrm{d}x\,\mathrm{d}t.$$

We say that the measure M is the double-scale limit of the sequence $\big(\mu_{n_k}(t)\big)_{k\in\mathbb{N}}$.

This proposition is nothing but a consequence of the Banach–Alaoglu theorem applied to the sequence of measures M_n defined by

$$\int_I\int_{\mathbb{R}^N}\int_Y\varphi(t,x;\tau,y)\,M_n(t,x;\tau,y)\,\mathrm{d}(\tau,y)\,\mathrm{d}x\,\mathrm{d}t:=\int_I\int_{\mathbb{R}^N}\varphi(t,x;t/\epsilon_n^2,x/\epsilon_n)\,\mu_n(t,x)\,\mathrm{d}x\,\mathrm{d}t.$$

The double-scale limit captures the periodic oscillations of μ_n with frequency $1/\epsilon_n$ in x and $1/\epsilon_n^2$ in t. By using test functions of the form $\varphi(t,x;\tau,y)=\psi(t,x)$, we also obtain immediately that the vague limit $\mu(t)$ of $\mu_{n_k}(t)\,\mathrm{d}t$ is given by the marginal

$$\mu(t,x)=\int_Y M(t,x;\tau,y)\,\mathrm{d}(\tau,y)\in\mathcal{M}^1(I\times\mathbb{R}^N).$$

4.2. Cell problems. Next, we are concerned with cell problems. In particular we discuss properties (A) and (B) stated in the previous section.

PROPOSITION 4.5. *Let $u^0\in C_\#^{[2;1,\alpha]}([0,\infty)\times\mathbb{R}^N\times\mathbb{R}^{N+1})$. Then the following assertions hold:*

(i) *There exists a unique function $\Theta\in C_\#^{[2;2,\alpha]}([0,\infty)\times\mathbb{R}^N\times\mathbb{R}^{N+1})$ such that*

$$\begin{cases}\mathcal{T}_0(\Theta)=\big(\partial_\tau\Theta+\mathrm{div}_y(u^0\ \Theta)-\eta\ \Delta_y\Theta\big)(t,x;\tau,y)=0\\ \qquad\forall(t,x;\tau,y)\in[0,\infty)\times\mathbb{R}^N\times\mathbb{R}^N,\\ \displaystyle\int_Y\Theta(t,x;\tau,y)\,\mathrm{d}(\tau,y)=1\qquad\forall(t,x)\in[0,\infty)\times\mathbb{R}^N.\end{cases}$$

Furthermore, we have $\Theta(t,x;\tau,y)>0$ for all $(t,x;\tau,y)$.

(ii) *For any* $S \in C_{\#}^{[p;0,\alpha]}([0,\infty)\times\mathbb{R}^N\times\mathbb{R}^{N+1})$ *with* $p=0,1,2$ *the cell problem*

$$\mathcal{T}_0(R) = \partial_\tau R + \operatorname{div}_y(u^0 R) - \eta\Delta_y R = S, \quad \int_Y R\,\mathrm{d}(\tau,y) = 0$$

has a unique solution $R \in C_{\#}^{[p;2,\alpha]}([0,\infty)\times\mathbb{R}^N\times\mathbb{R}^{N+1})$ *under the necessary and sufficient orthogonality condition*

$$\int_Y S(t,x;\tau,y)\,\mathrm{d}(\tau,y) = 0 \quad \forall (t,x)\in[0,\infty)\times\mathbb{R}^N. \tag{4.1}$$

(iii) *For any* $H \in C_{\#}^{[p;0,\alpha]}([0,\infty)\times\mathbb{R}^N\times\mathbb{R}^{N+1})$ *with* $p=0,1,2$ *the cell problem*

$$-\partial_\tau\Phi - u^0\cdot\nabla_y\Phi - \eta\Delta_y\Phi = H, \quad \int_Y \Phi\,\mathrm{d}(\tau,y) = 0$$

has a unique solution $\Phi \in C_{\#}^{[p;2,\alpha]}([0,\infty)\times\mathbb{R}^N\times\mathbb{R}^{N+1})$ *under the necessary and sufficient orthogonality condition*

$$\int_Y H(t,x;\tau,y)\ \Theta(t,x;\tau,y)\,\mathrm{d}(\tau,y) = 0 \quad \forall (t,x)\in[0,\infty)\times\mathbb{R}^N. \tag{4.2}$$

Proof. Let $(t,x)\in[0,\infty)\times\mathbb{R}^N$ be fixed. We denote by

$$\mathcal{T}_0^\star(\Phi) = -\partial_\tau\Phi - u^0\cdot\nabla_y\Phi - \eta\Delta_y\Phi$$

the adjoint operator of $\mathcal{T}_0$. Obviously, constants belong to the kernel of $\mathcal{T}_0^\star$. For $\lambda>0$ large enough we shall see that the resolvant $R_\lambda = (\lambda+\mathcal{T}_0^\star)^{-1}$ (resp., $S_\lambda = (\lambda+\mathcal{T}_0)^{-1}$) is well defined and we rewrite the problems $\mathcal{T}_0^\star(\Phi)=H$ (resp., $\mathcal{T}_0(R)=S$) as

$$(I-\lambda R_\lambda)\Phi = R_\lambda H \tag{4.3}$$

(resp., $(I-\lambda S_\lambda)\Phi = S_\lambda H$). We wish to conclude by applying the Fredholm alternative. We are thus led to investigate compactness and spectral properties of the resolvant operator R_λ (resp., S_λ).

Hence, we are interested in the problem

$$(\mathcal{T}_0^\star+\lambda)\Phi = H \in L^2_{\#}(Y) \tag{4.4}$$

for some $\lambda>0$ and domain $D(\mathcal{T}_0^\star)$, which will be made precise later on. Assuming existence in $H^1_{\#}(Y)$, uniqueness of the solution follows from the positivity of the energy functional

$$\begin{aligned} a_0(\Phi,\Phi) &= \int_Y (\mathcal{T}_0^\star+\lambda)\Phi\ \Phi\,\mathrm{d}(\tau,y) \\ &= \int_Y \left(\lambda + \frac{1}{2}\operatorname{div}_y u^0\right)\Phi^2\,\mathrm{d}(\tau,y) + \eta\int_Y |\nabla_y\Phi|^2\,\mathrm{d}(\tau,y), \end{aligned}$$

provided λ satisfies $\lambda > \max_{(\tau,y)\in Y}\max(0,-\frac{1}{2}\operatorname{div}_y u^0)$. From now on, we fix λ, which fulfills this relation. It remains to justify the existence of a solution of (4.4).

To this end, we introduce an elliptic regularization of $\mathcal{T}_0^\star$. Let ${\mathcal{T}_0^\star}_\mu$, $\mu>0$, be the operator

$${\mathcal{T}_0^\star}_\mu(\Phi) = -\mu\partial^2_{\tau,\tau}\Phi - \partial_\tau\Phi - u^0\cdot\nabla_y\Phi - \eta\Delta\Phi$$

with domain $D({\mathcal{T}_0^\star}_\mu) = H^2_\#(Y)$, the space of functions Y-periodic, belonging to $H^2_{\rm loc}(\mathbb{R}^{N+1})$. For $\lambda > \frac{1}{2}\max_{\tau,y}\{0, -{\rm div}_y u_0\}$, the associated energy functional

$$a_\mu(\Phi,\Phi) = \int_Y \Phi({\mathcal{T}_0^\star}_\mu(\Phi) + \lambda\Phi)\,{\rm d}(\tau,y)$$
$$= \mu\int_Y |\partial_\tau\Phi|^2\,{\rm d}(\tau,y) + \int_Y \left(\lambda + \frac{1}{2}{\rm div}_y(u^0)\right)\,\Phi^2\,{\rm d}(\tau,y) + \eta\int_Y |\nabla_y\Phi|^2\,{\rm d}(\tau,y)$$

is coercive on the space $H^1_\#(Y)$. A direct application of the Lax–Milgram theorem guarantees, for any data $H \in L^2_\#(Y)$, the existence-uniqueness of a solution $\Phi \in H^1_\#(Y)$ of the cell problem ${\mathcal{T}_0^\star}_\mu(\Phi_\mu) + \lambda\Phi_\mu = H$ in $H^1_\#(Y)$. Then $\mu\partial^2_{\tau,\tau}\Phi + \eta\Delta\Phi \in L^2_\#(Y)$, which implies that Φ_μ belongs to $H^2_\#(Y) = D({\mathcal{T}_0^\star}_\mu)$. We also have $\Phi_\mu \geq 0$ when $H \geq 0$ by the maximum principle. Then (4.4) will be solved by passing to the limit $\mu \to 0$.

By using the relation $a_\mu(\Phi_\mu,\Phi_\mu) = \int_Y H\ \Phi_\mu\,{\rm d}(\tau,y)$, we obtain that Φ_μ and $\nabla_y\Phi_\mu$ are uniformly bounded in $L^2_\#(Y)$. Multiplying the equation ${\mathcal{T}_0^\star}_\mu(\Phi_\mu) + \lambda\Phi_\mu = H$ by $\partial_\tau\Phi_\mu$ and integrating by parts give

$$\int_Y |\partial_\tau\Phi_\mu|^2\,{\rm d}(\tau,y) = -\int_Y u_0\cdot\nabla_y\Phi_\mu\partial_\tau\Phi_\mu\,{\rm d}(\tau,y) - \int_Y H\ \partial_\tau\Phi_\mu\,{\rm d}(\tau,y).$$

We deduce that $\partial_\tau\Phi_\mu$ is also uniformly bounded in $L^2_\#(Y)$. Hence, Φ_μ is bounded in $H^1_\#(Y)$. The cluster points of the sequence Φ_μ as $\mu \to 0$ are solutions in $H^1_\#(Y)$ of (4.4). Such a solution satisfies $\Delta_y\Phi \in L^2_\#(Y)$. Defining $D(\mathcal{T}_0^\star) = \{\Phi \in H^1_\#(Y),\ \Delta_y\Phi \in L^2_\#(Y)\}$, we have obtained the existence of a compact resolvant $R_\lambda = (\mathcal{T}_0^\star + \lambda)^{-1}$.

Then, in order to apply the Fredholm alternative for (4.3), it remains to determine the compatibility relation, and in particular the dimension of the eigenspace ${\rm Ker}(\mathcal{T}_0^\star) = {\rm Ker}(R_\lambda - 1/\lambda))$ (resp., ${\rm Ker}(\mathcal{T}_0) = {\rm Ker}(S_\lambda - 1/\lambda)$). We use the fact that R_λ preserves nonnegativity: for $H \geq 0$, $\Phi = R_\lambda(H) \geq 0$. The Krein–Rutman theorem (see [25]) applies and the spectral radius ρ of R_λ is an eigenvalue, associated to a nonnegative eigenfunction. Reasoning similarly for the adjoint operator, we obtain the existence of a function $\Theta \in H^1_\#(Y)$, $\Theta \geq 0$, verifying $S_\lambda\Theta = \rho\Theta$. However, we have remarked that $(1/\lambda, \mathbb{1})$ is an eigenpair for R_λ. Hence, we get

$$\rho\int_Y \Theta\,{\rm d}(\tau,y) = \int_Y \mathbb{1}\ S_\lambda(\Theta)\,{\rm d}(\tau,y) = \int_Y R_\lambda(\mathbb{1})\ \Theta\,{\rm d}(\tau,y) = \frac{1}{\lambda}\int_Y \Theta\,{\rm d}(\tau,y);$$

i.e., $\rho = 1/\lambda$ is the principal eigenvalue of R_λ, S_λ and Θ satisfies $\mathcal{T}_0(\Theta) = 0$.

Then we make use of regularity results and the maximum principle for parabolic equations. We first remark that the functions of $D(\mathcal{T}_0) = D(\mathcal{T}_0^\star)$ are continuous with respect to time τ with value in $L^2_\#([0,1]^N)$, and their first derivatives with respect to y belong to $L^2_\#(Y)$. With the notation of [27], such functions belong to $V_2^{1,0}$. Therefore, since the coefficients are smooth we can apply [27, Theorem 12.1, p. 223]. We deduce that $R_\lambda(H) = \Phi \in C^{[2,\alpha]}_\#(Y)$ ($H^{2+\alpha,1+\alpha/2}(Y)$ in the terminology of [27]) for $H \in C^{[0,\alpha]}_\#(Y)$. Similarly, we have $\Theta \in C^{[2,\alpha]}_\#(Y)$. Thus, the maximum principle applies (see [16, Theorem 5, p. 39]) and we obtain $\Theta > 0$, and $R_\lambda(H) = \Phi > 0$ for $H \geq 0$. Suppose now there exists a $\Phi \in {\rm Ker}(\mathcal{T}_0^\star)$ with $\int_Y \Phi\,{\rm d}(\tau,y) = 0$. Then Φ is also in $C^{[2,\alpha]}_\#(Y)$ and the positive and negative parts $\Phi_\pm$ are continuous functions nonidentically zero. By the maximum principle $R_\lambda\Phi_\pm > 0$. We deduce that $|\Phi| <$

$\lambda R_\lambda(|\Phi|)$. It would imply $\int_Y \Theta|\Phi| \,\mathrm{d}(\tau,y) = \lambda \int_Y \Theta R_\lambda(|\Phi|) \,\mathrm{d}(\tau,y) > \int_Y \Theta|\Phi| \,\mathrm{d}(\tau,y)$, a contradiction. Thus, the kernel of $\mathcal{T}_0^\star$ (resp., $\mathcal{T}_0$) is monodimensional and spanned by $\mathbb{1}$ (resp., $\Theta > 0$).

We conclude by applying the Fredholm alternative. The solution of $\mathcal{T}_0(R) = S$ exists if and only if S is orthogonal to the kernel of the adjoint that is $\int_Y S \,\mathrm{d}(\tau,y) = 0$. The solution is unique if we impose $\int_Y R \;\mathrm{d}(\tau,y) = 0$. The same is true for $\mathcal{T}_0^\star$: the equation $\mathcal{T}_0^\star(\Phi) = S$ has a solution if and only if $\int_Y S\,\Theta \;\mathrm{d}(\tau,y) = 0$ and the uniqueness is assured by the condition $\int_Y R \;\mathrm{d}(\tau,y) = 0$. By applying [27, Theorem 12.1, p. 223], we obtain that Θ, R, and Φ belong to $C_\#^{[2,\alpha]}(Y)$ if S, $H \in C_\#^{[0,\alpha]}(Y)$ for fixed (t,x).

We study now the dependency with respect to these parameters (t,x). Letting δ_h be a finite difference operator in the variables (t,x), we have

$$\mathcal{T}_0^\star(\delta_h(\Phi)) = -\delta_h(u^0)\cdot\nabla_y(\Phi) + \delta_h(H).$$

The right-hand side is bounded in $C_\#^{[0,\alpha]}(Y)$ uniformly with respect to h. Then the sequence $\big(\delta_h(\Phi)\big)_{h>0}$ is bounded in $C_\#^{2,\alpha}(\mathbb{R}^N)$ uniformly with respect to h. This implies that $(t,x) \longmapsto \Phi(t,x,.)$ is differentiable with respect to (t,x) with values in $C_\#^{[2,\alpha]}(Y)$. In particular $\Phi \in C^0([0,\infty)\times\mathbb{R}^N; C_\#^{[2,\alpha]}(Y))$. If $\partial\Phi$ denotes a derivative with respect to t or x, we have, letting $h \to 0$,

$$\mathcal{T}_0^\star(\partial\Phi) = -\partial u^0\cdot\nabla_y(\Phi) + \partial H.$$

Since the right-hand side is continuous with respect to the parameter (t,x), it yields the continuity of $\partial\Phi$. We obtain the same result for the derivatives up to order 2 by repeating the argument once. The regularity of the solution R and of the eigenfunction Θ are obtained in the same way. ∎

Remark 5. Of course, when the velocity field u^0 does not oscillate with respect to time (i.e., $u^{0,\epsilon} = u^0(t,x;x/\epsilon)$), the corresponding solution of the cell problem Θ does not depend on the fast time variable τ.

4.3. Effective coefficients. Our proof of the homogenization result is based on the expansion of the dual equation. Therefore, we obtain the diffusion matrix and the drift velocity, given by (3.3) and (3.4), respectively, by means of dual formulae. Let us introduce $\chi^\star$, solution of the dual problem of (3.1), namely,

$$\mathcal{T}_0^\star(\chi^\star) = -\partial_\tau\chi^\star - u^0\cdot\nabla_y\chi^\star - \eta\Delta_y\chi^\star = u^0, \qquad \int_Y \chi^\star \,\mathrm{d}(\tau,y) = 0. \tag{4.5}$$

We can express D and v with $\chi^\star$ as follows. On the one hand, we have

$$\begin{aligned} D &= \int_Y u^0\otimes\chi \,\mathrm{d}(\tau,y) = \int_Y \mathcal{T}_0^\star(\chi^\star)\otimes\chi \,\mathrm{d}(\tau,y) \\ &= \int_Y \chi^\star\otimes\mathcal{T}_0(\chi) \,\mathrm{d}(\tau,y) = \int_Y \chi^\star\otimes(\Theta u^0 - 2\eta\nabla_y\Theta) \,\mathrm{d}(\tau,y), \end{aligned} \tag{4.6}$$

and on the other hand,

$$\begin{aligned} v - \int_Y u^1\Theta \,\mathrm{d}(\tau,y) &= \int_Y u^0\;\kappa \,\mathrm{d}(\tau,y) = \int_Y \mathcal{T}_0^\star(\chi^\star)\;\kappa \,\mathrm{d}(\tau,y) = \int_Y \chi^\star\mathcal{T}_0(\kappa) \,\mathrm{d}(\tau,y) \\ &= \int_Y \chi^\star\Big(- \mathrm{div}_x(\Theta u^0) - \mathrm{div}_y(\Theta u_1) + 2\eta\nabla_x\cdot\nabla_y\Theta\Big) \,\mathrm{d}(\tau,y). \end{aligned} \tag{4.7}$$

Of course, the crucial question relies on the positiveness of the diffusion coefficient. Actually, we shall see that (the symmetric part of) D can be nonpositive, while the sum $\eta+D$ remains nonnegative. To this end, it is particularly illuminating to consider the following fundamental examples.

The divergence free case. Suppose that $\mathrm{div}_y(u^0)=0$. In this case, $\Theta=1$ and the vanishing ballistic velocity condition is nothing but

$$\int_Y u^0 \,\mathrm{d}(\tau,y)=0.$$

The function χ satisfies

$$\mathcal{T}_0(\chi)=u^0\Theta=u^0.$$

Then, for $\xi\in\mathbb{R}^N\backslash\{0\}$, we have

$$\begin{aligned} D\xi\cdot\xi &= \int_Y u^0\cdot\xi\ \chi\cdot\xi\,\mathrm{d}(\tau,y)=\int_Y \mathcal{T}_0(\chi\cdot\xi)\ \chi\cdot\xi\ \ \mathrm{d}(\tau,y)\\ &= \eta\int_Y |\nabla_y(\chi\cdot\xi)|^2\ \ \mathrm{d}(\tau,y)\geq 0. \end{aligned}$$

The potential case. Suppose that u_0 does not depend on the variable τ and is of the form $u^0(t,x;y)=\nabla_y V(t,x;y)$ for some potential function $V:\mathbb{R}^+\times\mathbb{R}^N\times\mathbb{R}^N\to\mathbb{R}$, $[0,1]^N$-periodic with respect to the last variable. Then the operator $\mathcal{T}_0$ recasts in the symmetric form

$$\mathcal{T}_0(R)=\partial_\tau R-\eta\ \mathrm{div}_y\Big(e^{V/\eta}\ \nabla_y(e^{-V/\eta}\ R)\Big)$$

so that

$$\int_Y \mathcal{T}_0(R)R\ e^{-V/\eta}\,\mathrm{d}(\tau,y)=\eta\int_Y e^{V/\eta}\ |\nabla_y(Re^{-V/\eta})|^2\,\mathrm{d}(\tau,y).$$

We deduce that

$$\mathrm{Ker}(\mathcal{T}_0)=\mathrm{Span}\{e^{V/\eta}\},\qquad \Theta=Ze^{V/\eta},$$

$Z=(\int_Y e^{V/\eta}\,\mathrm{d}(\tau,y))^{-1}$ being a normalization constant. Notice that the vanishing ballistic velocity condition is fulfilled since

$$\int_Y u^0\ \Theta\ \ \mathrm{d}(\tau,y)=\eta Z\int_Y \nabla_y(e^{V/\eta})\ \ \mathrm{d}(\tau,y)=0.$$

Finally, the equation for χ reads

$$\mathcal{T}_0(\chi)=u^0\Theta-2\eta\nabla_y\Theta=-Z\nabla_y V\ e^{V/\eta}=-u^0\Theta.$$

(Notice the change of sign in the right-hand side in comparison to the divergence-free case.) Then, for $\xi\in\mathbb{R}^N\backslash\{0\}$, we have

$$\begin{aligned} D\xi\cdot\xi &= \int_Y u^0\cdot\xi\ \chi\cdot\xi\ \ \mathrm{d}(\tau,y)=-\int_Y(-u^0\Theta\cdot\xi)\ \chi\cdot\xi\ \Theta^{-1}\ \ \mathrm{d}(\tau,y)\\ &= -\int_Y \mathcal{T}_0(\chi\cdot\xi)\ \chi\cdot\xi\ \Theta^{-1}\,\mathrm{d}(\tau,y)=-\eta\int_Y \Theta\ |\nabla_y(\chi\cdot\xi\Theta^{-1})|^2\,\mathrm{d}(\tau,y)\leq 0, \end{aligned}$$

which reveals an antidiffusive effect. However, Avellaneda and Vergassola [5] pointed out that, in this potential case, the total diffusivity remains nonnegative, $(\eta+D)\xi\cdot\xi \geq 0$. Their argument can be generalized, as shown by the following statement.

PROPOSITION 4.6. *Let D be the matrix defined in* (3.3). *Then, for any $\xi \in \mathbb{R}^N\backslash\{0\}$, one has*

$$(\eta + D)\xi \cdot \xi > 0.$$

Proof. The proof starts with the dissipativity property of $\mathcal{T}_0^\star$, using an inner product with weight Θ. It was pointed out to us by Collet [10] that this property is general and applies for any elliptic operator of order 2 and more generally for any operator coming from Markovian processes (see Kubo [26]). We have

$$\begin{aligned}
\int_Y \mathcal{T}_0^\star(\Phi)\ \Phi\ \Theta\,\mathrm{d}(\tau,y) &= \int_Y (-\partial_\tau\Phi - u^0\cdot\nabla_y\Phi - \eta\Delta_y\Phi)\Phi\ \Theta\,\mathrm{d}(\tau,y)\\
&= -\int_Y \left(\partial_\tau\left(\frac{\Phi^2}{2}\right) + u^0\cdot\nabla_y\left(\frac{\Phi^2}{2}\right)\right)\ \Theta\ \mathrm{d}(\tau,y)\\
&\quad + \eta\int_Y \nabla_y\left(\frac{\Phi^2}{2}\right)\cdot\nabla_y\Theta\ \mathrm{d}(\tau,y) + \eta\int_Y |\nabla_y\Phi|^2\ \Theta\ \mathrm{d}(\tau,y)\\
&= \int_Y \frac{\Phi^2}{2}\ \mathcal{T}_0\Theta\,\mathrm{d}(\tau,y) + \eta\int_Y |\nabla_y\Phi|^2\ \Theta\,\mathrm{d}(\tau,y)\\
&= \eta\int_Y |\nabla_y\Phi|^2\ \Theta\,\mathrm{d}(\tau,y) \geq 0.
\end{aligned}$$

Then let us use the expression of D obtained in (4.6). We get

$$\begin{aligned}
D\xi\cdot\xi &= \int_Y \chi^\star\cdot\xi\ u^0\cdot\xi\ \Theta\ \mathrm{d}(\tau,y) - 2\eta\int_Y \chi^\star\cdot\xi\ \nabla_y\Theta\cdot\xi\ \mathrm{d}(\tau,y)\\
&= \int_Y \chi^\star\cdot\xi\ \mathcal{T}_0^\star(\chi^\star\cdot\xi)\ \Theta\ \mathrm{d}(\tau,y) + 2\eta\int_Y \Theta\ \xi\cdot\nabla_y\big(\chi^\star\cdot\xi\big)\,\mathrm{d}(\tau,y)\\
&= \eta\int_Y |\nabla_y(\chi^\star\cdot\xi)|^2\ \Theta\,\mathrm{d}(\tau,y) + 2\eta\int_Y \Theta\ \xi\cdot\nabla_y\big(\chi^\star\cdot\xi\big)\,\mathrm{d}(\tau,y).
\end{aligned}$$

Thus, we obtain

$$\begin{aligned}
(\eta + D)\xi\cdot\xi &= \eta\int_Y |\xi + \nabla_y(\chi^\star\cdot\xi)|^2\Theta\,\mathrm{d}(\tau,y) \geq 0\\
&= \eta\left(\int_Y (I + \nabla_y\chi^\star)\ (I + \nabla_y\chi^\star)^T\ \Theta\,\mathrm{d}(\tau,y)\right)\xi\cdot\xi,
\end{aligned}$$

since $\int_Y \Theta\,\mathrm{d}(\tau,y) = 1$. The above integral vanishes when for a.e. $(\tau,y)\in Y$, $\nabla_y(\chi^\star\cdot\xi) = -\xi$ does not depend on τ, y. By using the periodicity, it follows that $0 = \int_Y \nabla_y(\chi^\star\cdot\xi)\,\mathrm{d}(\tau,y) = -\xi|Y|$; thus $\xi = 0$. We conclude that the symmetric part of D is positive definite. ∎

Remark 6. It should be pointed out that it is crucial to take into account a positive diffusivity coefficient $\eta > 0$, even for the simple divergence-free case. In this case, we have $\Theta = 1$. Suppose there exists a vector-valued function χ solution of $\partial_\tau\chi + u^0\cdot\nabla_y\chi = u^0$. Then if $\eta = 0$, it is readily checked that $D\xi\cdot\xi = 0$, i.e., the "effective diffusion matrix" vanishes! Turbulent diffusivity strongly depends on the molecular diffusivity.

We have two kinds of results depending on whether or not u^0 satisfies the orthogonality relation $\int_Y u^0\ \Theta\,\mathrm{d}(\tau,y)=0$. We have seen that if u^0 is divergence free, this condition is nothing but the fact that u^0 has a zero mean value and that the condition is always satisfied in the potential case. When the orthogonality relation is satisfied, the result is complete. When it is not, the result is not complete in the sense that we do not know if the effective equations we obtain give rise to a well-posed problem.

4.4. Vanishing ballistic velocity. Let us first investigate the case of a vanishing ballistic velocity.

THEOREM 4.7. *Let $\eta>0$. Let $u^0,\ u^1\in C^{[2;1,\alpha]}_{\#}([0,\infty)\times\mathbb{R}^N\times\mathbb{R}^{N+1})$. Let Θ be defined as in Proposition* 4.5 *and suppose that*

$$\int_Y u^0(t,x;\tau,y)\ \Theta(t,x;\tau,y)\ \mathrm{d}(\tau,y)=0\qquad\forall(t,x)\in[0,\infty)\times\mathbb{R}^N \tag{4.8}$$

holds. Let $\left(\rho_I^\epsilon\right)_{\epsilon>0}$ be a bounded sequence of nonnegative measures on $\mathbb{R}^N$. We suppose that $\left(\rho_I^\epsilon\right)_{\epsilon>0}$ converges vaguely to ρ_I. Let $\rho^\epsilon(t)$ be the solutions of the advection diffusion problem (2.1), (2.2) *with*

$$u^\epsilon(t,x)=\frac{1}{\epsilon}\left(u^0\left(t,x,\frac{t}{\epsilon^2},\frac{x}{\epsilon}\right)+\epsilon u^1\left(t,x,\frac{t}{\epsilon^2},\frac{x}{\epsilon}\right)\right).$$

Then, up to the extraction of a subsequence, $\rho^\epsilon(t)\ge 0$ converges vaguely to $\rho(t)\ge 0$ locally and uniformly, and the limit $\rho(t)$ is a solution of

$$\partial_t\rho(t,x)+\mathrm{div}_x\big(\rho\ v\big)(t,x)-\mathrm{div}_x\left((D+\eta I_N)\cdot\nabla_x\rho\right)(t,x)=0 \tag{4.9}$$

in the sense of distributions with the Cauchy data $\rho(0)=\rho_I$. The coefficients are given by

$$\begin{aligned}
v(t,x)&=\int_Y u^1\ \Theta(t,x;\tau,y)\ \mathrm{d}(\tau,y)\\
&\quad-\int_Y\chi^\star\Big(\mathrm{div}_x(\Theta\ u^0)+\mathrm{div}_y(\Theta\ u^1)-2\eta\nabla_x\cdot\nabla_y\Theta\Big)(t,x;\tau,y)\ \mathrm{d}(\tau,y),\\
D(t,x)&=\int_Y\chi^\star\otimes(\Theta\ u^0-2\eta\nabla_y\Theta)(t,x;\tau,y)\ \mathrm{d}(\tau,y),
\end{aligned}$$

and $\chi^\star$ is the solution of

$$-\partial_\tau\chi^\star-u^0\cdot\nabla_y\chi^\star-\eta\Delta_y\chi^\star=u^0,\quad\int_Y\chi^\star\,\mathrm{d}(\tau,y)=0.$$

Remark 7. The statement can be strengthened when the velocity fields satisfy the divergence-free condition $\mathrm{div}_{x,y}u^{0,1}=0$. In this case, recall that $\Theta=1$. Assume the initial condition is converging in $L^2(\mathbb{R}^N)$: $\rho_I^\epsilon\to\rho_I$. Then one has L^2-estimates on both ρ^ϵ and $\nabla_x\rho^\epsilon$. Consequently, one has convergence of the solutions ρ^ϵ to ρ in $L^2(0,T;H^1(\mathbb{R}^N))$ weakly, and in $C^0([0,T];L^2(\mathbb{R}^N)\text{-}weak)$, and for a.e. $t\in[0,T]$, $\rho^\epsilon(t)\to\rho(t)$ strongly in $L^2(K)$ for any compact set $K\subset\mathbb{R}^N$.

Proof. By using standard results for parabolic equations, the solutions ρ^ϵ satisfy

$$\begin{cases}\forall(t,x)\in[0,\infty)\times\mathbb{R}^N,\ \rho^\epsilon(t,x)\ge 0,\\ \forall t\in[0,\infty),\ \displaystyle\int_{\mathbb{R}^N}\rho^\epsilon(t,x)\,\mathrm{d}x\le\int_{\mathbb{R}^N}\rho_I^\epsilon(x)\,\mathrm{d}x\le M,\end{cases} \tag{4.10}$$

where $M = \sup_{\epsilon>0}\ \rho_I^\epsilon(\mathbb{R}^N)$ is finite by assumption. Note that the conservation of mass $\int_{\mathbb{R}^N} \rho^\epsilon(t,x)\,\mathrm{d}x = \int_{\mathbb{R}^N} \rho_I^\epsilon(x)\,\mathrm{d}x$, which can be expected from formal arguments, is not obvious at all. To obtain this conservation law, an additional bound on $|u^\epsilon|$ is necessary. For instance, we have to assume $|u^0(t,x;\tau,y)| + |u^1(t,x;\tau,y)| \le C|x|$ for some constant $C > 0$. We will not address this problem in this work.

We cannot immediately use Proposition 4.3 for $\rho^\epsilon(t)$ since the equicontinuity with respect to time is far from clear: there is no obvious bound for $\partial_t\rho^\epsilon$ because of the singular term $\frac{1}{\epsilon}\mathrm{div}_x(u^0\rho^\epsilon)$. Actually we will use crucially the condition (4.8) to prove that the sequence $\rho^\epsilon(t)$ is equicontinuous. The idea is to recover the formal asymptotics of the previous subsection, but this time working on the test functions for the adjoint equation. Hence, let us define for any $\varphi \in C^2_{\#}(\mathbb{R}^N \times \mathbb{R}^{N+1})$ the following operators:

$$\begin{aligned}
\mathcal{T}_0^\star(\varphi)(t,x;\tau,y) &= -\partial_\tau\varphi(x;\tau,y) - u^0(t,x;\tau,y)\cdot\nabla_y\varphi(x;\tau,y) - \eta\Delta_y\varphi(x;\tau,y),\\
\mathcal{T}_1^\star(\varphi)(t,x;\tau,y) &= -u^1(t,x;\tau,y)\cdot\nabla_y\varphi(x;\tau,y) - u^0(t,x;\tau,y)\cdot\nabla_x\varphi(x;\tau,y)\\
&\quad -2\eta\nabla_x\cdot\nabla_y\varphi(x;\tau,y),\\
\mathcal{T}_2^\star(\varphi)(t,x;\tau,y) &= -u^1(t,x;\tau,y)\cdot\nabla_x\varphi(x;\tau,y) - \eta\Delta_x\varphi(x;\tau,y).
\end{aligned}$$

Step 1. Let $\varphi \in C^2_{c,\#}([0,\infty)\times\mathbb{R}^N\times\mathbb{R}^{N+1})$. By multiplying the equation by $\epsilon^2\,\varphi(t,x;t/\epsilon^2,x/\epsilon)$, we get

(4.11)
$$\int_{\mathbb{R}^N} \mathcal{T}_0^\star(\varphi)(t,x;t/\epsilon^2,x/\epsilon)\ \rho^\epsilon(t,x)\,\mathrm{d}x = -\epsilon\int_{\mathbb{R}^N}(\mathcal{T}_1^\star+\epsilon\mathcal{T}_2^\star)(\varphi)(t,x;t/\epsilon^2,x/\epsilon)\ \rho^\epsilon(t,x)\,\mathrm{d}x$$
$$-\epsilon^2\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^N}\varphi(t,x;t/\epsilon^2,x/\epsilon)\ \rho^\epsilon(t,x)\,\mathrm{d}x + \epsilon^2\int_{\mathbb{R}^N}\partial_t\varphi(t,x;t/\epsilon^2,x/\epsilon)\ \rho^\epsilon(t,x)\,\mathrm{d}x.$$

Then we obtain

$$\lim_{\epsilon\to0}\left(\int_{\mathbb{R}^N}\mathcal{T}_0^\star(\varphi)(t,x;t/\epsilon^2,x/\epsilon)\ \rho^\epsilon(t,x)\,\mathrm{d}x\right) = 0 \qquad \text{in } \mathcal{D}'(0,\infty) \tag{4.12}$$

for any $\varphi \in C^2_{c,\#}([0,\infty)\times\mathbb{R}^N\times\mathbb{R}^{N+1})$.

Step 2. In order to get rid of the leading term in the dual equation (4.11), we choose the test function of the form $\varphi(t,x;t/\epsilon^2,x/\epsilon) = \psi(x) + \epsilon\ \phi(t,x;t/\epsilon^2,x/\epsilon)$. We obtain

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^N}(\psi+\epsilon\ \phi)\,(t,x;t/\epsilon^2,x/\epsilon)\rho^\epsilon(t,x)\,\mathrm{d}x\\
&\qquad = -\frac{1}{\epsilon}\int_{\mathbb{R}^N}\left(\mathcal{T}_1^\star(\psi)+\mathcal{T}_0^\star(\phi)\right)(t,x;t/\epsilon^2,x/\epsilon)\ \rho^\epsilon(t,x)\,\mathrm{d}x\\
&\qquad\quad -\int_{\mathbb{R}^N}\left(\mathcal{T}_2^\star(\psi)+\mathcal{T}_1^\star(\phi)\right)(t,x;t/\epsilon^2,x/\epsilon)\ \rho^\epsilon(t,x)\,\mathrm{d}x\\
&\qquad\quad +\epsilon\int_{\mathbb{R}^N}\left(\partial_t\phi-\mathcal{T}_2^\star(\phi)\right)(t,x;t/\epsilon^2,x/\epsilon)\ \rho^\epsilon(t,x)\,\mathrm{d}x.
\end{aligned} \tag{4.13}$$

Since ψ does not depend on the fast variables (τ,y), we remark that $\mathcal{T}_1^\star(\psi) = -u^0\cdot\nabla_x\psi$. Consequently, multiplying by ϵ, passing to the limit, and using (4.12) yield

$$\begin{aligned}
&\lim_{\epsilon\to0}\left(\int_{\mathbb{R}^N}\left(\mathcal{T}_1^\star(\psi)+\mathcal{T}_0^\star(\phi)(t,x;t/\epsilon^2,x/\epsilon)\right)\ \rho^\epsilon(t,x)\,\mathrm{d}x\right) = 0\\
&\qquad = \lim_{\epsilon\to0}\left(\int_{\mathbb{R}^N}u^0(t,x;t/\epsilon^2,x/\epsilon)\cdot\nabla_x\psi(x)\ \rho^\epsilon(t,x)\,\mathrm{d}x\right).
\end{aligned} \tag{4.14}$$

Up to now, condition (4.8) has not been used. We shall use the condition to write $u^0 \cdot \nabla_x \psi$ as $\mathcal{T}_0^\star(\phi)$ for a convenient choice of the function ϕ. Therefore, we realize that (4.14) is already contained in (4.12).

In order to go further we have to get rid of the term of order $1/\epsilon$ in (4.13). Let us introduce the vector-valued function $\chi^\star(t,x;\tau,y) \in \left(C_\#^{[2;2,\alpha]}([0,\infty)\times\mathbb{R}^N\times\mathbb{R}^{N+1})\right)^N$ as the solution of the adjoint cell problem

$$\mathcal{T}_0^\star(\chi^\star) = -\partial_\tau - u^0\cdot\nabla_y\chi^\star - \eta\Delta_y\chi^\star = u^0, \qquad \int_Y \chi^\star \,\mathrm{d}(\tau,y) = 0.$$

The function $\chi^\star$ is well defined thanks to Proposition 4.5 and to condition (4.8). We choose the test function ϕ depending on ψ as follows:

$$\phi(t,x;\tau,y) = \chi^\star(t,x;\tau,y)\cdot\nabla_x\psi(x).$$

Assuming $\psi \in C_c^3(\mathbb{R}^N)$, we have $\phi \in C_{c,\#}^{[2;2,\alpha]}([0,T]\times\mathbb{R}^N\times\mathbb{R}^{N+1})$ for any $T>0$. The crucial fact is that $\mathcal{T}_0^\star(\phi) = \mathcal{T}_0^\star(\chi^\star)\cdot\nabla_x\psi = u^0\cdot\nabla_x\psi = -\mathcal{T}_1^\star(\psi)$. Consequently, we recover (4.12) from (4.14). Furthermore, the $\frac{1}{\epsilon}$ term in (4.13) vanishes. We deduce the estimate

$$\left|\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^N}(\psi+\epsilon\,\phi)\,(t,x;t/\epsilon^2,x/\epsilon)\;\rho^\epsilon(t,x)\,\mathrm{d}x\right| \le C(\psi). \tag{4.15}$$

Moreover, passing to the limit $\epsilon\to 0$, we are led to

$$\lim_{\epsilon\to 0}\left(\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^N}\psi(x)\;\rho^\epsilon(t,x)\,\mathrm{d}x + \int_{\mathbb{R}^N}\left(\mathcal{T}_2^\star(\psi)+\mathcal{T}_1^\star(\phi)\right)(t,x;t/\epsilon^2,x/\epsilon)\;\rho^\epsilon(t,x)\,\mathrm{d}x\right) = 0 \tag{4.16}$$

for any $\psi\in C_c^3(\mathbb{R}^N)$ with $\phi(t,x;\tau,y) = \chi^\star(t,x;\tau,y)\cdot\nabla_x\psi(x)$.

Step 3. We are now ready to obtain the equicontinuity of $\rho^\epsilon(t)$. Indeed, the bound (4.15) shows that for any $\psi\in C_c^3(\mathbb{R}^N)$ the sequence of functions

$$t\longmapsto\int_{\mathbb{R}^N}(\psi+\epsilon\,\phi)\,(t,x;t/\epsilon^2,x/\epsilon)\;\rho^\epsilon(t,x)\,\mathrm{d}x$$

is equicontinuous on $[0,\infty)$. Since the family of functions

$$t\longmapsto\int_{\mathbb{R}^N}\psi(x)\;\rho^\epsilon(t,x)\,\mathrm{d}x$$

is close, up to ϵ, to the previous sequence, it is also equicontinuous. The density of $C_c^3(\mathbb{R}^N)$ in $C_c^0(\mathbb{R}^N)$ allows us to conclude that $\rho^\epsilon(t)$ is vaguely equicontinuous on $[0,\infty)$. Thanks to (4.10) we can apply Propositions 4.3 and 4.4 to the sequence $\rho^\epsilon(t)$.

Step 4. Up to a subsequence there is a double-scale limit R of $\rho^\epsilon(t)$ and the vague limit of $\rho^\epsilon(t)$ is given by the marginal $\rho(t,x) = \int_Y R(t,x;\tau,y)\,\mathrm{d}(\tau,y)$. The limits (4.12) and (4.16) now become

$$\int_0^\infty\int_{\mathbb{R}^N}\int_Y \mathcal{T}_0^\star(\varphi)(t,x;\tau,y)\;R(t,x;\tau,y)\,\mathrm{d}(\tau,y)\,\mathrm{d}x\,\mathrm{d}t = 0 \qquad \forall\varphi\in C_{c,\#}^2([0,\infty)\times\mathbb{R}^N\times\mathbb{R}^{N+1}), \tag{4.17}$$

and, in the distribution sense on $[0,\infty)$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^N}\psi(x)\ \rho(t,x)\,\mathrm{d}x+\int_{\mathbb{R}^N}\int_Y\left(\mathcal{T}_2^\star(\psi)+\mathcal{T}_1^\star(\phi)\right)\ R\ (t,x;\tau,y)\,\mathrm{d}(\tau,y)\,\mathrm{d}x=0$$
$$\forall\psi\in C_c^3(\mathbb{R}^N)\text{ with }\phi=\chi^\star(t,x;\tau,y)\cdot\nabla_x\psi(x), \tag{4.18}$$

respectively. By formally using Proposition 4.5, (4.17) means that the double-scale limit R lies in the orthogonal set of $\mathrm{Ran}(\mathcal{T}_0^\star)=(\mathrm{Ker}(\mathcal{T}_0))^\perp=(\mathrm{Span}\{\Theta\})^\perp$; thus $R\in\mathrm{Span}\{\Theta\}$, which corresponds to the first step of the formal asymptotics. Let us make this argument rigorous (it does not work due to the lack of regularity of the limit R).

Thanks to Proposition 4.5, we obtain from (4.17) that, for any $H\in C_\#^{[0;0,\alpha]}([0,\infty)\times\mathbb{R}^N\times\mathbb{R}^{N+1})$ verifying $\int_Y H\ \Theta\,\mathrm{d}(\tau,y)=0$, we have

$$\int_0^\infty\int_{\mathbb{R}^N}\int_Y H(t,x;\tau,y)\ R(t,x;\tau,y)\,\mathrm{d}(\tau,y)\,\mathrm{d}x\,\mathrm{d}t=0.$$

Let $\varphi\in C_\#^{[0;0,\alpha]}([0,\infty)\times\mathbb{R}^N\times\mathbb{R}^{N+1})$. We write

$$\varphi(t,x;\tau,y)=c_\varphi(t,x)\Theta(t,x;\tau,y)+(\varphi(t,x;\tau,y)-c_\varphi(t,x)\Theta(t,x;\tau,y)),$$

where

$$c_\varphi(t,x)=\left(\int_Y\varphi(t,x;z)\ \Theta(t,x;z)\,\mathrm{d}z\right)\Big/\left(\int_Y\ \Theta(t,x;\tau,y)^2\,\mathrm{d}(\tau,y)\right).$$

Since $\int_Y(\varphi-c_\varphi\Theta)\ R\ \mathrm{d}(\tau,y)=0$, we obtain

$$\begin{aligned}\int_0^\infty\int_{\mathbb{R}^N}\int_Y&\ \varphi\,(t,x;\tau,y)\ R(t,x;\tau,y)\,\mathrm{d}(\tau,y)\,\mathrm{d}x\,\mathrm{d}t\\&=\int_0^\infty\int_{\mathbb{R}^N}c_\varphi(t,x)\left(\int_Y\Theta(t,x;\tau,y)R(t,x;\tau,y)\,\mathrm{d}(\tau,y)\right)\mathrm{d}x\,\mathrm{d}t\\&=\int_0^\infty\int_{\mathbb{R}^N}\int_Y\varphi(t,x;\tau,y)\ \widetilde{\rho}(t,x)\Theta(t,x;\tau,y)\,\mathrm{d}(\tau,y)\,\mathrm{d}x\,\mathrm{d}t,\end{aligned}$$

where $\widetilde{\rho}(t,x)=\int_Y\Theta(t,x;\tau,y)R(t,x;\tau,y)\,\mathrm{d}(\tau,y)\times\big(\int_Y\ \Theta(t,x;\tau,y)^2\,\mathrm{d}(\tau,y)\big)^{-1}$. Hence, it shows that $R(t,x;\tau,y)=\widetilde{\rho}(t,x)\Theta(t,x;\tau,y)$. Furthermore, using $\int_Y\Theta\,\mathrm{d}(\tau,y)=1$, we have $\widetilde{\rho}=\rho$, the weak limit of ρ^ϵ. Plugging this result into (4.18) gives the desired equation for $\rho(t,x)$ and ends the proof of Theorem 4.7. Note that in this way we obtain the diffusion matrix and the drift velocity with the dual formulae (4.6) and (4.7), respectively. ∎

4.5. Nonvanishing ballistic velocity. We can follow step by step the same strategy without assuming condition (4.8). Of course, we cannot define in this situation $\chi^\star$ by (4.5) since the right-hand side does not fulfill the solvability condition. We have to take into account the ballistic velocity

$$c(t,x)=\int_Y u^0\ \Theta(t,x;\tau,y)\,\mathrm{d}(\tau,y)\neq 0.$$

Then we are led to the following statement.

THEOREM 4.8. *Let the assumptions of Theorem* 4.7 *be fulfilled, except for* (4.8). *Then, up to a subsequence,* $\rho^\epsilon \geq 0$ *converges to* ρ *in the vague sense for measures on* $[0,\infty)\times\mathbb{R}^N$. *The limit* ρ *satisfies*

$$\mathrm{div}_x(c\ \rho) = 0$$

in the sense of distributions on $[0,\infty)\times\mathbb{R}^N$. *Moreover, for any* $\psi \in C_c^2([0,\infty)\times\mathbb{R}^N)$ *which satisfies* $c(t,x)\cdot\nabla_x\psi(t,x) = 0$ *for all* $(t,x)\in[0,\infty)\times\mathbb{R}^N$, *we have*

$$\int_{\mathbb{R}^N} \psi(t,x)\ \rho^\epsilon(t,x)\,\mathrm{d}x \xrightarrow[\epsilon\to 0]{} \int_{\mathbb{R}^N} \psi(t,x)\ \rho(t,x)\,\mathrm{d}x,$$

uniformly on any interval $[0,T]$, $0<T<\infty$. *Furthermore, for any such test function* ψ, *the limit satisfies*

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^N}\psi\rho(t,x)\,\mathrm{d}x = \int_{\mathbb{R}^N}\Big(\partial_t\psi\ + v\cdot\nabla_x\psi + \mathrm{div}_x\big(D^T\cdot\nabla_x\psi\big)\ + \eta\Delta_x\psi\Big)\ \rho\,\mathrm{d}x,$$

$$\int_{\mathbb{R}^N}\psi\rho(0,x)\,\mathrm{d}x = \int_{\mathbb{R}^N}\psi\rho_I(x)\,\mathrm{d}x,$$

where D^T *is the transpose of the matrix* D. *Here* D *and* v *are defined as in Theorem* 4.7 *but with* $\chi^\star$ *as the solution of*

$$\mathcal{T}_0^\star(\chi^\star) = u^0 - c, \qquad \int_Y \chi^\star\,\mathrm{d}(\tau,y) = 0. \tag{4.19}$$

Proof. The $L^\infty(0,T;L^1(\mathbb{R}^N))\subset L^1((0,T)\times\mathbb{R}^N)$ estimate on ρ^ϵ allows us to assume that, for a subsequence, $\rho^\epsilon \rightharpoonup \rho$ vaguely in $\mathcal{M}^1((0,T)\times\mathbb{R}^N)$. In Step 2, we rewrite (4.14) as

$$\int_0^T\int_{\mathbb{R}^N}\Big((c-u^0)\cdot\nabla_x\psi + \mathcal{T}_0^\star(\phi)\Big)(t,x;t/\epsilon^2,x/\epsilon)\ \rho^\epsilon\,\mathrm{d}x\,\mathrm{d}t$$

$$-\int_0^T\int_{\mathbb{R}^N} c\cdot\nabla_x\psi\ \rho^\epsilon\,\mathrm{d}x\,\mathrm{d}t \to 0 \quad \text{as } \epsilon\to 0.$$

For a given function ψ, we choose ϕ such that the first integral in this expression vanishes. Namely, we set $\phi(t,x;\tau,y) = \chi^\star(t,x;\tau,y)\cdot\nabla_x\psi(t,x)$, $\chi^\star$ being defined by (4.19). We deduce that

$$\int_0^T\int_{\mathbb{R}^N} c\cdot\nabla_x\psi\ \rho\,\mathrm{d}x\,\mathrm{d}t = 0$$

for any test function ψ. Then, considering now a test function verifying $c\cdot\nabla_x\psi = 0$, we can reproduce the arguments of the proof of Theorem 4.7. ∎

Let q be a (scalar) distribution on $(0,T)\times\mathbb{R}^N$. Clearly, $T = \mathrm{div}_x(c\ q)$ belongs to the orthogonal of $E = \{\psi : \mathbb{R}\times\mathbb{R}^N\to\mathbb{R}$, such that $c\cdot\nabla_x\psi = 0\}$. At least formally, the elements T of $E^\perp$ always have this form. Indeed, let us introduce the symmetric degenerate elliptic operator $\mathcal{A}(\psi) = -\mathrm{div}_x(c\otimes c\nabla_x\psi)$. Remarking that $\int_{\mathbb{R}^N}\mathcal{A}(\psi)\ \psi\,\mathrm{d}x = \int_{\mathbb{R}^N}|c\cdot\nabla_x\psi|^2\,\mathrm{d}x$, we have $E = \mathrm{Ker}(\mathcal{A})$. Thus, formally, a distribution $T\in E^\perp$ lies in the range of $\mathcal{A}$. This means that $T = \mathcal{A}(p) = \mathrm{div}_x(cq)$, with $q = -c\nabla_x p$. In these conditions the equations for ρ read

$$\begin{cases} \mathrm{div}_x(c\ \rho) = 0, \\ \partial_t\rho + \mathrm{div}_x(v\ \rho) - \mathrm{div}_x\Big((D+\eta I_N)\nabla_x\rho\Big) = \mathrm{div}_x(c\ q) \end{cases}$$

and q appears as a Lagrange multiplier associated to the constraint $\mathrm{div}_x(c\ \rho) = 0$.

Actually it is possible that the second part of the statement is meaningless. For instance, consider the simple case $c = (1, 0, \dots) \in \mathbb{R}^N$. Thus, $\mathrm{div}_x(c\rho) = 0 = \partial_{x_1}\rho = 0$ means that ρ does not depend on the first variable. Since ρ is a bounded measure on $\mathbb{R}^N$, this implies that $\rho = 0$! According to this example, the set of compactly supported test functions verifying $c \cdot \nabla_x \psi = 0$ can be reduced to $\{0\}$ (this is the case in the autonomous case with characteristics curves verifying $|X(s,x)| \to \infty$ as $s \to \infty \dots$).

The simplest way to treat this difficulty consists of changing the time scale. Let us define $s = t/\epsilon$. Then we study

$$\frac{1}{\epsilon}\partial_s \rho^\epsilon(s,x) + \frac{1}{\epsilon}\mathrm{div}_x\big((u^0 + \epsilon u^1)(\epsilon s, x; s/\epsilon, x/\epsilon)\ \rho^\epsilon(s,x)\big) = \eta\ \Delta_x \rho^\epsilon(s,x)$$

instead of (2.1), (2.3). At least formally we can replace $u^i(\epsilon s)$ by $u^i(0)$. In this case we have only to consider velocities fields independent on the time scale s. For the sake of generality we consider the problem

$$\frac{1}{\epsilon}\partial_s \rho^\epsilon(s,x) + \frac{1}{\epsilon}\mathrm{div}_x\big((u^0 + \epsilon u^1)(s, x; s/\epsilon, x/\epsilon)\ \rho^\epsilon(s,x)\big) = \eta\ \Delta_x \rho^\epsilon(s,x) \tag{4.20}$$

$$\forall (s,x) \in (0,\infty) \times \mathbb{R}^N,$$

$$\rho^\epsilon(0,x) = \rho_I^\epsilon(x) \qquad \forall x \in \mathbb{R}^N. \tag{4.21}$$

The initial data ρ_I^ϵ is still assumed to be a bounded sequence of nonnegative measures which converges vaguely to ρ_I. Therefore, the sequence of solution satisfies the uniform estimate (4.10). The behavior as ϵ goes to 0 is then simply described by the transport equation with velocity $c(s,x) = \int_Y u^0\ \Theta\,\mathrm{d}(\tau, y)$.

THEOREM 4.9. *Let ρ^ϵ be the solution of* (4.20), (4.21). *We assume that ρ_I^ϵ is a bounded sequence of nonnegative measures which converges vaguely to ρ_I. Then ρ^ϵ converges vaguely locally, uniformly on $\mathbb{R}^+$ to ρ, the solution of the transport equation*

$$\partial_s \rho + \mathrm{div}_x(c\rho) = 0$$

with initial data ρ_I.

Proof. We still follow the strategy of the proof of Theorem 4.7. Multiplying (4.20) by $\varphi(s,x;s/\epsilon,x/\epsilon)$ yields

$$\int_{\mathbb{R}^N} \mathcal{T}_0^\star(\varphi)(s,x;s/\epsilon,x/\epsilon)\ \rho^\epsilon(s,x)\,\mathrm{d}x = -\epsilon \int_{\mathbb{R}^N} (\mathcal{T}_1^\star + \epsilon \mathcal{T}_2^\star)(\varphi)(s,x;s/\epsilon,x/\epsilon)\ \rho^\epsilon(s,x)\,\mathrm{d}x \tag{4.22}$$
$$-\epsilon\ \frac{\mathrm{d}}{\mathrm{d}s}\int_{\mathbb{R}^N} \varphi(s,x;s/\epsilon,x/\epsilon)\ \rho^\epsilon(s,x)\,\mathrm{d}x + \epsilon\ \int_{\mathbb{R}^N} \partial_s\varphi(s,x;s/\epsilon,x/\epsilon)\ \rho^\epsilon(s,x)\,\mathrm{d}x.$$

Hence, we recover (4.12). Then we use $\varphi(s,x;\tau,y) = \psi(x) + \epsilon\phi(s,x;\tau,y)$ as a test function so that the ϵ^0 term in (4.22) disappears. We get

$$\frac{\mathrm{d}}{\mathrm{d}s}\int_{\mathbb{R}^N} (\psi + \epsilon\phi)(s,x;s/\epsilon,x/\epsilon)\ \rho^\epsilon(s,x)\,\mathrm{d}x - \epsilon\int_{\mathbb{R}^N} \partial_s\phi(s,x;s/\epsilon,x/\epsilon)\ \rho^\epsilon(s,x)\,\mathrm{d}x \tag{4.23}$$
$$= \int_{\mathbb{R}^N} (u^0\cdot\nabla_x\psi - \mathcal{T}_0^\star(\phi))(s,x;s/\epsilon,x/\epsilon)\ \rho^\epsilon(s,x)\,\mathrm{d}x$$
$$+ \epsilon\int_{\mathbb{R}^N} \Big((u^1\cdot\nabla_x\psi + \eta\Delta_x\psi) + (\mathcal{T}_1^\star + \epsilon\mathcal{T}_2^\star)(\phi)\Big)(s,x;s/\epsilon,x/\epsilon)\ \rho^\epsilon(s,x)\,\mathrm{d}x.$$

We deduce that $(s \mapsto \int_{\mathbb{R}^N} \psi \rho^\epsilon \, \mathrm{d}x)_{\epsilon>0}$ is equicontinuous on $\mathbb{R}^+$, and thus we prove the compactness of ρ^ϵ. Next, the last term in (4.23) goes to 0 as $\epsilon \to 0$. The first term in the right-hand side can be recast as

$$\int_{\mathbb{R}^N} \Big((u^0 - c) \cdot \nabla_x \psi - \mathcal{T}_0^\star(\phi) \Big)(s, x; s/\epsilon, x/\epsilon) \ \rho^\epsilon(s, x) \, \mathrm{d}x + \int_{\mathbb{R}^N} c(s, x) \cdot \nabla_x \psi \ \rho^\epsilon(s, x) \, \mathrm{d}x.$$

Hence, we choose ϕ depending on ψ so that the first integral vanishes. Namely, we set $\phi = \chi^\star \cdot \nabla_x \psi$ with $\mathcal{T}_0^\star(\chi^\star) = (u^0 - c)$. Then, letting $\epsilon \to 0$, we obtain the relation

$$\frac{\mathrm{d}}{\mathrm{d}s} \int_{\mathbb{R}^N} \psi \ \rho(s, x) \, \mathrm{d}x - \int_{\mathbb{R}^N} c(s, x) \cdot \nabla_x \psi \ \rho(s, x) \, \mathrm{d}x = 0,$$

which ends the proof. □

Remark 8. The situation we observe is reminiscent to the hydrodynamic limits in kinetic theory. There, the small parameter is related to the Knudsen number, i.e., the ratio of the mean-free path over some characteristic length. When the flux associated to the equilibrium states of the equation vanishes, a parabolic scaling has to be considered. It corresponds to the vanishing ballistic velocity in this work. Conversely, when the flux does not vanish (nonvanishing ballistic velocity), a hyperbolic scaling has to be used which corresponds to a faster time scale. We refer to the lecture notes of Golse [17] for a presentation of these questions.

More precise information can be given when we restrict ourselves to purely periodic velocity fields. Namely, we assume that $u^{0,1}$ do not depend on t, x but only on the fast variables $(\tau, y) \in Y$. Consequently, the ballistic velocity c is constant. In such a case, there is no mixing of the time scales, and we can give a result incorporating both the transport through the ballistic velocity and the diffusion at the parabolic time scale. A similar approach has been introduced by Mellet in kinetic theory; see [18]. We also mention the interesting attempts due to Capdeboscq [8, 9].

THEOREM 4.10. *Let $\eta > 0$ and let $u^0, \ u^1 \in C_\#^{[1,\alpha]}(\mathbb{R}^{N+1})$. Let $\left(\rho_I^\epsilon\right)_{\epsilon>0}$ be a bounded sequence of nonnegative measures on $\mathbb{R}^N$. We suppose that $\left(\rho_I^\epsilon\right)_{\epsilon>0}$ converges vaguely to ρ_I. Let $\rho^\epsilon(t)$ be the solutions of the advection diffusion problem* (2.1), (2.2) *with*

$$u^\epsilon(t, x) = \frac{1}{\epsilon} \left(u^0 \left(\frac{t}{\epsilon^2}, \frac{x}{\epsilon} \right) + \epsilon u^1 \left(\frac{t}{\epsilon^2}, \frac{x}{\epsilon} \right) \right).$$

Then, up to a subsequence, $\widetilde{\rho}^\epsilon(t, x) = \rho^\epsilon(t, x + ct/\epsilon)$ converges vaguely locally, uniformly on $\mathbb{R}^+$ to $\widetilde{\rho}$, the solution of the drift-diffusion equation

$$\partial_t \widetilde{\rho} + \mathrm{div}_x(v\widetilde{\rho} - D\nabla_x \widetilde{\rho}) = 0$$

with initial data ρ_I. The velocity v is defined as in Theorem 4.7, *while*

$$D = \int_Y \chi^\star \otimes \big(\Theta(u^0 - c) - 2\eta \nabla_y \Theta \big) \, \mathrm{d}(\tau, y)$$

with $\chi^\star$ solution of $\mathcal{T}_0^\star(\chi^\star) = u^0 - c$.

Proof. We remark that $\widetilde{\rho}^\epsilon$ is the solution of the advection-diffusion problem (2.1), (2.2), with u^ϵ replaced by

$$\widetilde{u}^\epsilon(t, x) = \frac{1}{\epsilon} \left(\widetilde{u}^0 \left(\frac{t}{\epsilon^2}, \frac{x}{\epsilon} \right) + \epsilon \widetilde{u}^1 \left(\frac{t}{\epsilon^2}, \frac{x}{\epsilon} \right) \right),$$
$$\widetilde{u}^0(\tau, y) = u^0(\tau, y + c\tau) - c, \quad \widetilde{u}^1(\tau, y) = u^1(\tau, y + c\tau).$$

The functions $\widetilde{u}^i,\ i=1,2$, are not Y-periodic. But if $(e_0,e_1,\dots,e_N)$ is the canonical basis of $\mathbb{R}^{N+1}$, the new cell basis is $(e_0-c,e_1,\dots,e_N)$ corresponding to the cell

$$\widetilde{Y}=\{(\tau,y)/\tau\in(0,1),\ y\in -c\,\tau+(0,1)^N\}.$$

The functions $\widetilde{u}^i,\ i=1,2$, are $\widetilde{Y}$-periodic. It is easily checked that the null function of the new cell problem is then $\widetilde{\Theta}(\tau,y)=\Theta(\tau,y+c\tau)$. As a consequence we have

$$\int_{\widetilde{Y}}\widetilde{u}^0\ \widetilde{\Theta}(\tau,y)\ \mathrm{d}(\tau,y)=\int_Y(u^0-c)\Theta(\tau,y)\ \mathrm{d}(\tau,y)=0.$$

Then $\widetilde{u}^0$ satisfies the vanishing ballistic velocity condition and we can apply Theorem 4.7. This leads to the result. Note that it is crucial that $u^{0,1}$ depends only on (τ,y) and c is constant. ∎

5. Dissipation properties. The only immediate estimate on $\rho^\epsilon(t)$ is in the space L^1, by using the conservative form of the equation. The asymptotic behavior stated in Theorem 4.7 is obtained for a limit $\rho(t)$, which is a priori only a measure. In this functional framework, there is no uniqueness of the solution of the limit effective advection-diffusion problem. In this subsection we want to recover uniqueness (and regularity) of the limit. Assuming more bounds on the initial data, it is possible to obtain more involved dissipation properties of the equation. Actually, quantities like

$$\int H(\rho^\epsilon/\Psi^\epsilon)\ \Psi^\epsilon\,\mathrm{d}x$$

can be uniformly bounded for convex functions H and a suitable choice of $\Psi^\epsilon>0$.

As a preliminary, we establish a general dissipation property which is a consequence of dissipativity of Markovian processes (see [26]) as it is explained by Collet in [10].

PROPOSITION 5.1. *Let*

$$\mathcal{T}=\partial_t\cdot+\sum_{i=1}^N\partial_{x_i}(u_i(t,x)\cdot)-\sum_{i,j=1}^N\partial_{x_i}(a_{ij}(t,x)\partial_{x_j}\cdot)$$

with bounded coefficients u_i,a_{ij} verifying $\sum_{ij}a_{ij}\xi_i\xi_j\geq 0$ for any $\xi\in\mathbb{R}^N$. Let $H:\mathbb{R}\to\mathbb{R}$ be a C^2 convex function. Let ρ and Ψ be in $H^1_{\mathrm{loc}}(\mathbb{R}^{N+1})$, with $\Psi>0$. Then we have

$$\begin{aligned}\mathcal{T}\left(H\left(\frac{\rho}{\Psi}\right)\ \Psi\right)-H'\left(\frac{\rho}{\Psi}\right)\mathcal{T}(\rho)-G\left(\frac{\rho}{\Psi}\right)\mathcal{T}(\Psi)\\ =-\sum_{i,j=1}^N H''\left(\frac{\rho}{\Psi}\right)\ \Psi\ a_{ij}\partial_i\left(\frac{\rho}{\Psi}\right)\ \partial_j\left(\frac{\rho}{\Psi}\right)\leq 0,\end{aligned}$$

with $G(s)=H(s)-sH'(s)$.

The assumptions on ρ and Ψ allow us to use the chain rule for the functions $\rho,\ \Psi$, and $\frac{\rho}{\Psi}$. Then the above proposition is a consequence of two computations whose results are given in the following lemma.

LEMMA 5.2. *Let $\rho,\Psi:\mathbb{R}^{N+1}\to\mathbb{R}$ be as in Proposition* 5.1. *Let $a:\mathbb{R}^{N+1}\to\mathbb{R}$ be a bounded coefficient. Let $H:\mathbb{R}\to\mathbb{R}$ be a C^2 function. Denote by ∂ any derivative in $\mathbb{R}^{N+1}$. We have*

$$\partial\left(H\left(\frac{\rho}{\Psi}\right)\ a\ \Psi\right)=G\left(\frac{\rho}{\Psi}\right)\ \partial(a\Psi)+H'\left(\frac{\rho}{\Psi}\right)\ \partial(a\rho)$$

with $G(s)=H(s)-sH'(s)$.

Proof. We compute

$$\begin{aligned}\partial\left(H\left(\frac{\rho}{\Psi}\right)\ a\ \Psi\right) &= H'\left(\frac{\rho}{\Psi}\right)\ \partial\left(\frac{\rho}{\Psi}\right)\ a\ \Psi + H\left(\frac{\rho}{\Psi}\right)\ \partial(a\Psi)\\ &= H'\left(\frac{\rho}{\Psi}\right)\ \left(\partial\left(\frac{\rho}{\Psi}\ a\ \Psi\right) - \frac{\rho}{\Psi}\partial(a\Psi)\right) + H\left(\frac{\rho}{\Psi}\right)\ \partial(a\Psi)\\ &= \left(H\left(\frac{\rho}{\Psi}\right) - \frac{\rho}{\Psi}H'\left(\frac{\rho}{\Psi}\right)\right)\partial(a\Psi) + H'\left(\frac{\rho}{\Psi}\right)\ \partial(a\rho). \qquad \square\end{aligned}$$

LEMMA 5.3. *We keep the notation of Lemma* 5.2. *Denote by* ∂_i, ∂_j *any derivatives in* $\mathbb{R}^{N+1}$. *We have*

$$\begin{aligned}\partial_i\left(a\partial_j\left[H\left(\frac{\rho}{\Psi}\right)\ \Psi\right]\right) &= G\left(\frac{\rho}{\Psi}\right)\ \partial_i(a\partial_j\Psi) + H'\left(\frac{\rho}{\Psi}\right)\ \partial_i(a\partial_j\rho)\\ &\quad + H''\left(\frac{\rho}{\Psi}\right)\ a\Psi\ \partial_i\left(\frac{\rho}{\Psi}\right)\ \partial_j\left(\frac{\rho}{\Psi}\right).\end{aligned}$$

Proof. We compute

$$\begin{aligned}&\partial_i\left(a\partial_j\left[H\left(\frac{\rho}{\Psi}\right)\ \Psi\right]\right) = \partial_i\left(aH'\left(\frac{\rho}{\Psi}\right)\ \partial_j\left(\frac{\rho}{\Psi}\right)\ \Psi + aH\left(\frac{\rho}{\Psi}\right)\partial_j\Psi\right)\\ &= a\Psi\ H''\left(\frac{\rho}{\Psi}\right)\partial_i\left(\frac{\rho}{\Psi}\right)\ \partial_j\left(\frac{\rho}{\Psi}\right) + H'\left(\frac{\rho}{\Psi}\right)\ \left[\partial_i\left(a\partial_j\left(\frac{\rho}{\Psi}\right)\ \Psi\right) + \partial_i\left(\frac{\rho}{\Psi}\right)\ a\partial_j\Psi\right]\\ &\quad + H\left(\frac{\rho}{\Psi}\right)\ \partial_i(a\partial_j\Psi).\end{aligned}$$

Then we observe that

$$\begin{aligned}\partial_i\left(a\partial_j\left(\frac{\rho}{\Psi}\right)\ \Psi\right) &= \partial_i\left(a\partial_j\left(\frac{\rho}{\Psi}\ \Psi\right)\right) - a\frac{\rho}{\Psi}\partial_j\Psi\Big)\\ &= \partial_i(a\partial_j\rho) - \frac{\rho}{\Psi}\partial_i(a\partial_j\Psi) - \partial_i\left(\frac{\rho}{\Psi}\right)\ a\partial_j\Psi.\end{aligned}$$

Hence, we get

$$\begin{aligned}\partial_i\left(a\partial_j\left[H\left(\frac{\rho}{\Psi}\right)\ \Psi\right]\right) &= a\Psi\ H''\left(\frac{\rho}{\Psi}\right)\partial_i\left(\frac{\rho}{\Psi}\right)\ \partial_j\left(\frac{\rho}{\Psi}\right)\\ &\quad + H'\left(\frac{\rho}{\Psi}\right)\ \left[\partial_i(a\partial_j\rho) - \frac{\rho}{\Psi}\partial_i(a\partial_j\Psi)\right] + H\left(\frac{\rho}{\Psi}\right)\ \partial_i(a\partial_j\Psi)\\ &= a\Psi\ H''\left(\frac{\rho}{\Psi}\right)\partial_i\left(\frac{\rho}{\Psi}\right)\ \partial_j\left(\frac{\rho}{\Psi}\right) + H'\left(\frac{\rho}{\Psi}\right)\ \partial_i(a\partial_j\rho)\\ &\quad + \left(H\left(\frac{\rho}{\Psi}\right) - \frac{\rho}{\Psi}H'\left(\frac{\rho}{\Psi}\right)\right)\ \partial_i(a\partial_j\Psi). \qquad \square\end{aligned}$$

Then we apply Proposition 5.1 with the operator $\mathcal{T}_\epsilon$, ρ^ϵ, and $\Psi^\epsilon = (\Theta + \epsilon\kappa)(t, x; t/\epsilon^2, x/\epsilon)$, where we recall that κ is the solution of the cell problem $\mathcal{T}_0(\kappa) = -\mathcal{T}_1(\Theta)$. We know (Proposition 4.5) that Θ and κ are continuous function and that Θ is positive. Therefore on every compact K of $\mathbb{R}^{n+1}$ they are bounded and Θ is bounded from below. It guarantees that on K, for ϵ small enough, Ψ^ϵ is bounded from above and below. Of course, positivity of Ψ^ϵ is not guaranteed in the whole space. For that we have to assume a uniform behavior of $u^{0,1}$ at infinity. This problem disappears if the problem (2.1), (2.2) is posed on a bounded domain with periodic or Dirichlet conditions. We deduce the following corollary.

COROLLARY 5.4. *With the same assumption as in Theorem* 4.7, *let* ρ^ϵ *be the solution of the problem* (2.1), (2.2) *on* $[0,\infty)\times\Omega$. *Let us assume the following:*

- $\Omega=\mathbb{R}^N$, *and the functions* Θ, κ *and their derivatives are bounded and* Θ *is bounded from below; or*
- Ω *is a cell or a smooth bounded domain, and the problem is completed by periodic or Dirichlet conditions.*

Set $\Psi^\epsilon(t,x)=(\Theta+\epsilon\kappa)(t,x;t/\epsilon^2,x/\epsilon)>0$, *which is bounded from above and below for* $t\in(0,T)$, $x\in\mathbb{R}^N$, $\epsilon\in(0,\epsilon_0)$. *Let* H *be a* C^2 *convex function, satisfying, for all* $s\geq 0$, $|sH'(s)|\leq CH(s)$ *for some constant* $C>0$. *Suppose that initially*

$$\sup_{\epsilon>0}\int H\left(\frac{\rho_I^\epsilon}{\Psi^\epsilon}\right)\ \Psi^\epsilon\,\mathrm{d}x\leq C<\infty.$$

Then the quantities

$$\begin{cases}\displaystyle\int H\left(\frac{\rho^\epsilon}{\Psi^\epsilon}\right)\ \Psi^\epsilon\,\mathrm{d}x,\\ \displaystyle\int_0^t\int H''\left(\frac{\rho^\epsilon}{\Psi^\epsilon}\right)\ \left|\nabla_x\left(\frac{\rho^\epsilon}{\Psi^\epsilon}\right)\right|^2\,\mathrm{d}x\,\mathrm{d}s\end{cases}$$

are bounded on $(0,T)$, *uniformly with respect to* $0<\epsilon<\epsilon_0$.

Proof. We integrate with respect to x the relation given by Proposition 5.1 and get

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int H\left(\frac{\rho^\epsilon}{\Psi^\epsilon}\right)\ \Psi^\epsilon\,\mathrm{d}x+\int H''\left(\frac{\rho^\epsilon}{\Psi^\epsilon}\right)\ \left|\nabla_x\left(\frac{\rho^\epsilon}{\Psi^\epsilon}\right)\right|^2\,\mathrm{d}x&=\int G\left(\frac{\rho^\epsilon}{\Psi^\epsilon}\right)\ \mathcal{T}_\epsilon(\Psi^\epsilon)\,\mathrm{d}x\\ &\leq(1+C)\int H\left(\frac{\rho^\epsilon}{\Psi^\epsilon}\right)\ \Psi^\epsilon\ \left|\frac{\mathcal{T}_\epsilon(\Psi^\epsilon)}{\Psi^\epsilon}\right|\mathrm{d}x.\end{aligned}$$

We have $\mathcal{T}_\epsilon(\Psi^\epsilon)=(\mathcal{T}_2(\Theta+\epsilon\kappa)+\mathcal{T}_1(\kappa))(t,x;\frac{t}{\epsilon^2},\frac{x}{\epsilon})$, which is a bounded function. Then the Gronwall lemma concludes the proof. ∎

This result allows us to improve the regularity of the limit function. Indeed, with $H(s)=s^2/2$, we get that $\mu^\epsilon(t,x)=\rho^\epsilon/\Psi^\epsilon$ is bounded in $L^\infty(0,T;L^2(\Omega))\cap L^2(0,T;H^1(\Omega))$ (actually, we can obtain bound in any L^p). The double-scale limit of μ^ϵ is

$$R(t,x;\tau,y)/\Theta(t,x;\tau,y)=\rho(t,x).$$

Hence μ^ϵ converges to $\rho(t,x)$, which thus belongs to $L^\infty(0,T;L^2(\Omega))\cap L^2(0,T;H^1(\Omega))$ (or $L^2(0,T;H_0^1(\Omega))$). In this class, there is uniqueness of the solution of the effective advection-diffusion problem. As a consequence, we deduce that the whole sequence converges to a unique cluster point. Let us also remark that $\rho^\epsilon=\mu^\epsilon\ \Psi^\epsilon$ is bounded in $L^\infty(0,T;L^2(\Omega))$, which implies as in the proof of Theorem 4.7 the convergence of ρ^ϵ in $C^0([0,T];L^2(\Omega)\text{-}weak)$. We point out that we cannot have a convergence in L^2-strong due to the oscillating factor Ψ^ϵ.

COROLLARY 5.5. *With the same assumption as in Corollary* 5.4, *suppose that the initial data is bounded in* $L^2(\Omega)$. *Then* ρ^ϵ *converges to* ρ *in* $C^0([0,T];L^2(\Omega)\text{-}weak)$, *the unique solution in* $L^\infty(0,T;L^2(\Omega))\cap L^2(0,T;H^1(\Omega))$ *of* (4.9) *with the corresponding boundary conditions and the initial condition* ρ_I.

Acknowledgments. The authors are particularly grateful to Alain Pumir for fruitful discussions and nice comments about this problem. They also thank Grégoire Allaire and Sylvie Méléard for their help during the preparation of the manuscript.

REFERENCES

[1] G. Allaire, *Homogenization and two-scale convergence*, SIAM J. Math. Anal., 23 (1992), pp. 1482–1518.
[2] Y. Amirat, K. Hamdache, and A. Ziani, *Homogénéisation d'équations hyperboliques du premier ordre et application aux écoulements miscibles en milieux poreux*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 6 (1989), pp. 397–417.
[3] Y. Amirat, K. Hamdache, and A. Ziani, *Some results on homogenization of convection-diffusion equations*, Arch. Ration. Mech. Anal., 114 (1989), pp. 155–178.
[4] M. Avellaneda and A. Majda, *An integral representation and bounds on the effective diffusivity in passive advection by laminar and turbulent flows*, Comm. Math. Phys., 138 (1991), pp. 339–391.
[5] M. Avellaneda and M. Vergassola, *Scalar transport in compressible flow*, Phys. D, 106 (1997), pp. 148–166.
[6] A. Bensoussan, J.-L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, Stud. Math. Appl. 5, North-Holland, Amsterdam, 1978.
[7] Y. Brenier, *Remarks on some linear hyperbolic equations with oscillatory coefficients*, in Proceedings of the Third International Conference on Hyperbolic Problems: Theory, Numerical Methods and Applications, Vol. I, II, B. Engquist and B. Gustafsson, eds., Studentlitteratur, Lund, 1991, pp. 119–130.
[8] Y. Capdeboscq, *Homogenization of a spectral problem with drift*, Proc. Roy. Soc. Edinburgh Sect. A, 132 (2002), pp. 567–594.
[9] Y. Capdeboscq, *Homogénéisation des Modèles de Diffusion en Neutronique*, Thèse Université Paris 6, 1999.
[10] J. F. Collet, work in preparation.
[11] D. Cioranescu and P. Donato, *An Introduction to Homogenization*, Oxford Lecture Ser. Math. Appl. 17, Oxford University Press, New York, 1999.
[12] R. Di Perna and A. Majda, *Oscillations and concentrations in weak solutions of the incompressible fluid equations*, Comm. Math. Phys., 108 (1987), pp. 667–689.
[13] W. E, *Homogenization of linear and nonlinear transport equations*, Comm. Pure Appl. Math., 65 (1990), pp. 301–326.
[14] L. C. Evans, *The perturbed test function method for viscosity solutions of nonlinear PDE*, Proc. Roy. Soc. Edinburgh Sect. A, 111 (1989), pp. 359–375.
[15] L. C. Evans, *Periodic homogenization of certain fully nonlinear partial differential equations*, Proc. Roy. Soc. Edinburgh Sect. A, 120 (1992), pp. 245–265.
[16] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.
[17] F. Golse, *From kinetic to macroscopic models*, in Kinetic Equations and Asymptotic Theory, B. Perthame and L. Desvillettes, eds., Ser. Appl. Math. 4, Gauthier-Villars, Paris, 2000, pp. 41–126.
[18] T. Goudon and A. Mellet, *Homogenization and diffusion asymptotics of the linear Boltzmann equation*, ESAIM Control Optim. Calc. Var., 9 (2003), pp. 371–398.
[19] K. Hamdache, *Homogénéisation non locale d'équations hyperboliques*, in Non Linear PDE's and Their Applications, Collège de France Seminar, Vol. XII., Pitman Res. Notes in Math. 302, Longman Scientific and Technical, Harlow, UK, 1994, pp. 97–112.
[20] K. Hamdache, *Equations de transport, homogénéisation*, Notes de cours, Université Bordeaux 1, Talence, France, 1994.
[21] T. Y. Hou and X. Xin, *Homogenization of linear transport equations with oscillatory vector fields*, SIAM J. Appl. Math., 52 (1992), pp. 34–45.
[22] V. V. Jikov, S. M. Kozlov, and O. A. Oleinik, *Homogenization of Differential Operators and Integral Functionals*, Springer-Verlag, Berlin, 1994.
[23] H. Kesten and G. Papanicolaou, *A limit theorem for turbulent diffusion*, Comm. Math. Phys., 65 (1979), pp. 97–128.
[24] P. Kramer and A. Majda, *Simplified models for turbulent diffusion: Theory, numerical modeling, and physical phenomena*, Phys. Rep., 314 (1999), pp. 237–574.
[25] M. Krein and M. Rutman, *Linear operator leaving invariant a cone in a Banach space*, Amer. Math. Soc. Translation, 10 (1962), pp. 199–325.

[26] R. Kubo, *H-theorems for Markoffian processes*, in Perspectives in Statistical Physics, H. J. Raveché, ed., North-Holland, Amsterdam, 1981, pp. 101–110.
[27] O. A. Ladyzhenskaya, V. A. Solonnikov, and N. N. Uralt'tseva, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monogr. 23, AMS, Providence, RI, 1968.
[28] L. Mascarenhas, *A linear homogenization problem with time dependent coefficient*, Trans. Amer. Math. Soc., 281 (1984), pp. 179–195.
[29] D. W. McLaughlin, G. C. Papanicolaou, and O. R. Pironneau, *Convection of microstructure and related problems*, SIAM J. Appl. Math., 45 (1985), pp. 780–797.
[30] G. Nguetseng, *A general convergence result for a functional related to the theory of homogenization*, SIAM J. Math. Anal., 20 (1989), pp. 608–623.
[31] L. Tartar, *Remarks on homogenization* in Homogenization and Effective Moduli of Material and Media, IMA Vol. Math. Appl., Springer-Verlag, New York, 1986, pp. 228–246.
[32] L. Tartar, *Nonlocal effects induced by homogenization*, in Partial Differential Equations and the Calculus of Variations, Vol. II, Progr. Nonlinear Differential Equations Appl. 2, Birkhäuser Boston, Boston, MA, 1989, pp. 925–938.

SIAM J. MATH. ANAL.
Vol. 36, No. 3, pp. 882–915

A FREE BOUNDARY PROBLEM WITH UNILATERAL CONSTRAINTS DESCRIBING THE EVOLUTION OF A TUMOR CORD UNDER THE INFLUENCE OF CELL KILLING AGENTS*

ALESSANDRO BERTUZZI†, ANTONIO FASANO‡, AND ALBERTO GANDOLFI†

Abstract. A system of tumor cords is schematized by an array of identical cords, each one having approximately a rotational symmetry around its central blood vessel. A mathematical model for the evolution of the cord is presented, taking into account the influence of a limiting nutrient on the proliferation and death of the cells, the volume reduction of the necrotic material due to fluid loss from the cord, and the influence of chemotherapy or radiation treatment. Both the steady state and the evolution problem are considered, showing existence and uniqueness of the solution. A peculiar feature of the evolution model is that the boundary conditions for nutrient concentration on the interface between viable cord and the necrotic region may change during the response to treatment, depending on whether or not new cells enter the necrotic region.

Key words. tumor growth, cancer treatment, free boundary problems for PDEs

AMS subject classifications. 35R35, 92C37, 92C50

DOI. 10.1137/S003614002406060

1. Introduction and model formulation.

In some human and experimental tumors, tumor cells appear to be arranged in cylindrical structures around central blood vessels, generally surrounded by necrosis. These structures are named tumor cords [19, 15, 18]. Oxygen and/or nutrient deprivation in cells remote from the central vessel are likely to play a decisive role in the decrease of cell proliferation rate within the cord and in the occurrence of necrosis. Mathematical models, describing the spatial distribution of proliferating and quiescent cells and the outward directed flux of cells induced by proliferation in a tumor cord at the stationary state, have been recently proposed [3, 5, 12]. The authors represented the proliferating cells as an age- or maturity-structured cell population or as discrete maturity compartments. Existence and uniqueness of the steady-state age density of the cell population within the cord have been shown in [20]. The growth of an isolated tumor cord within the normal tissue, when nutrient is supplied by the central vessel and by a distributed peripheral source that mimics surrounding vessels, has been analyzed in [4].

In the present paper we propose a mathematical model that describes, using the continuum approach, the behavior of a fully developed system of tumor cords under the influence of a therapeutic treatment. The existence of a unique stationary state in the absence of therapy will be shown, as well as the existence and uniqueness of the solution of the evolutive problem that arises following the perturbation of the stationary state. This problem is characterized by the presence of free boundaries, the most important being the ones that confine the necrotic zone: the external boundary is

*Received by the editors April 23, 2002; accepted for publication (in revised form) October 24, 2003; published electronically October 14, 2004. This work was supported in part by CIMAB, by the Progetto Strategico CNR "Metodi e Modelli Matematici nello Studio dei Fenomeni Biologici," and by the National Research Project "Problemi a Frontiera Libera," cofinanced by MIUR.
http://www.siam.org/journals/sima/36-3/40606.html

†Istituto di Analisi dei Sistemi ed Informatica "A. Ruberti" – CNR, Viale Manzoni 30, 00185 Rome, Italy (bertuzzi@iasi.rm.cnr.it, gandolfi@iasi.rm.cnr.it).

‡Dipartimento di Matematica "U. Dini," Università di Firenze, Viale Morgagni 67/A, 50134 Firenze, Italy (fasano@math.unifi.it).

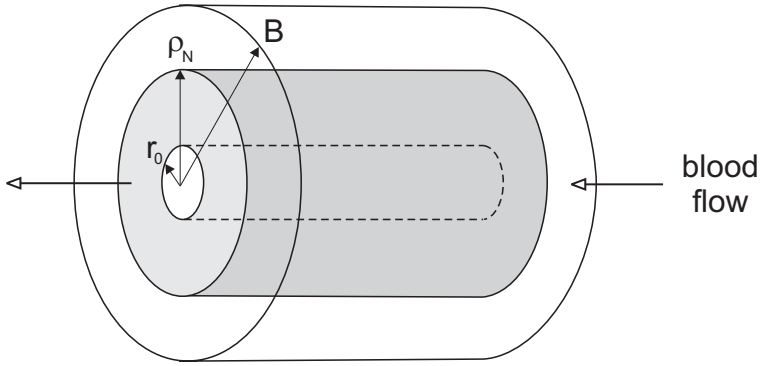

FIG. 1. *Schematic geometry of the tumor cord with the viable region indicated in gray (symbols explained in the text).*

always a no-flux material surface, whereas the internal boundary may be a nonmaterial or a material surface, depending on the evolution of the whole cord.

We concentrate on one cord in the core of the system, supposing that (i) we have symmetry around the axis of the central blood vessel of the cord; (ii) all the quantities describing the cord structure and the concentrations of the various chemical species are independent of the axial coordinate; (iii) there is a cylindrical boundary around the cord where there is no radial exchange of matter (cells, necrotic material, and diffusible chemicals) with the environment. Such a geometry of the outer boundary can be considered a reasonable approximation by viewing the cord inside an array of parallel, identical cords. Figure 1 shows a schematic representation of a tumor cord. As in previous works on the growth of spherical tumors [14, 1, 11, 13] and in [4, 7], we consider for simplicity just one species of "nutrient" with concentration σ. Here we keep the simplified picture in which the system is considered a continuum where we define the concentrations of the various diffusing substances in a global sense, i.e., without distinguishing interstitial and intracellular concentrations. Cells are assumed to die if σ reaches a death threshold σ_N. Moreover, we assume a certain degree of spontaneous cell death within the cord, according to a rate $\mu(\sigma)$, in addition to cell death induced by the treatment. Accordingly, within the cord we have viable cells, dead cells, and extracellular fluids, the respective volume fractions ν_V, ν_N, and ν_E adding to one.

Cells proliferate at the maximum rate χ_0 when $\sigma \geq \sigma_P > \sigma_N$ and ν_E is beyond some threshold $\bar{\nu}_E$. We introduce another threshold $\sigma_Q \in (\sigma_N, \sigma_P)$ below which the progression of cells across cell cycle is arrested and all cells become quiescent, maintaining, however, the capacity to resume the proliferation. The properties of the proliferation rate $\chi(\sigma)$ and of the death rate $\mu(\sigma)$ are stated as follows (see Figure 2):

(H1) $\chi(\sigma), \mu(\sigma)$ continuous and piecewise C^1 functions in $[\sigma_N, \sigma^\star]$, with bounded first derivatives and with $\sigma^\star > \sigma_P$

(H2) $$\chi(\sigma) = \chi_0 \text{ for } \sigma \geq \sigma_P, \text{ and } \chi(\sigma) = 0 \text{ for } \sigma \leq \sigma_Q$$

(H3) $$\chi'(\sigma) > 0 \text{ for } \sigma \in (\sigma_Q, \sigma_P)$$

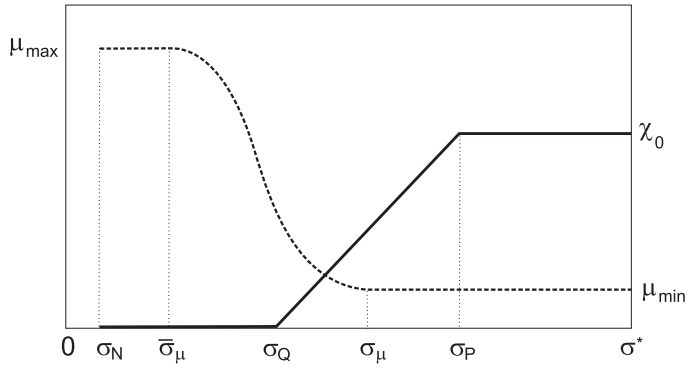

FIG. 2. *Possible graphs of χ (solid line) and μ (dashed line) as functions of σ. For the meaning of the parameters see the text.*

(H4) $$\mu(\sigma)=\mu_{min}\geq 0 \text{ for } \sigma\geq\sigma_\mu, \text{ with } \sigma_\mu\leq\sigma_P$$

(H5) $\mu'(\sigma)\leq 0$; if $\sigma_\mu>\sigma_N$, $\mu'(\sigma)<0$ only in an interval $(\bar\sigma_\mu,\sigma_\mu)$, with $\sigma_\mu>\bar\sigma_\mu\geq\sigma_N$

(H6) $$\chi_0>\mu_{min}\,.$$

Some generalizations are possible, but (H1)–(H6) are physically natural and simplify the exposition. Owing to (i)–(ii) we may consider that the volume fractions and the concentrations of the chemicals depend on the time t and on the radial coordinate r, measured from the axis of the central blood vessel of radius r_0. In the system we distinguish the following regions from inner to outer: P (fully proliferating zone), $\sigma(r,t)\geq\sigma_P$; T (transition zone), $\sigma_P>\sigma(r,t)>\sigma_Q$; Q (quiescent zone), $\sigma_Q\geq\sigma(r,t)>\sigma_N$; and N (purely necrotic zone), $\nu_V=0$. We denote by $\rho_N(t)$ the radius of the interface between the viable cord (the P∪T∪Q region) and the N region, and by $B(t)$ the radius of the exterior boundary of N. The necrotic zone has $\sigma(r,t)=\sigma_N$ in the usually observed conditions, that is, in the steady state of untreated cord, and, as a matter of fact, σ never goes below σ_N if $\sigma(r_0,t)$ remains above σ_N. However, we shall see that during the treatment σ may exceed σ_N in N. Concerning the dead cells, we assume that they decay to a liquid material at a constant rate:

(H7) $$\mu_N>0 \text{ in } \mathrm{P}\cup\mathrm{T}\cup\mathrm{Q},\ \tilde\mu_N>0 \text{ in N.}$$

Such different values reflect the different modes of cell death, i.e., apoptosis within the cord versus necrosis when σ falls to the death threshold. We shall also consider in section 2 the case in which the region N is absent. We denote by $\mathbf{u}$ the velocity field of the cellular component (here assumed to be the same for both living and dead cells) and by $\mathbf{v}$ the velocity of the extracellular fluid. Assuming equal densities for viable cells, dead cells, and extracellular fluid, the increment of cellular volume during proliferation is due to the incorporation of an equal volume of extracellular material. Thus, the governing equations for the three volume fractions in P∪T∪Q are

(1.1) $$\frac{\partial\nu_V}{\partial t}+\nabla\cdot(\mathbf{u}\,\nu_V)=\chi(\sigma)\nu_V-[\mu(\sigma)+\mu_C(c,\sigma)+\mu_R(\sigma,t)]\,\nu_V\,,$$

$$\frac{\partial \nu_N}{\partial t} + \nabla \cdot (\mathbf{u}\, \nu_N) = [\mu(\sigma) + \mu_C(c,\sigma) + \mu_R(\sigma,t)]\, \nu_V - \mu_N \nu_N \,, \tag{1.2}$$

$$\frac{\partial \nu_E}{\partial t} + \nabla \cdot (\mathbf{v}\, \nu_E) = -\chi(\sigma)\nu_V + \mu_N \nu_N \,, \tag{1.3}$$

as long as $\nu_E > \bar{\nu}_E$ (otherwise $\chi(\sigma)$ should be reduced by a factor tending to zero when $\nu_E \to 0$). In (1.1)–(1.3), c is the concentration of a cytotoxic chemical, and $\mu_C(c,\sigma)$ is the chemically induced death rate. The dependence of μ_C on σ allows us to represent the different sensitivity to treatment of cycling cells with respect to quiescent cells. The last term in (1.1) describes the cell killing rate by radiation: the dependence of μ_R on t takes into account the schedule of radiation treatment and the delayed effects following the delivery of a single dose. The death rates μ_C and μ_R are assumed bounded. In the region N, since $\nu_V = 0$, the balance equations reduce to

$$\frac{\partial \nu_N}{\partial t} + \nabla \cdot (\mathbf{u}\, \nu_N) = -\tilde{\mu}_N \nu_N \,, \tag{1.4}$$

$$\frac{\partial \nu_E}{\partial t} + \nabla \cdot (\mathbf{v}\, \nu_E) = \tilde{\mu}_N \nu_N \,. \tag{1.5}$$

Summing up (1.1)–(1.3) and (1.4)–(1.5), and imposing $\nu_V + \nu_N + \nu_E = 1$, we find

$$\nabla \cdot [\mathbf{u}(\nu_V + \nu_N) + \mathbf{v}\nu_E] = 0 \,, \tag{1.6}$$

which expresses total mass conservation. To model the transport of the nutrient we assume, as previously mentioned, a common concentration within the cells and in the extracellular fluid, and uniform diffusivity throughout the system. This proves to be the case for fast diffusing substances like oxygen. In such a way we can write the mass balance equation as follows:

$$\frac{\partial \sigma}{\partial t} - D\Delta\sigma + \nabla \cdot (\sigma[\mathbf{u}(\nu_V + \nu_N) + \mathbf{v}\nu_E]) = -\varphi(\sigma)\nu_V \,, \tag{1.7}$$

where $\varphi(\sigma)$ is the consumption rate of viable cells and D is the diffusion coefficient; $\varphi(\sigma)$ may be assumed to be a function of Michaelis–Menten type. Thus,

(H8) $\varphi(\sigma)$ is a bounded, twice continuously differentiable increasing function for

$$\sigma \geq \sigma_N, \text{ and } \varphi(\sigma_N) > 0.$$

To express the velocity fields in (1.1)–(1.7), one should describe the dynamics of the mixture of cells and extracellular fluids, writing the momentum balance and including the interactions among the components [2]. To avoid new assumptions necessary to express the stress tensor and to take full advantage of the simplified geometry, we decided instead to remain in a purely kinematic framework, introducing the further approximation $\nu_E = \text{constant}$. This assumption appears to be reasonable in view of experimental observations in the untreated cord [18], although ν_E is likely to take different values in P∪T∪Q and in N, and to transiently increase during treatments inducing cell death. Thus we set

$$\nu_E = 1 - \nu^\star, \quad \nu^\star = \text{constant},$$

which amounts to saying that both living and dead cells, despite their volume loss, keep a uniform spatial arrangement. As a result of this simplification, from (1.1), (1.2), and (1.4) we can deduce the following equation for the velocity field $\mathbf{u}$:

$$\nabla \cdot \mathbf{u} = \begin{cases} \chi(\sigma)\dfrac{\nu_V}{\nu^\star} - \mu_N(1 - \dfrac{\nu_V}{\nu^\star}) & \text{in P} \cup \text{T} \cup \text{Q}, \\ -\tilde{\mu}_N \quad \text{in N}. \end{cases} \tag{1.8}$$

From now on, we will set

$$\frac{\nu_V}{\nu^\star} = \nu\,, \quad \nu \in [0,1]\,. \tag{1.9}$$

Moreover, we assume that $\mathbf{u}$ has a negligible component along the axis of the cord (this simplification is justifiable away from the ends of the cord). Thus, denoting by $u(r,t)$ the radial component of $\mathbf{u}$, we write

$$\operatorname{div} u = \frac{1}{r}\frac{\partial}{\partial r}(ru) = \begin{cases} (\chi(\sigma) + \mu_N)\nu - \mu_N & \text{in P} \cup \text{T} \cup \text{Q}, \\ -\tilde{\mu}_N \quad \text{in N}. \end{cases} \tag{1.10}$$

Inserting (1.10) in $\operatorname{div}(\nu u) = \nu \operatorname{div} u + u\partial\nu/\partial r$, we finally get for ν the following equation:

$$\frac{\partial \nu}{\partial t} + u\frac{\partial \nu}{\partial r} + \nu\big[\mu + \mu_C + \mu_R - (\chi + \mu_N)(1-\nu)\big] = 0 \quad \text{in P} \cup \text{T} \cup \text{Q}\,, \tag{1.11}$$

and, integrating (1.10) with the condition $u(r_0,t) = 0$, we find

$$ru = \int_{r_0}^{r} r'[(\chi(\sigma) + \mu_N)\nu - \mu_N]\, dr'\,. \tag{1.12}$$

Since the necrotic material cannot be converted back to living cells, the following condition on $\rho_N(t)$ must be satisfied:

$$u(\rho_N, t) \geq \dot{\rho}_N\,, \tag{1.13}$$

with $u(\rho_N,t) - \dot{\rho}_N$ being the feeding rate per unit surface of the necrotic zone. Condition (1.13) has a central role in the model.

Concerning the equation for σ, by (1.6) and the assumption $\partial\sigma/\partial z = 0$ (z being the axial coordinate) that eliminates from (1.7) the axial component of $\mathbf{v}$, we derive

$$\frac{\partial \sigma}{\partial t} - D\Delta\sigma + \frac{\partial \sigma}{\partial r}[u\nu^\star + \bar{v}(1-\nu^\star)] = -\varphi(\sigma)\nu^\star\nu\,, \tag{1.14}$$

where by $\bar{v}(r,t)$ we denote the average along the axial direction of the radial component of the field $\mathbf{v}$ over the cord length. We may indirectly estimate the value of $\bar{v}$ at the vessel wall from the observation of a fluid loss rate from the vasculature ranging from 0.14 to 0.22 cm^3/h per gram of tissue in experimental tumors [8]. Assuming an overall surface of exchanging vasculature of 20 cm^2/gram, we obtain $(1-\nu^\star)\bar{v} \simeq 70 \div 110$ μm/h, while a typical value for u is of the order of 1 μm/h. Thus we can compare the coefficients of $\partial\sigma/\partial r$ in (1.14), concluding that diffusion with $D > 5{\cdot}10^{-8}$ cm^2/s ($D \simeq 2{\cdot}10^{-5}$ cm^2/s for oxygen [19]) is dominant in cords whose radius is of the order of 100 μm. Moreover, in the typical time scale of cord evolution,

the whole transport process can be considered quasi-stationary. Therefore (1.14) is effectively replaced by

$$\Delta\sigma = f(\sigma)\nu\,, \tag{1.15}$$

where $f(\sigma)=\varphi(\sigma)\nu^\star/D$. The interfaces $r=\rho_P(t)$, $r=\rho_Q(t)$ bounding the zone T are defined implicitly as

$$\sigma(\rho_P(t),t)=\sigma_P\,,\quad \sigma(\rho_Q(t),t)=\sigma_Q\,. \tag{1.16}$$

At the inner boundary, i.e., at the vessel wall, we prescribe

$$\sigma(r_0,t)=\sigma^\star\,. \tag{1.17}$$

When the cells at the boundary $\rho_N(t)$ enter the necrotic zone, that is, when

$$u(\rho_N,t)>\dot\rho_N\,, \tag{1.18}$$

the free boundary between viable cord and N carries the conditions

$$\sigma(\rho_N(t),t)=\sigma_N, \tag{1.19}$$

$$\sigma_r(\rho_N(t),t)=0\,, \tag{1.20}$$

and so the interface is defined implicitly and is not a material surface. We stress that the massive death occurring when the cells cross ρ_N can only be caused by insufficient nutrient, because the treatment was assumed to kill the cells according to bounded rate constants. Thus (1.19) must hold. However, if the treatment kills a sufficiently large number of cells in a sufficiently short time, the sudden decrease of nutrient absorption tends to push outward the free boundary defined by (1.19)–(1.20) with a speed that can easily exceed that of the cells. This possibility cannot be allowed because of (1.13). Therefore, when inequality (1.18) tends to be reversed, we must modify the free boundary conditions, giving up condition (1.19) and replacing it by the equation

$$\dot\rho_N=u(\rho_N,t)\,, \tag{1.21}$$

which expresses that the interface has become a material surface. Condition (1.20) is maintained as the second free boundary condition. When such a switch intervenes, $\sigma(\rho_N,t)$ will rise above σ_N: it is in fact the relative nutrient abundance which is responsible for the switch. Of course, the new constraint

$$\sigma(\rho_N(t),t)\geq\sigma_N \tag{1.22}$$

must be imposed, because no cell can be alive if $\sigma<\sigma_N$, and thus when (1.22) tends to be violated we must revert to the previous formulation. We remark that the possibility for the free boundary $r=\rho_N(t)$ to be nonmaterial or material at different times is a particular feature of this model.

We must also define the exterior boundary, $B(t)$, where there is no exchange of matter. This boundary is defined as

$$\dot B=u(B,t), \tag{1.23}$$

$$B(0) = B_0 \tag{1.24}$$

and is an additional free boundary in the problem.

The discussion about the initial condition for (1.11) and the selection of B_0 is rather delicate (note that (1.11) requires no boundary condition for $r = r_0$, where u vanishes). We denote by $\nu_0(r)$, $\sigma_0(r)$, B_0 the steady-state solution of the system (1.11), (1.12), (1.15), (1.17), (1.19), (1.20), (1.23) in the absence of treatment ($\mu_C = \mu_R = 0$). We will devote section 2 to determining the triple (σ_0, ν_0, B_0). To describe the response to a treatment starting at $t = 0$ of a fully developed tumor cord, we assume

$$\nu(r, 0) = \nu_0(r), \tag{1.25}$$

which implies

$$\sigma(r, 0) = \sigma_0(r)\,. \tag{1.26}$$

The evolution problem must be complemented with the transport equation for the drug concentration c. Also for the cytotoxic chemical we do not distinguish the concentrations inside and outside the cells, and we assume uniform diffusivity. We can thus perform a discussion parallel to the one made for σ. However, we must remark that 1) the boundary condition at r_0 can be rapidly changing due to the pharmacokinetics of the drug (for instance, the half-life in plasma of the drug 5-fluorouracyl is around 10 min), so the process cannot in general be considered quasi-stationary; 2) if the drug diffusivity, D_C, is lower than $5{\cdot}10^{-8}$ cm^2/s, the model should be considerably modified because the field $\bar{v}$ could become important. Neglecting the convective term (as possible, for instance, for the drug tirapazamine, which has $D_C = 7.0 \cdot 10^{-7}$ cm^2/s [16]), we may write for c the following diffusion-absorption equation:

$$\frac{\partial c}{\partial t} - D_C \Delta c = -\varphi_C(c, \sigma)\nu^\star \nu - \lambda c \tag{1.27}$$

with

$$c(r_0, t) = c^\star(t), \tag{1.28}$$

$$c_r(B(t), t) = 0\,, \tag{1.29}$$

$$c(r, 0) = 0\,. \tag{1.30}$$

In (1.27), $\varphi_C(c, \sigma)$ is a continuously differentiable function, positive for $c > 0$ and vanishing for $c = 0$, that represents drug loss by cellular uptake and metabolism. Through the dependence of φ_C on σ it is possible to take into account the different drug uptake into cycling and quiescent cells, whereas the dependence on c accounts for the modality of uptake (for instance, the dependence on c may be of Michaelis–Menten type). The coefficient $\lambda \geq 0$ may be associated with a possible natural decay of c (if the substance is chemically unstable). The function $c^\star(t)$ in (1.28) will represent the pharmacokinetics of the drug in the tumor vasculature.

We summarize here the statement of the evolution problem.

Problem P. Find the following field functions and interfaces with the specified regularity and satisfying the quoted equations:

- $\nu(r,t)$: differential equation (1.11), initial condition (1.25), $\nu(r,t) \in [0,1]$, $\nu \in C^1$ for $r \in [r_0, \rho_N(t)]$ and $t \in [0,T]$;
- $u(r,t)$: integral equation (1.12) for $r_0 \le r \le \rho_N(t)$ and $\mathrm{div} u = -\tilde{\mu}_N$ for $\rho_N(t) < r < B(t)$, continuous across the interface, $\partial u/\partial r$ continuous separately in the two domains;
- $\sigma(r,t)$: differential equation (1.15), condition on the fixed boundary (1.17), either free boundary conditions (1.19)–(1.20) under the constraint (1.18), or (1.20)–(1.21) if the latter is violated, with (1.22) becoming the new constraint; σ, $\partial\sigma/\partial r$, $\partial^2\sigma/\partial r^2$ continuous, and $\partial\sigma/\partial t$ piecewise continuous w.r.t. t, for $r \in [r_0, \rho_N(t)]$ and $t \in [0,T]$;
- $c(r,t)$: satisfies the differential equation (1.27) in the classical sense separately in the N region and its complement, with initial condition (1.30) and boundary conditions (1.28)–(1.29). Moreover, $c \in H^{1+\alpha,(1+\alpha)/2}$ in the whole domain for any $\alpha \in (0,1)$ (the notation of the functional space is taken from [17]);
- $\rho_N(t)$: defined either implicitly through the Cauchy conditions (1.19)–(1.20) or as a material surface by (1.21), continuous and piecewise continuously differentiable;
- $B(t)$: differential equation (1.23), initial condition (1.24), $B(t)$ having Lipschitz continuous first derivative.

The interfaces $\rho_P(t)$, $\rho_Q(t)$ are implicitly defined as the level curves $\sigma = \sigma_P$, $\sigma = \sigma_Q$.

2. The stationary solution. In this section we investigate the stationary solution of the untreated system, to be used as initial data for the evolution problem. For ease of notation in this section we will drop the subscript "0" previously used to denote the stationary solution. Although the typical experimental situation, to which we made specific reference in the formulation of the evolution problem, is characterized by the presence of a necrotic region around viable cells, in the study of the steady state we also envisage the possibility that such a necrotic region may be absent. Both cases may occur, depending on the values of the parameters involved.

Case I*: The stationary problem with a necrotic region.* Find the triple (σ, ν, u) and the boundaries ρ_N, B of the necrotic zone, with $B > \rho_N$, $\sigma(r) > 0$, $\nu(r) \in [0,1]$, and $u(r) > 0$ in (r_0, B), satisfying

$$\Delta\sigma = f(\sigma)\nu\,, \quad r_0 < r < \rho_N\,, \tag{2.1}$$

$$\sigma(r_0) = \sigma^\star\,, \tag{2.2}$$

$$\sigma(\rho_N) = \sigma_N\,, \tag{2.3}$$

$$\sigma_r(\rho_N) = 0\,, \tag{2.4}$$

$$\frac{\partial\nu}{\partial r} + A\nu = 0\,, \quad r_0 < r < \rho_N \tag{2.5}$$

with A given by

$$A = \frac{1}{u}[\mu(\sigma) - (\chi(\sigma) + \mu_N)(1-\nu)]\,, \tag{2.6}$$

$$ru = \begin{cases} \displaystyle\int_{r_0}^{r} r'[(\chi(\sigma)+\mu_N)\nu - \mu_N]\,dr'\,, & r_0 \le r \le \rho_N, \\ \rho_N u(\rho_N) - (\tilde{\mu}_N/2)(r^2 - \rho_N^2)\,, & \rho_N < r \le B\,, \end{cases} \tag{2.7}$$

$$u(B) = 0\,. \tag{2.8}$$

Case II*: The stationary problem without a necrotic region.* The unknowns are (σ, ν, u) and the outer boundary B, again with $\sigma(r) > 0$, $\nu(r) \in [0,1]$, and $u(r) > 0$ in (r_0, B), such that the following equations are satisfied: (2.1), (2.5), (2.6), the first equation in (2.7), all for $r_0 < r < B$, (2.2), (2.8), and

$$\sigma(B) \geq \sigma_N\,, \qquad \sigma > \sigma_N \quad \text{for } r_0 \leq r < B\,, \tag{2.3$'$}$$

$$\sigma_r(B) = 0\,. \tag{2.4$'$}$$

We expect that the latter case occurs when μ_N is sufficiently large and μ is sufficiently large in Q.

We remark that assumption (H8), i.e., $f(\sigma_N) > 0$, is necessary in Case I (otherwise problem (2.1), (2.3), (2.4) for any finite ρ_N can have only the constant solution $\sigma \equiv \sigma_N$). Moreover, in both cases we may say a priori that $\sigma_r < 0$ (in (r_0, ρ_N) or in (r_0, B), respectively) and that, as we shall see very soon, the right derivative of u at r_0 is equal to $\chi_0 - \mu_{min}$, pointing out that $\chi_0 > \mu_{min}$ is a necessary condition for a positive velocity field.

We shall prove an existence and uniqueness theorem treating Cases I and II simultaneously.

THEOREM 2.1. *Under the previously stated assumptions* (H1)–(H8)*, the stationary problem has one unique solution.*

Since u vanishes for $r = r_0$, (2.5) becomes degenerate and we cannot prescribe ν for $r = r_0$. We circumvent this difficulty by noting that when $\sigma > \sigma_P$ equations (2.5)–(2.6) are satisfied by $\nu = \nu_{max}$ with

$$\nu_{max} = 1 - \frac{\mu_{min}}{\chi_0 + \mu_N}\,. \tag{2.9}$$

If $\mu_{min} = 0$, $\nu_{max} = 1$; otherwise $\nu_{max} \in (0,1)$ thanks to (H6). We can state the following lemma that will prove fundamental in establishing uniqueness.

LEMMA 2.2. *When* $\chi(\sigma) = \chi_0$ *and* $\mu(\sigma) = \mu_{min}$, (2.9) *is the only nontrivial bounded solution of* (2.5), (2.6).

Proof. If the limit of ν for $r \to r_0^+$ exists bounded and different from ν_{max} and zero, then the derivative of ν has a nonintegrable singularity near r_0, contradicting the existence of a bounded limit for ν. We can also exclude $\nu \to 0$ because $\partial\nu/\partial r$ and ν would have the opposite sign in a neighborhood of r_0, where $A > 0$. If ν has no limit, ν can oscillate only if all its maxima are equal to ν_{max} and all its minima are equal to zero. However, if $\nu = \nu_{max}$ at any point separated from r_0 (A is then bounded), the only compatible solution of (2.5) is $\nu \equiv \nu_{max}$. Therefore, we must examine only the case in which $\nu \neq \nu_{max}$ for $r > r_0$ and ν tends to ν_{max}. In that case, in the proximity of r_0, $u \simeq (\chi_0 - \mu_{min})(r - r_0)$ and from (2.5) we can see that ν cannot exceed ν_{max}. Setting $\nu = \nu_{max} - \tilde{\nu}$ and using the above first-order approximation for u, we can write a differential equation that describes the behavior of $\tilde{\nu}$ at the leading order in $r - r_0$, on which we impose the condition $\tilde{\nu} \to 0$ as $r \to r_0^+$. Integrating this equation backward from a point $\bar{r}_0 > r_0$, where we suppose $\tilde{\nu}(\bar{r}_0) = \tilde{\nu}_0 < \nu_{max}$ with $\tilde{\nu}_0 > 0$, we obtain

$$\frac{1}{\nu_{max}} \log \left| \frac{\tilde{\nu}(\nu_{max} - \tilde{\nu}_0)}{\tilde{\nu}_0(\nu_{max} - \tilde{\nu})} \right| = -\frac{\chi_0 + \mu_N}{\chi_0 - \mu_{min}} \log \frac{r - r_0}{\bar{r}_0 - r_0}\,,$$

and for any $\tilde{\nu}_0 > 0$ we have a sign incompatibility in the limit $r \to r_0^+$. Hence, it must necessarily be $\tilde{\nu}_0 = 0$ and $\nu \equiv \nu_{max}$. □

Thus we put $\nu \equiv \nu_{max}$ as long as $\sigma \geq \sigma_P$, i.e., for $r \in [r_0, \rho_P]$, and we work with the (nondegenerate) system (2.5), (2.6) for $r > \rho_P$, with the condition

$$\nu(\rho_P) = \nu_{max}\,. \tag{2.10}$$

A useful a priori result concerning the stationary solution is the following.

LEMMA 2.3. *For any solution of the stationary problem, in Case* I ν *is positive, continuously differentiable, and nonincreasing in* $[r_0, \rho_N]$. *More precisely,*

(a) *if* $\mu \equiv 0$, *then* $\nu \equiv 1$ *in* $[r_0, \rho_N]$;
(b) *if* $\mu = 0$ *for* $\sigma \geq \sigma_\mu$, *with* $\sigma_\mu \in (\sigma_N, \sigma_P]$ *and* $\mu > 0$ *for* $\sigma < \sigma_\mu$, *then* $\nu(r) = 1$ *for* $r_0 \leq r \leq \rho_\mu$, *with* ρ_μ *defined by* $\sigma(\rho_\mu) = \sigma_\mu$, *and* $\nu'(r) < 0$ *for* $\rho_\mu < r < \rho_N$;
(c) *if* $\mu = \mu_{min} > 0$ *for* $\sigma \geq \sigma_\mu$, *with* $\sigma_\mu \in (\sigma_N, \sigma_P]$, *then* $\nu'(r) < 0$ *for* $\rho_P < r < \rho_N$.

Moreover, if $\mu_N > \mu_{max}$, *with* μ_{max} *being the maximal value of* $\mu(\sigma)$, *then* $\nu > \nu_{min}$ *with*

$$\nu_{min} = 1 - \frac{\mu_{max}}{\mu_N}\,. \tag{2.11}$$

In Case II, ν *is positive, continuously differentiable, and nonincreasing in* $[r_0, B)$. *Either* (b) *or* (c) *holds after substituting* B *for* ρ_N.

Proof. The continuity of $\nu'(r)$ is an immediate consequence of $\nu = \nu_{max}$ in $[r_0, \rho_P]$ and of (2.5)–(2.7). We start with Case I. First of all, dealing a priori with a solution, we may use the properties $u(r) \geq \hat{u} > 0$ and $\sigma_r < 0$ in (ρ_P, ρ_N). Therefore, the formal integration of (2.5) for $r > \rho_P$ with $\nu(\rho_P) = \nu_{max}$ provides $\nu > 0$ and, in particular, $\nu(\rho_N) > 0$. Case (a) is trivial: since $\chi + \mu_N > 0$, when $\mu \equiv 0$ the only solution of (2.5) with $\nu(\rho_P) = \nu_{max} = 1$ is $\nu \equiv 1$. We note that when $\mu \equiv 0$, only a solution of the type of Case I is possible. To deal with (b) and (c), let us consider the function $\hat{\nu}(r)$ defined by the condition $A(r) = 0$, that is,

$$\hat{\nu}(r) = 1 - \frac{\mu(\sigma(r))}{\chi(\sigma(r)) + \mu_N}\,. \tag{2.12}$$

Since $\mu_N > 0$, $\hat{\nu}(r)$ also is defined in Q.

In case (b) we have $\nu = \hat{\nu} = 1$ up to $r = \rho_\mu$. Computing

$$\hat{\nu}'(r) = \frac{\sigma_r}{\chi + \mu_N}\left[\frac{\mu\,\chi'}{\chi + \mu_N} - \mu'\right],$$

we see that $\hat{\nu}'(r) < 0$ in the union $\mathcal{I}$ of the intervals where $\mu\,\chi' > 0$ and/or $\mu' < 0$, which by (H5) includes a right neighborhood of ρ_μ and is connected. If $\sigma_\mu \in (\sigma_Q, \sigma_P]$, it is indeed $\mathcal{I} = (\rho_\mu, \max[\rho_Q, \bar{\rho}_\mu])$, with $\bar{\rho}_\mu$ being such that $\sigma(\bar{\rho}_\mu) = \bar{\sigma}_\mu$, whereas if $\sigma_\mu \in (\sigma_N, \sigma_Q]$, it is $\mathcal{I} = (\rho_\mu, \bar{\rho}_\mu)$. Now we prove that as long as $\hat{\nu}' < 0$ we must have $\nu > \hat{\nu}$ and consequently $A > 0$, implying $\nu' < 0$. Suppose that for some $\bar{r} \in \mathcal{I}$ we have $\nu(\bar{r}) < \hat{\nu}(\bar{r})$, implying $A(\bar{r}) < 0$ and $\nu'(\bar{r}) > 0$. As a consequence, there must be a point $r^\star \in (\rho_\mu, \bar{r})$ in which ν has a local minimum and thus $A(r^\star) = 0$ and $\nu(r^\star) < \hat{\nu}(r^\star)$, which is impossible. Also, we can exclude that ν equals $\hat{\nu}$ at some point $\bar{r}$ of $\mathcal{I}$, because at such a point $A = 0$ and thus we are back to the previous contradiction in a left neighborhood of $\bar{r}$. Therefore $\nu > \hat{\nu}$ and $\nu' < 0$ in $\mathcal{I}$. If $\bar{\rho}_\mu = \rho_N$, then $\nu' < 0$ for $r \in (\rho_\mu, \rho_N)$. If not, in the interval $[\rho^\star, \rho_N)$, where $\rho^\star = \max[\rho_Q, \bar{\rho}_\mu]$,

we have $\mu = \mu_{max} > 0$ and $\hat{\nu} = 1 - \mu_{max}/\mu_N$. We investigate the behavior of ν in the interval $[\rho^\star, \rho_N)$. If $\mu_N \le \mu_{max}$, it is $\hat{\nu} \le 0$ and necessarily $\nu > \hat{\nu}$ and $\nu' < 0$ in $[\rho^\star, \rho_N)$. If $\mu_N > \mu_{max}$, it is $\hat{\nu} = \nu_{min}$ and, if $\nu(\rho^\star) > \nu_{min}$, we have $\nu > \nu_{min}$ and $\nu' < 0$ in $(\rho^\star, \rho_N)$, because otherwise, integrating (2.5) backward from a point where $\nu = \nu_{min}$, we conclude that $\nu(\rho^\star) = \nu_{min}$ against our assumption. Suppose now that $\nu(\rho^\star) = \nu_{min}$, which would imply $\nu = \nu_{min}$ in $[\rho^\star, \rho_N)$, and take the difference $\tilde{\nu} = \nu - \nu_{min}$. Recalling that $\nu_{min} = 1 - \mu_{max}/\mu_N$, for $r < \rho^\star$ we have

$$\tilde{\nu}' + \tilde{A}\tilde{\nu} = \delta$$

with

$$\tilde{A} = \frac{1}{u}\left[\mu - \chi\left(\frac{\mu_{max}}{\mu_N} - \tilde{\nu} - \nu_{min}\right) - 2\mu_{max} + \mu_N + \mu_N\tilde{\nu}\right],$$

$$\delta = \frac{1}{u}\left(\mu_{max} - \mu + \chi\frac{\mu_{max}}{\mu_N}\right)\nu_{min}\,.$$

For $\rho^\star - r > 0$ sufficiently small, we have $\tilde{A} > 0$, because $\tilde{A}(\rho^\star) > 0$ since $\mu - \mu_{max} = \mu_N\tilde{\nu} = 0$ for $r = \rho^\star$. Also, $\delta > 0$ in the same interval. Therefore, integrating the ODE for $\tilde{\nu}$ backward from $r = \rho^\star$ with $\tilde{\nu}(\rho^\star) = 0$, we obtain $\tilde{\nu} < 0$, contradicting the already established result $\tilde{\nu} > 0$ for $r < \rho^\star$.

In case (c) the set $\mathcal{I}$ is not necessarily connected, but it includes at least the interval (ρ_P, ρ_Q); thus $\nu' < 0$ in the region T. If μ is constant, then $\mathcal{I}$ coincides with T and, to extend the result $\nu' < 0$ to Q, we may argue as in case (b). If there is a gap between T and the set where $\mu' < 0$, the same argument applies there, leading to the same conclusion.

In Case II, as in Case I, since it is a priori known that $u(r) > 0$ for $r_0 < r < B$, the formal integration of (2.5) starting from $r = \rho_P$, where $\nu = \nu_{max}$, gives $\nu > 0$ for $r \in (r_0, B)$. We note preliminarily that in case (b) it is $B > \rho_\mu$, because u is positive for $r_0 < r \le \rho_\mu$. The same cannot be guaranteed in case (c). Following the above arguments with the necessary slight modifications, the stated properties can be demonstrated. □

The proof of Theorem 2.1 is based on the following argument. First we consider the auxiliary problem in which, in place of conditions (2.3), (2.4), we prescribe

$$\sigma_r(r_0) = \Sigma^\star < 0 \tag{2.13}$$

and we look for (σ, ν, u) having the required properties up to $r = \hat{\rho}$, which is the minimum among the points where u or σ_r vanishes for the first time or where σ takes the value σ_N. In such a way σ is never increasing and $\sigma_{rr} > 0$ as long as $\sigma_r < 0$, and $u > 0$ for $r_0 < r < \hat{\rho}$.

A basic property of the auxiliary problem is that, setting $\nu = \nu_{max}$, we can reduce the system (2.1), (2.2), (2.13) for σ to the nonlinear Volterra integral equation

$$\sigma = \sigma^\star + r_0\Sigma^\star\log\frac{r}{r_0} + \nu_{max}\int_{r_0}^{r} r' f(\sigma)\log\frac{r}{r'}\,dr'\,, \tag{2.14}$$

up to the point $\rho_P(\Sigma^\star)$ at which σ takes the value σ_P, which is defined if $\Sigma^\star$ is less than some negative constant. We also find that in the same interval (r_0, ρ_P)

$$\sigma_r = \Sigma^\star\frac{r_0}{r} + \nu_{max}\int_{r_0}^{r} f(\sigma)\frac{r'}{r}\,dr'\,, \tag{2.15}$$

$$ru = \frac{1}{2}(\chi_0 - \mu_{min})(r^2 - r_0^2)\,. \tag{2.16}$$

Therefore, ν and u are independent of $\Sigma^\star$ as long as $\sigma \geq \sigma_P$, and differentiating (2.14) w.r.t. $\Sigma^\star$ we obtain a linear integral equation in $\partial\sigma/\partial\Sigma^\star$:

$$\frac{\partial\sigma}{\partial\Sigma^\star} = r_0 \log\frac{r}{r_0} + \nu_{max}\int_{r_0}^{r} r' f'(\sigma)\frac{\partial\sigma}{\partial\Sigma^\star}\log\frac{r}{r'}\,dr'\,, \tag{2.17}$$

showing that $\partial\sigma/\partial\Sigma^\star > 0$. The fact that $\partial\sigma_r/\partial\Sigma^\star > 0$ in the same interval is now a consequence of (2.15). It is also easy to conclude that $\rho_P(\Sigma^\star)$ is increasing.

Then, we shall go through the following steps:

1) We show that, for $r > \rho_P(\Sigma^\star)$ and $\Sigma^\star$ in a suitable interval, we can continue the solution (σ, ν, u) in a unique way up to $r = \hat{\rho}$.
2) We prove that σ, σ_r, ν, and u depend monotonically on $\Sigma^\star$ also for $r > \rho_P$.
3) We establish that there is a unique choice of $\Sigma^\star$ such that (2.3), (2.4) or (2.3′), (2.4′) are satisfied.

We start by looking for a priori bounds on $\Sigma^\star$.

LEMMA 2.4. *The value of* $\Sigma^\star$ *such that* (2.3), (2.4) *or* (2.3′), (2.4′) *are satisfied lies in a suitable interval* (Σ_1, Σ_2) *which can be computed a priori.*

Proof. If we consider the Cauchy problem

$$\Delta\sigma = f(\sigma)\nu_{max}\,, \qquad \sigma(r_0) = \sigma^\star\,, \quad \sigma_r(r_0) = \Sigma^\star\,,$$

we realize that both σ and σ_r depend monotonically on $\Sigma^\star$ and that we can choose $\Sigma^\star = \Sigma_2$ in such a way that $\partial\sigma/\partial r$ vanishes where σ takes the value σ_P. From (2.16) we note that $u(\rho_P) > 0$. If (2.3), (2.4) or (2.3′), (2.4′) are to be fulfilled, we must necessarily have $\Sigma^\star < \Sigma_2$. We note that, for all $\Sigma^\star < \Sigma_2$, the function $\rho_P(\Sigma^\star)$ is uniquely defined in a monotone fashion.

Let us now establish a lower bound for $\Sigma^\star$. For any fixed $\Sigma^\star < \Sigma_2$, as long as $\sigma > \sigma_P$, the auxiliary problem is reduced to the integral equation (2.14). We also know that beyond ρ_P the volume fraction ν does not exceed ν_{max}. We compare the continuation of $\sigma(r)$ for $r > \rho_P$ with the function $\omega(r)$ satisfying for $r > \rho_P$

$$\Delta\omega = f(\sigma_P)\nu_{max}\,, \qquad \omega(\rho_P) = \sigma_P\,, \quad \omega_r(\rho_P) = \bar{\Sigma} < 0\,, \tag{2.18}$$

with $\bar{\Sigma}$ chosen in such a way that ω_r vanishes where $\omega = \sigma_N$. If $\sigma_r(\rho_P) \leq \bar{\Sigma}$, then $\omega > \sigma$, $\omega_r > \sigma_r$, and therefore $\sigma_r < 0$ where $\sigma = \sigma_N$. If we denote by Σ_1 a Cauchy datum for $\sigma_r(r_0)$ which produces $\sigma_r(\rho_P) = \bar{\Sigma}$, then we have $\Sigma^\star > \Sigma_1$.

To prove the existence of Σ_1 we write explicitly the function ω for any $\rho_P > r_0$ and any $\bar{\Sigma} < 0$:

$$\omega = \sigma_P + \rho_P\bar{\Sigma}\log\frac{r}{\rho_P} + \frac{1}{2}S\left(\frac{r^2 - \rho_P^2}{2} - \rho_P^2\log\frac{r}{\rho_P}\right)$$

with $S = f(\sigma_P)\nu_{max}$. Imposing that ω_r vanishes where $\omega = \sigma_N$, we obtain an algebraic system for the pair $(\bar{\rho}_N, \bar{\Sigma})$, with $\bar{\rho}_N$ being the point such that $\omega(\bar{\rho}_N) = \sigma_N$. Setting $y = \bar{\rho}_N/\rho_P$, $y > 1$, such a system can be written in the following form:

$$\bar{\Sigma} = -\frac{1}{2}S\rho_P(y^2 - 1)\,, \tag{2.19}$$

$$y^2 \log y - \frac{y^2-1}{2} = \frac{2}{S\rho_P^2}(\sigma_P - \sigma_N)\,, \tag{2.20}$$

the derivative of the left-hand side of (2.20) w.r.t. y being positive for $y>1$. Thus for $\rho_P \in (r_0, \rho_P(\Sigma_2))$ there is a one-to-one mapping between ρ_P and y, through which we can define the continuous function $\bar\Sigma = h(\rho_P)$ with range in a finite interval $\bar\Sigma_{min} \le \bar\Sigma \le \bar\Sigma_{max} < 0$. Going back to the determination of Σ_1, we note that it corresponds to finding a value of $\Sigma^\star$ with the property that the function $\Sigma_\star$, defined as $\Sigma_\star(\Sigma^\star) = \sigma_r(\rho_P(\Sigma^\star))$, takes precisely the value of $\bar\Sigma$ corresponding to $\rho_P(\Sigma^\star)$. Since

$$\Sigma_\star = \Sigma^\star \frac{r_0}{\rho_P} + \frac{\nu_{max}}{\rho_P}\int_{r_0}^{\rho_P} f(\sigma) r\, dr < 0$$

for $\Sigma^\star \in (-\infty, \Sigma_2)$, it is easy to see that $d\Sigma_\star/d\Sigma^\star$ is positive, and we conclude that $\Sigma_\star$ grows from $-\infty$ to 0 as $\Sigma^\star$ varies from $-\infty$ to Σ_2. Hence, we can define a C^1 function $\rho_P = g(\Sigma_\star)$, monotonically increasing from r_0 to $\rho_P(\Sigma_2)$, over the interval $(-\infty, 0)$. Therefore, in the plane $(\bar\Sigma, \rho_P)$ the two graphs $\rho_P = g(\bar\Sigma)$ and $\bar\Sigma = h(\rho_P)$ must have at least one intersection. To each intersection we associate a value of Σ_1 via the mapping $\rho_P \to \Sigma^\star$, and our final definition of Σ_1 is the largest in the set of the values above. □

Now we turn our attention to the solution of the auxiliary problem (step 1).

LEMMA 2.5. *The auxiliary problem* (2.1), (2.2), (2.13), (2.5), (2.6), (2.7) *is uniquely solvable for any* $\Sigma^\star \in (\Sigma_1, \Sigma_2)$ *up to* $r = \hat\rho$.

Proof. For each $\Sigma^\star \in (\Sigma_1, \Sigma_2)$, we find $\sigma(r)$ in $(r_0, \rho_P(\Sigma^\star))$, and beyond ρ_P we consider the continuation $\omega(r)$ obtained by solving (2.18) with $\bar\Sigma = \sigma_r(\rho_P^-)$. For any given function $\nu(r)$ taking values in $(0, \nu_{max}]$ the solution of $\Delta\sigma = f(\sigma)\nu$ with the same Cauchy data in ρ_P as for ω is such that $\sigma \le \omega$, $\sigma_r \le \omega_r$. In particular, σ is decreasing as long as ω is decreasing. So we have an estimate (ρ_P, r_1), with r_1 being such that $\omega_r(r_1) = 0$, of the interval in which σ is decreasing. Also, we note that $u(\rho_P) > 0$ can be computed from (2.16), and that for $r > \rho_P$

$$u(r) > \frac{1}{r}\left[\frac{\chi_0 - \mu_{min}}{2}(\rho_P^2 - r_0^2) - \frac{\mu_N}{2}(r^2 - \rho_P^2)\right] = F(r)\,. \tag{2.21}$$

Thus for any $u_m \in (0, u(\rho_P))$ we can define r_2 such that $F(r_2) = u_m$.

At this point we set up a fixed point argument to prove existence in $(\rho_P, \bar r)$, with $\bar r = \min(r_1, r_2)$. Let us introduce the set of functions

$$\mathcal{N} = \left\{\nu \in C([\rho_P, \bar r]) \,|\, \nu(\rho_P) = \nu_{max},\ \nu \text{ nonincreasing},\ \nu \in [0, \nu_{max}], \mathrm{Lip}\,\nu \le \frac{\mu_{max}\nu_{max}}{u_m}\right\},$$

which, if $\mu = 0$, reduces to the only element $\nu = 1$. For ν given in $\mathcal{N}$, solve the problem

$$\Delta\sigma = f(\sigma)\nu\,, \qquad \sigma(\rho_P) = \sigma_P\,, \quad \sigma_r(\rho_P^+) = \sigma_r(\rho_P^-)\,, \quad r > \rho_P \tag{2.22}$$

and define $\tilde\nu(r)$ as the solution of

$$\frac{\partial\tilde\nu}{\partial r} + \tilde A\tilde\nu = 0\,, \quad \tilde A = \frac{1}{u}[\mu(\sigma) - (\chi(\sigma) + \mu_N)(1-\tilde\nu)]/u\,, \qquad \tilde\nu(\rho_P) = \nu_{max} \tag{2.23}$$

with

$$ru = \rho_P u(\rho_P) + \int_{\rho_P}^{r} r'[(\chi(\sigma) + \mu_N)\nu - \mu_N]\, dr'\,, \tag{2.24}$$

where σ is the solution of (2.22) and ν is the chosen element of $\mathcal{N}$. If $\mu = 0$, the trivial fixed point is $\nu = 1$. We know that in $(\rho_P, \bar{r})$ σ is decreasing and $u > u_m > 0$. Rereading the proof of Lemma 2.2, we see that this is all we need to conclude that $\tilde{\nu}$ is nonincreasing and with range in $(0, \nu_{max}]$. In addition, since $\tilde{A} \geq 0$, we can say that $\tilde{A} \leq \mu_{max}/u_m$ and therefore $\tilde{\nu} \in \mathcal{N}$.

Now take ν_1, $\nu_2 \in \mathcal{N}$ and consider the corresponding functions $\tilde{\nu}_1$, $\tilde{\nu}_2$ as well as $\tilde{A}_1$, $\tilde{A}_2$, σ_1, σ_2, u_1, u_2. We set $\delta = \nu_1 - \nu_2$, $\tilde{\delta} = \tilde{\nu}_1 - \tilde{\nu}_2$. It is not difficult to show that $\tilde{\delta}$ satisfies

$$\frac{\partial \tilde{\delta}}{\partial r} + \left[\tilde{A}_1 + \frac{\tilde{\nu}_2}{u_2}(\chi_2 + \mu_N)\right]\tilde{\delta} = \frac{\tilde{\nu}_2}{u_2}\Big[\overline{\chi}'(\sigma_1 - \sigma_2)(1 - \tilde{\nu}_1) + (u_1 - u_2)\tilde{A}_1 - \overline{\mu}'(\sigma_1 - \sigma_2)\Big], \tag{2.25}$$

with $\overline{\chi}'$ and $\overline{\mu}'$ evaluated at values between σ_1 and σ_2 and $\tilde{\delta}(\rho_P) = 0$, and that the estimates

$$\sup_{(\rho_P, r)} |\sigma_1 - \sigma_2| \leq C_1(r) \int_{\rho_P}^{r} \delta(r')\, dr', \tag{2.26}$$

$$\sup_{(\rho_P, r)} |u_1 - u_2| \leq C_2(r) \int_{\rho_P}^{r} \delta(r')\, dr', \tag{2.27}$$

where $C_1(r)$ and $C_2(r)$ are known increasing functions of r vanishing for $r = \rho_P$, can be obtained by standard arguments. Thus we obtain the inequality

$$|\tilde{\delta}(r)| \leq \int_{\rho_P}^{r} C_3(r') \int_{\rho_P}^{r'} \delta(r'')\, dr''\, dr', \tag{2.28}$$

concluding that the mapping $\nu \to \tilde{\nu}$ is continuous in the sup-norm and contractive for r close enough to ρ_P, which provides existence and uniqueness up to $\bar{r}$. To conclude the proof, we apply the same procedure for $r > \bar{r}$, redefining $\omega(r)$ as the solution of

$$\Delta\omega = f(\sigma(\bar{r}))\nu(\bar{r}), \qquad \omega(\bar{r}^+) = \sigma(\bar{r}^-), \quad \omega_r(\bar{r}^+) = \sigma_r(\bar{r}^-), \tag{2.29}$$

thus shifting r_1 to the right, and redefining $F(r)$ as $F(r) = [\bar{r}u(\bar{r}) - \frac{\mu_N}{2}(r^2 - \bar{r}^2)]/r$, which provides a new value of r_2 through $F(r_2) = u_m$, $u_m \in (0, u(\bar{r}))$. Since we can take u_m arbitrarily close to zero, by repeating this procedure we obtain precisely the desired result. □

We remark that the function $\nu(r)$ obtained as the solution of the auxiliary problem is positive and nonincreasing in $(r_0, \hat{\rho})$. Moreover, if $\mu_{min} > 0$, it is $\nu' < 0$ in $(\rho_P, \hat{\rho})$, whereas if $\mu_{min} = 0$, it is $\nu' < 0$ in $(\rho_\mu, \hat{\rho})$ for the values of $\Sigma^\star$ such that $\sigma(\hat{\rho}) < \sigma_\mu$. These properties can be checked following the argument of Lemma 2.2.

The monotonicity result (step 2) is now stated by the following lemma.

LEMMA 2.6. *The functions σ, ν, and u solving the auxiliary problem depend monotonically on $\Sigma^\star$. Indeed, $\partial\sigma/\partial\Sigma^\star > 0$, $\partial\sigma_r/\partial\Sigma^\star > 0$, $\partial\nu/\partial\Sigma^\star \geq 0$ ($\partial\nu/\partial\Sigma^\star > 0$ in the interval in which ν is decreasing), and $\partial u/\partial\Sigma^\star \geq 0$.*

Proof. The auxiliary problem is equivalently rewritten as

$$\sigma = \sigma^\star + r_0\Sigma^\star \log\frac{r}{r_0} + \int_{r_0}^{r} r' f(\sigma)\nu \log\frac{r}{r'}\, dr', \quad r_0 \leq r \leq \hat{\rho}(\Sigma^\star), \tag{2.30}$$

$$\nu = \nu_{max} \exp\left(- \int_{\rho_P}^{r} A \, dr' \right), \quad \rho_P \le r \le \hat{\rho}(\Sigma^\star), \tag{2.31}$$

together with the expression (2.7) for u. We recall that A contains u, σ and ν, so that (2.31) is just a formal way of representing ν. Differentiating (2.30), (2.7), and (2.31) with respect to $\Sigma^\star$, we obtain

$$\frac{\partial \sigma}{\partial \Sigma^\star} = r_0 \log \frac{r}{r_0} + \int_{r_0}^{r} r' \left[f'(\sigma) \frac{\partial \sigma}{\partial \Sigma^\star} \nu + f(\sigma) \frac{\partial \nu}{\partial \Sigma^\star} \right] \log \frac{r}{r'} \, dr' \,, \tag{2.32}$$

$$r \frac{\partial u}{\partial \Sigma^\star} = \int_{r_0}^{r} r' \left[\chi' \frac{\partial \sigma}{\partial \Sigma^\star} \nu + (\chi + \mu_N) \frac{\partial \nu}{\partial \Sigma^\star} \right] dr' \,, \tag{2.33}$$

$$\frac{\partial \nu}{\partial \Sigma^\star} = -\nu \int_{\rho_P}^{r} \frac{\partial A}{\partial \Sigma^\star} \, dr' \,, \tag{2.34}$$

where we have used $A(\rho_P) = 0$. Next we compute

$$\frac{\partial A}{\partial \Sigma^\star} = -\frac{A}{u} \frac{\partial u}{\partial \Sigma^\star} - \frac{1}{u} \left[\chi' \frac{\partial \sigma}{\partial \Sigma^\star} (1 - \nu) - (\chi + \mu_N) \frac{\partial \nu}{\partial \Sigma^\star} - \mu' \frac{\partial \sigma}{\partial \Sigma^\star} \right] \tag{2.35}$$

in which we may eliminate $\partial u / \partial \Sigma^\star$ making use of (2.33). We recall that the problem for σ is uncoupled up to $r = \rho_P$ (because $\nu = \nu_{max}$) and that $\partial \sigma / \partial \Sigma^\star > 0$ in that interval, as it is easily deduced from (2.32) itself and as it was already mentioned. Thus the monotone dependence must in fact be shown for $r > \rho_P(\Sigma^\star)$ (ρ_P is an increasing function of $\Sigma^\star$). For this reason, for $r > \rho_P$ we rewrite (2.32) in the form

$$\frac{\partial \sigma}{\partial \Sigma^\star} = \left. \frac{\partial \sigma}{\partial \Sigma^\star} \right|_{r = \rho_P(\Sigma^\star)} + r_0 \log \frac{r}{\rho_P} + \int_{\rho_P}^{r} r' \log \frac{r}{r'} f' \nu \frac{\partial \sigma}{\partial \Sigma^\star} \, dr' + \int_{\rho_P}^{r} r' \log \frac{r}{r'} f \frac{\partial \nu}{\partial \Sigma^\star} \, dr' \,, \tag{2.36}$$

while (2.34) becomes, using (2.35) and after some algebra,

$$\frac{1}{\nu} \frac{\partial \nu}{\partial \Sigma^\star} = \int_{\rho_P}^{r} \left[\left(r' F(r', r) \nu + \frac{1}{u} (1 - \nu) \right) \chi' - \frac{\mu'}{u} \right] \frac{\partial \sigma}{\partial \Sigma^\star} \, dr'$$

$$+ \int_{\rho_P}^{r} \left[r' F(r', r) - \frac{1}{u} \right] (\chi + \mu_N) \frac{\partial \nu}{\partial \Sigma^\star} \, dr' \tag{2.37}$$

with

$$F(r', r) = \int_{r'}^{r} \frac{1}{r''} \frac{A}{u} \, dr'' \,, \quad r > r' \,. \tag{2.38}$$

As we said, the term $\partial \sigma / \partial \Sigma^\star |_{r = \rho_P(\Sigma^\star)}$ in (2.36) is strictly positive, while $\partial \nu / \partial \Sigma^\star$ is zero at the same point. We have obtained a system of linear Volterra integral equations for the pair $\partial \sigma / \partial \Sigma^\star$, $\partial \nu / \partial \Sigma^\star$ and we restrict r to staying far from the possible singularity of $1/u$. Let us distinguish the same three cases (a), (b), and (c)

as in the proof of Lemma 2.2. Case (a): $\mu \equiv 0$ implies $\partial\nu/\partial\Sigma^\star \equiv 0$, implying also that $\partial\sigma/\partial\Sigma^\star > 0$. Case (b): $\partial\nu/\partial\Sigma^\star = 0$ up to $r = \rho_\mu$, and we may rewrite (2.37) replacing ρ_P by ρ_μ. Note that the problem for σ is uncoupled in (r_0, ρ_μ), where we can easily see that $\partial\sigma/\partial\Sigma^\star > 0$. We consider the modified version of (2.37) for $r > \rho_\mu$ and sufficiently close to ρ_μ, so that $\partial\sigma/\partial\Sigma^\star > 0$. If $(1-\nu)\chi' - \mu' = -\mu' > 0$ for $r = \rho_\mu^+$, then $\int_{\rho_\mu}^r (1/u)[(1-\nu)\chi' - \mu'](\partial\sigma/\partial\Sigma^\star)\, dr' > 0$ is the only term of order $r - \rho_\mu$, all the remaining ones being at least $O[(r-\rho_\mu)^2]$. Thus $\partial\nu/\partial\Sigma^\star > 0$ for a sufficiently small interval on the right of ρ_μ. If $\mu' = 0$ at $r = \rho_\mu^+$, we rewrite (2.37) in the form

$$Y(r) = \Theta(r) + \Xi(r) - \int_{\rho_\mu}^r \frac{1}{u}\nu(\chi + \mu_N) Y\, dr' , \tag{2.39}$$

having defined

$$Y(r) = \frac{1}{\nu}\frac{\partial\nu}{\partial\Sigma^\star} , \qquad \Theta(r) = \int_{\rho_\mu}^r \frac{1}{u}[(1-\nu)\chi' - \mu']\frac{\partial\sigma}{\partial\Sigma^\star}\, dr' ,$$

$$\Xi(r) = \int_{\rho_\mu}^r r' F(r', r)\nu\left[\frac{\partial\sigma}{\partial\Sigma^\star} + (\chi + \mu_N)Y\right] dr' .$$

We note that $\Theta(r) > 0$, at least as long as $\partial\sigma/\partial\Sigma^\star > 0$, and $\Xi(r) > 0$ for r not too far from ρ_μ (note that $A > 0$ for $r > \rho_\mu$ (see Lemma 2.2), implying $F > 0$, and $Y(\rho_\mu) = 0$). Hence (2.39) can be written as

$$Y(r) + \int_{\rho_\mu}^r a(r') Y(r')\, dr' = Z(r) \tag{2.40}$$

with $Z = \Theta + \Xi > 0$, $a = \nu(\chi + \mu_N)/u \geq 0$. From (2.40) we conclude that Y has the same sign as Z in a neighborhood of ρ_μ. In case (c), either $(1-\nu)\chi' - \mu'$ is positive for $r = \rho_P$, or it is positive in a right neighborhood of it and we can repeat the argument above. Clearly $\partial\sigma/\partial\Sigma^\star$ remains positive if $\partial\nu/\partial\Sigma^\star > 0$ and even in a larger interval beyond the possible sign inversion of $\partial\nu/\partial\Sigma^\star$.

Let us now suppose that $\partial\nu/\partial\Sigma^\star$ vanishes for the first time in some $r = \hat{r}$ after it has become positive. Consider first the case (c) with $\hat{r} \in (\rho_P, \rho_Q]$. From (2.37) we have

$$\int_{\rho_P}^{\hat{r}} \left[\left(r' F(r', \hat{r})\nu + \frac{1}{u}(1-\nu)\right)\chi' - \frac{\mu'}{u}\right]\frac{\partial\sigma}{\partial\Sigma^\star}\, dr' + \int_{\rho_P}^{\hat{r}} \left[r' F(r', \hat{r}) - \frac{1}{u}\right](\chi + \mu_N)\frac{\partial\nu}{\partial\Sigma^\star}\, dr' = 0$$

and for $r > \hat{r}$ we can write

$$\frac{1}{\nu}\frac{\partial\nu}{\partial\Sigma^\star} = \int_{\rho_P}^{\hat{r}} r'\Phi(r, \hat{r})\nu\chi'\frac{\partial\sigma}{\partial\Sigma^\star}\, dr' + \int_{\rho_P}^{\hat{r}} r'\Phi(r, \hat{r})(\chi + \mu_N)\frac{\partial\nu}{\partial\Sigma^\star}\, dr'$$

$$+ \int_{\hat{r}}^r \left[\left(r' F(r', r)\nu + \frac{1}{u}(1-\nu)\right)\chi' - \frac{\mu'}{u}\right]\frac{\partial\sigma}{\partial\Sigma^\star}\, dr' + \int_{\hat{r}}^r \left[r' F(r', r) - \frac{1}{u}\right](\chi + \mu_N)\frac{\partial\nu}{\partial\Sigma^\star}\, dr' \tag{2.41}$$

with

$$\Phi(r,\hat r)=F(r',r)-F(r',\hat r)=\int_{\hat r}^{r}\frac{1}{r''}\frac{A}{u}\,dr''>0\,.$$

The first three terms are positive. Since $\chi'>0$ in some interval and $\Phi(r,\hat r)=O(r-\hat r)$ (remember that A is strictly positive near $\hat r$), we can say that the first and the third terms are $O(r-\hat r)$, thus dominating the last term, whose sign is uncertain. In other words, we have proved that $\partial(\partial\log\nu/\partial\Sigma^\star)/\partial r$ is positive in $\hat r$, contradicting the fact that $\partial\nu/\partial\Sigma^\star$ has attained a minimum there in $[\rho_P,\hat r]$. If we are still in case (c), but $\hat r>\rho_Q$, (2.41) can be simplified, taking into account that $\chi=\chi'=0$ for $\rho_Q<r<\hat r$, and a similar conclusion can be reached. Passing to case (b), we can argue in the same way if $\rho_\mu\in(\rho_P,\rho_Q)$. If instead $\rho_\mu\ge\rho_Q$, we have to modify (2.41) to

$$\frac{1}{\nu}\frac{\partial\nu}{\partial\Sigma^\star}=\int_{\rho_\mu}^{\hat r}r'\Phi(r,\hat r)\mu_N\frac{\partial\nu}{\partial\Sigma^\star}\,dr'+\int_{\hat r}^{r}\left[r'F(r',r)-\frac{1}{u}\right]\mu_N\frac{\partial\nu}{\partial\Sigma^\star}\,dr'-\int_{\hat r}^{r}\frac{\mu'}{u}\frac{\partial\sigma}{\partial\Sigma^\star}\,dr'\,,$$

and we infer the desired conclusion since $\mu_N>0$.

Once we have seen that $\partial\sigma/\partial\Sigma^\star>0$, $\partial\nu/\partial\Sigma^\star>0$, it is straightforward to check that $\partial\sigma_r/\partial\Sigma^\star>0$ by differentiating $r\sigma_r=r_0\Sigma^\star+\int_{r_0}^{r}r'f(\sigma)\nu\,dr'$ w.r.t. $\Sigma^\star$. □

Now we can complete the proof of Theorem 2.1, on the basis of the monotonicity results obtained in the previous lemma.

Proof of Theorem 2.1. In view of Lemma 2.4, we have obtained a one-parameter family of solutions of the auxiliary problem, including all possible solutions of the original problem. The reason why a solution of that family is not a solution of the original problem is related to its behavior at the terminal radial coordinate $\hat\rho$. Namely, we may distinguish the following three disjoint classes of solutions of the auxiliary problem not solving the original problem:

(α) $\sigma_r(\hat\rho)=0$, $\sigma(\hat\rho)>\sigma_N$, $u(\hat\rho)>0$;
(β) $\sigma_r(\hat\rho)<0$, $\sigma(\hat\rho)=\sigma_N$, $u(\hat\rho)>0$;
(γ) $\sigma_r(\hat\rho)<0$, $u(\hat\rho)=0$.

The class (α) is certainly not empty because it contains all the solutions with $\Sigma^\star$ sufficiently close to Σ_2. One of the classes (β) or (γ) may be empty. The class (α) may confine with (β) or with (γ). In the former case, (α) and (β) are generally separated by a solution of type I. In the latter case, (α) and (γ) are separated by a solution of Case II. Classes (β) and (γ) both may exist, but a boundary element, corresponding to some $\Sigma^\star=\Sigma_{\beta\gamma}$, generally belongs to ($\gamma$) and therefore is not a solution. However, there can be the exceptional case in which such a boundary element is precisely the limit solution of Case II characterized by $\sigma_r(\rho_N)=0$, $\sigma(\rho_N)=\sigma_N$, $u(\rho_N)=0$, and also confining with class (α). We can approach the solution of our problem from above or from below, making $\Sigma^\star$ decrease from Σ_2 or increase from Σ_1, respectively. We recall that $\partial\sigma/\partial\Sigma^\star>0$ and that $\partial\nu/\partial\Sigma^\star\ge0$ is not identically zero in $(r_0,\hat\rho)$, except for the trivial case $\mu\equiv0$. This implies that $\partial u/\partial\Sigma^\star$ (always nonnegative) is also not identically zero and in particular is strictly positive near the end point $\hat\rho$ (see (2.33)).

1) $\Sigma^\star$ *decreasing from* Σ_2. Obviously we are moving through the class (α). Taking into account that $\sigma_r(\hat\rho)=0$ and $\sigma(\hat\rho)$ is strictly decreasing as long as $u(\hat\rho)>0$, two cases are possible: either $u(\hat\rho)$ remains positive until $\sigma(\hat\rho)$ reaches σ_N or $u(\hat\rho)$ vanishes before (or possibly when $\sigma(\hat\rho)$ reaches σ_N). In the first case, for the corresponding value of $\Sigma^\star$, a solution of type I is found; otherwise we have obtained a solution of type II.

2) $\Sigma^\star$ *increasing from* Σ_1. For $\Sigma^\star$ close to Σ_1 we may have solutions in (β) or in (γ). Suppose we start with class (β). Increasing $\Sigma^\star$, either $u(\hat\rho)$ remains positive and then we reach exactly the same solution of type I approached from above, or $u(\hat\rho)$ vanishes for some $\Sigma^\star$, meaning that we are shifting to class (γ), unless $\sigma_r(\hat\rho)$ also vanishes, so that we have recovered the limit solution of type II having $\sigma_r(\hat\rho)=0$, $\sigma(\hat\rho)=\sigma_N$, $u(\hat\rho)=0$. Moving within (γ) (possibly from $\Sigma^\star=\Sigma_1$), an increase of $\Sigma^\star$ produces a positive velocity and $\hat\rho$ is shifted to the right. The procedure stops when we reach a $\hat\rho$ such that not only $u(\hat\rho)=0$, but also $\sigma_r(\hat\rho)=0$, so that (2.3′), (2.4′) are satisfied.

As a result of the discussion above, we can say that the interval (Σ_1,Σ_2) is partitioned in one of the following ways:

$$(\Sigma_1,\Sigma_{\beta\gamma})\cup[\Sigma_{\beta\gamma},\Sigma^{\rm II}_{sol})\cup\{\Sigma^{\rm II}_{sol}\}\cup(\Sigma^{\rm II}_{sol},\Sigma_2)\equiv I_\beta\cup I_\gamma\cup\{\Sigma^{\rm II}_{sol}\}\cup I_\alpha, \tag{2.42}$$

$$(\Sigma_1,\Sigma^{\rm I}_{sol})\cup\{\Sigma^{\rm I}_{sol}\}\cup(\Sigma^{\rm I}_{sol},\Sigma_2)\equiv I_\beta\cup\{\Sigma^{\rm I}_{sol}\}\cup I_\alpha\,. \tag{2.43}$$

In (2.42) the intervals I_β, I_γ, I_α correspond to solutions in the respective classes, and $\Sigma^{\rm II}_{sol}$ is the value of $\Sigma^\star$ providing a solution of type II. The interval I_β is possibly empty. In (2.43) $\Sigma^{\rm I}_{sol}$ is the value of $\Sigma^\star$ providing a solution of type I. It is evident that the monotone structure of the family of solutions has implied the uniqueness of the solution to the original free boundary problem. □

All we need to complete the description of the solution of Case I is to calculate the velocity field in the region N according to (2.7). As $u(B)=0$, this gives the steady-state value of B,

$$B=\left[\rho_N^2+\frac{2}{\tilde\mu_N}\rho_N u(\rho_N)\right]^{1/2}. \tag{2.44}$$

3. Existence and uniqueness for the evolution problem. We suppose that at $t=0$ the system is in equilibrium and a purely necrotic region is present (Case I). In addition to (H1)–(H8), we make the following assumptions:

(H9) $\mu_C(c,\sigma)$ is a nonnegative, twice continuously differentiable, bounded function, increasing with respect to c and vanishing for $c=0$.

(H10) $\mu_R(\sigma,t), t\ge 0$, is a nonnegative, twice continuously differentiable, bounded function, with $\mu_R(\sigma,0)=0$.

(H11) $c^\star(t), t\ge 0$, is a nonnegative, continuously differentiable and bounded function with $c^\star(0)=0$.

For technical reasons, we need to extend the definition of the consumption coefficient $\varphi(\sigma)$ for values of σ less than σ_N:

(H12) $\varphi(\sigma)$ is extended for $0\le\sigma<\sigma_N$ in such a way that it possesses the same regularity stated in (H8) and it remains strictly positive.

The aim of this section is to prove the existence and uniqueness of the solution of Problem P stated at the end of section 1. We have the following theorem.

THEOREM 3.1. *Under the assumptions* (H1)–(H12)*, Problem P has a solution* $(\nu,\sigma,u,\rho_N,B,c)$ *in an arbitrarily large time interval.*

THEOREM 3.2. *Under the assumptions* (H1)–(H12)*, and supposing that*

$$\|f'\|\left[R_2^2\log\frac{R_2}{r_0}-\frac{1}{2}(R_2^2-r_0^2)\right]<1\,, \tag{3.1}$$

where R_2 is an upper bound for $\rho_N(t)$, Problem P has one unique solution.

An upper bound for $\rho_N(t)$ will be found later (Lemma 3.2).

To prove existence, we approximate the solution of the evolution problem using a step-by-step procedure. For a given time interval $[0,T]$ we partition it into n equal parts; to simplify notation, we will omit the index "n" in the variables of the approximation of order n. Let us now describe our approximation scheme, starting from the first interval $[0,\theta]$, with $\theta = T/n$. All the quantities referring to the steady state are denoted with the subscript "0" as in section 1.

1. Compute the curves $\gamma(\hat r) : r = \eta(\hat r, t)$ integrating

$$\dot\eta = u_0(\eta)\,, \qquad \eta(\hat r, 0) = \hat r\,, \quad \hat r \in [r_0, B_0]\,. \tag{3.2}$$

Note that $\eta(B_0,t) = B_0$; that is, $B(t)$ is equal to B_0 in $[0,\theta]$. The characteristic lines do not intersect because (3.2) has a unique solution forward and backward (u_0 is indeed Lipschitz continuous), and the equation $r = \eta(\hat r, t)$ defines $\hat r = \zeta(r,t)$ uniquely. Moreover, $\partial\eta/\partial\hat r > 0$ and, more precisely, from $\partial(\partial\eta/\partial\hat r)/\partial t = u_0'(\eta)(\partial\eta/\partial\hat r)$, $\partial\eta/\partial\hat r|_{t=0} = 1$, we can say that $\partial\eta/\partial\hat r = \exp[\int_0^t u_0'(\eta(\hat r,\tau))\,d\tau]$, giving a positive lower (and upper) bound for $\partial\eta/\partial\hat r$. Also, we write $0 = (\partial\eta/\partial\hat r)(\partial\zeta/\partial t) + \partial\eta/\partial t$, giving $\partial\zeta/\partial t = -u_0/(\partial\eta/\partial\hat r)$, which is obviously a priori bounded. Similarly we have $1 = (\partial\eta/\partial\hat r)(\partial\zeta/\partial r)$, implying $\partial\zeta/\partial r = (\partial\eta/\partial\hat r)^{-1}$, positive and a priori bounded. Moreover, $\partial\zeta/\partial r$ and $\partial\zeta/\partial t$ are continuous.

2. In the domain $[r_0, B_0]\times[0,\theta]$ solve the problem (1.27)–(1.30) for c, with $B = B_0$ and omitting the consumption term because in the first step we put $c \equiv c(r,0) = 0$ in $\varphi_C(c,\sigma)$. This problem is standard, and it is well known that $0 \le c \le \sup_{[0,\theta]} c^\star(t)$ and that $|c_r|$ can be estimated in terms of $\sup_{[0,\theta]} |\dot c^\star(t)|$.

3. Integrate the equation

$$\begin{aligned} D_u\nu &= -\nu\big[\mu(\sigma_0) + \mu_R(\sigma_0,t) + \mu_C(c,\sigma_0) - (\chi(\sigma_0)+\mu_N)(1-\bar\nu^0)\big] \\ &\equiv -\nu H(t,c,\sigma_0,\bar\nu^0) \end{aligned} \tag{3.3}$$

along the curves $\gamma(\hat r)$, where D_u is the derivative along the characteristic lines (in this interval $D_u = \partial/\partial t + u_0\,\partial/\partial r$) and we have denoted $\bar\nu^0(r,t) = \nu_0(\zeta(r,t))$. The initial datum is $\nu(\hat r, 0) = \nu_0(\hat r)$, $\hat r \in [r_0, \rho_{N0}]$, and $\nu(\hat r, 0) = 0$, $\hat r \in (\rho_{N0}, B_0]$, with ρ_{N0} being the stationary value of ρ_N. Setting $\mathcal{H}(\hat r, t) = H|_{r=\eta(\hat r,t)}$, we have

$$\nu(r,t) = \nu_0(\zeta(r,t)) \exp\left(-\int_0^t \mathcal{H}(\zeta(r,t),\tau)\,d\tau\right). \tag{3.4}$$

Since ν_0 is strictly positive for $\hat r \in [r_0, \rho_{N0}]$ and $\mathcal{H}$ is bounded, $\nu(r,t)$ also is strictly positive for $r \le \eta(\rho_{N0}, t)$. From (3.4) we may calculate $\partial\nu/\partial r$ and $\partial\nu/\partial t$ as follows:

$$\frac{\partial\nu}{\partial r} = \nu\frac{\partial\zeta}{\partial r}\left(\frac{\nu_0'}{\nu_0} - \int_0^t \frac{\partial\mathcal{H}}{\partial\hat r}\,d\tau\right), \tag{3.5}$$

$$\frac{\partial\nu}{\partial t} = -\nu H + \nu\frac{\partial\zeta}{\partial t}\left(\frac{\nu_0'}{\nu_0} - \int_0^t \frac{\partial\mathcal{H}}{\partial\hat r}\,d\tau\right). \tag{3.6}$$

The right-hand sides of (3.5)–(3.6) contain σ_0, ν_0, c and their first derivatives w.r.t. r.

4. Find $\sigma(r,t)$ and $\tilde\rho_N(t)$ such that

$$\Delta\sigma = f(\sigma)\nu\,, \quad r_0 < r < \tilde\rho_N(t)\,, \tag{3.7}$$

$$\sigma(r_0,t)=\sigma^\star\,, \tag{3.8}$$

$$\sigma(\tilde\rho_N,t)=\sigma_N\,, \tag{3.9}$$

$$\sigma_r(\tilde\rho_N,t)=0\,. \tag{3.10}$$

Equations (3.7)–(3.10) are equivalent to

$$\sigma-\sigma_N=\int_r^{\tilde\rho_N(t)} r'\log\frac{r'}{r}f(\sigma)\nu\,dr'\,, \tag{3.11}$$

and

$$\sigma^\star-\sigma_N=\int_{r_0}^{\tilde\rho_N(t)} r\log\frac{r}{r_0}f(\sigma)\nu\,dr\,. \tag{3.12}$$

We want to show that the pair $(\sigma,\tilde\rho_N)$ can be found with $\tilde\rho_N(0)=\rho_{N0}$ and $\tilde\rho_N(t)<\eta(\rho_{N0},t)$ in some interval $(0,\hat t\,)$, $\hat t\le\theta$. It is easily seen that (3.7)–(3.10) have a unique solution $(\sigma,\tilde\rho_N)$ provided that ν does not approach zero. To fulfill this condition, for the moment we give ν a positive continuous extension for $r>\eta(\rho_{N0},t)$, setting $\nu(r,t)=\nu(\eta(\rho_{N0},t),t)$. Then we recall that $\nu\to\nu_0$ (provided that ν_0 is extended in the same way) and $\partial\nu/\partial t\to 0$ as $t\to 0$ in view of (1.30) and (H9)–(H11), and because of (3.6) and (2.5)–(2.6). We may also establish the continuity of $\partial\sigma/\partial t$, noting that it satisfies the equation $\Delta(\partial\sigma/\partial t)=f'(\sigma)(\partial\sigma/\partial t)\nu+f(\sigma)(\partial\nu/\partial t)$, $r_0<r<\tilde\rho_N(t)$, with zero boundary values at $r=r_0$, $r=\tilde\rho_N(t)$. In particular, since $\partial\nu/\partial t$ vanishes for $t=0$, so does $\partial\sigma/\partial t$. Now we compute $\dot{\tilde\rho}_N$ by differentiation of (3.12), obtaining

$$\dot{\tilde\rho}_N\,\tilde\rho_N[f(\sigma)\nu]\Big|_{r=\tilde\rho_N(t)}\log\frac{\tilde\rho_N}{r_0}=-\int_{r_0}^{\tilde\rho_N} r\log\frac{r}{r_0}\,\frac{\partial}{\partial t}[f(\sigma)\nu]\,dr\,. \tag{3.13}$$

Owing to the remarks above, (3.13) implies $\dot{\tilde\rho}_N\,(0)=0<u_0(\rho_{N0})$. Therefore, in some time interval $[0,\hat t\,]$, $\hat t\le\theta$, the pair $(\sigma,\tilde\rho_N)$ is actually the solution of (3.11)–(3.12), where no use is made of the extension of ν (in other words, ν is precisely the function calculated in step 3). Moreover, (3.13) shows the continuity of $\dot{\tilde\rho}_N$.

Starting from $t=0$, as long as $u_0(\tilde\rho_N)\ge\dot{\tilde\rho}_N$, we set $\rho_N(t)=\tilde\rho_N(t)$ and accept the solution σ given by (3.11) up to the time $\bar t$ such that either a right neighborhood of $\bar t$ exists in which $u_0(\tilde\rho_N)<\dot{\tilde\rho}_N$, or in any right neighborhood of $\bar t$ the difference $u_0(\tilde\rho_N)-\dot{\tilde\rho}_N$ undergoes infinite sign changes. In the first case, for $t>\bar t$ we force ρ_N to coincide with the characteristic line tangent to $r=\tilde\rho_N(t)$ at $t=\bar t$; that is, we set $\dot\rho_N=u_0(\rho_N)$. In this case ρ_N becomes known and we redefine σ by solving the problem (3.7)–(3.8) and (3.10) (with $\tilde\rho_N$ changed to ρ_N). We have

$$\sigma-\sigma(\rho_N(t),t)=\int_r^{\rho_N(t)} r'\log\frac{r'}{r}f(\sigma)\nu\,dr'\,, \tag{3.14}$$

$$\sigma^\star-\sigma(\rho_N(t),t)=\int_{r_0}^{\rho_N(t)} r\log\frac{r}{r_0}f(\sigma)\nu\,dr\,, \tag{3.15}$$

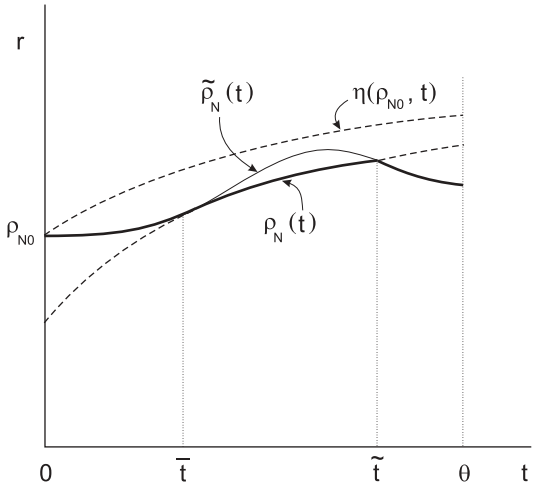

FIG. 3. *The construction of $\rho_N(t)$ (thick line) in the first time step. The characteristic lines are indicated by dashed lines and $\tilde{\rho}_N(t)$ by a thin continuous line. In this example $\hat{t}=\theta$.*

and it can be easily seen that $\sigma(\rho_N(t),t) > \sigma_N$ in an open right neighborhood of $\bar{t}$. In the second case, we must artificially reconstruct the possibility of computing the continuation of the approximate solution over a finite time interval. To this end, we select a velocity u_{tol} as a small fraction of $u_0(\rho_P)$ and we consider the time interval in which $u_0(\tilde{\rho}_N) > \dot{\tilde{\rho}}_N - u_{tol}/n$. Here we choose one of the zeros of the difference $u_0(\tilde{\rho}_N) - \dot{\tilde{\rho}}_N$ beyond which such a quantity becomes negative. At that time we switch to the condition $\dot{\rho}_N = u_0(\rho_N)$, replacing (3.11)–(3.12) with (3.14)–(3.15). Of course, we have to switch back to $\rho_N(t) = \tilde{\rho}_N(t)$ from the possible time $\tilde{t}$ after which (3.15) can be satisfied only with $\sigma(\rho_N(t),t) < \sigma_N$, and we will again have $u_0(\tilde{\rho}_N) > \dot{\tilde{\rho}}_N$ in an open right neighborhood of $\tilde{t}$ with a possible discontinuity of $\dot{\rho}_N$ at $t=\tilde{t}$. If after $\tilde{t}$ the difference $\sigma(\rho_N,t) - \sigma_N$ has infinitely many sign changes in any right neighborhood, after selecting $\sigma_{tol} < \sigma_N$, we consider the time interval in which $\sigma(\rho_N,t) > \sigma_N - \sigma_{tol}/n$, switching to $\rho_N(t) = \tilde{\rho}_N(t)$ at one of the zeros of $\sigma(\rho_N,t) - \sigma_N$ beyond which $\sigma(\rho_N,t) - \sigma_N < 0$.

We remark that if $\hat{t} < \theta$, or $\hat{t} = \theta$ and $\rho_N(\hat{t}) = \eta(\rho_{N0},\theta)$, the time $\bar{t} < \hat{t}$ previously defined will exist. Indeed, in this case we may identify $\hat{t}$ as the time instant such that

$$\sigma^\star - \sigma_N = \int_{r_0}^{\eta(\rho_{N0},\hat{t})} r \log \frac{r}{r_0} f(\sigma)\nu \, dr \,.$$

Therefore, for $t \in (0,\hat{t})$, we have $\eta(\rho_{N0},t) > \tilde{\rho}_N(t)$ and $\eta(\rho_{N0},0) = \tilde{\rho}_N(0)$, $\eta(\rho_{N0},\hat{t}) = \tilde{\rho}_N(\hat{t})$. In this situation the curve $r = \tilde{\rho}_N(t)$ becomes tangent to one of the characteristic lines $\eta(\hat{r},t)$ for some $\hat{r} < \rho_{N0}$ at some time smaller than $\hat{t}$. Since the curve $r = \tilde{\rho}_N(t)$ cannot lie on this characteristic line up to $\hat{t}$, it will leave the characteristic line at some $\bar{t} < \hat{t}$. However, $\bar{t}$ may exist even if $\hat{t} = \theta$ and $\rho_N(\hat{t}) < \eta(\rho_{N0},\theta)$. Figure 3 shows an example of the construction of $\rho_N(t)$.

5. We set ν equal to zero and $\sigma(r,t) = \sigma(\rho_N(t),t)$ for $r > \rho_N(t)$, and this is the final form of ν and σ in the step. Moreover, we continue $c(r,t)$ for $r > B$ by setting

$c(r,t)=c(B(t),t)$.

6. With the new values of σ, ν we compute the new velocity field on the basis of (1.10) as follows:

$$ru=\begin{cases}\displaystyle\int_{r_0}^{r} r'[(\chi(\sigma)+\mu_N)\nu-\mu_N]\,dr', & r_0\le r\le\rho_N(t),\\ \rho_N u(\rho_N)-(\tilde{\mu}_N/2)(r^2-\rho_N^2), & r>\rho_N(t),\end{cases} \tag{3.16}$$

where we have extended the definition of u beyond $r=B(t)$ because we may need it in what follows.

We are now ready to go to the second time step $(\theta,2\theta]$, in which we have

1. Continuation of the characteristic lines $r=\eta(\hat{r},t)$. Starting with the value $\eta(\hat{r},\theta)$ we integrate

$$\dot{\eta}(t)=u(\eta(t),t-\theta),\quad t\in(\theta,2\theta)\,. \tag{3.17}$$

From the continuation of the characteristic line $r=\eta(\hat{r},t)$, the function $\hat{r}=\zeta(r,t)$ is also defined for $t\in(\theta,2\theta]$. Likewise we continue the external boundary as

$$\dot{B}(t)=u(B(t),t-\theta)\,,\quad B(\theta^+)=B(\theta^-)\,. \tag{3.18}$$

2. Computation of c according to

$$\frac{\partial c}{\partial t}-D_C\Delta c=-\varphi_C(c^\theta,\sigma^\theta)\nu^\star\bar{\nu}^\theta-\lambda c\,, \tag{3.19}$$

$$c(r_0,t)=c^\star(t)\,, \tag{3.20}$$

$$c_r(B(t),t)=0\,, \tag{3.21}$$

$$c(r,\theta^+)=c(r,\theta^-)\,,\quad r\in(r_0,B(\theta)]\,. \tag{3.22}$$

Here and in what follows, we denote $c^\theta(r,t)=c(r,t-\theta)$, $\sigma^\theta(r,t)=\sigma(r,t-\theta)$, and $\bar{\nu}^\theta(r,t)=\nu(\eta(\zeta(r,t),t-\theta),t-\theta)$. Note that $\bar{\nu}^\theta(r,t)>0$ if $r\le\eta(\zeta(\rho_N(\theta),\theta),t)$; otherwise $\bar{\nu}^\theta=0$. Therefore, the consumption term in (3.19) is discontinuous. The solution exists in the space $W_q^{2,1}$ (for any $q>1$) and is in fact in the Hölder space $H^{1+\alpha,(1+\alpha)/2}$ for any $\alpha\in(0,1)$ (see [17, Chap. 4]).

3. Computation of ν:

$$D_u\nu=-\nu H(t,c,\sigma^\theta,\bar{\nu}^\theta)\,, \tag{3.23}$$

the initial values for ν being provided by continuity through $t=\theta$. Thus, (3.4) can be extended to the interval $(\theta,2\theta]$.

4. Computation of σ and ρ_N as described in the corresponding point of the first step, the velocity field now being $u(r,t-\theta)$. In the comparison between $u(\tilde{\rho}_N(t),t-\theta)$ and $\dot{\tilde{\rho}}_N(t)$, the expression of $\dot{\tilde{\rho}}_N(t)$ at the general step is still given by (3.13). Note that the existence of $\rho_N(t)$, such that $u(\rho_N,t-\theta)>\dot{\rho}_N(t)$ in a right neighborhood of $t=\theta$, is now not guaranteed.

5. The function ν is set equal to zero and $\sigma \equiv \sigma(\rho_N(t), t)$ for $r > \rho_N(t)$. The function $c(r,t)$ is also continued as in the first step.

6. Computation of u by means of (3.16).

Precisely the same scheme can be iterated up to $t = T$. In the following, when we refer to $\partial\nu/\partial t$, we mean that it is calculated in the positivity set of ν. In particular, the sup-norm $\|\partial\nu/\partial t\|$ is likewise referred to the support of ν.

Remark 3.1. In the approximating solutions constructed above, it is not difficult to see that the function $B(t)$ together with all the characteristic lines is $C^1[0,T]$, whereas the interface $\rho_N(t)$ is not in general continuously differentiable at the switching points. The function $\nu(r,t)$ is C^1 for $r \in [r_0, \rho_N]$ and for $r \in (\rho_N, \infty)$, $t \in [0,T]$. The functions $\sigma(r,t)$, $u(r,t)$, and $c(r,t)$ are continuous in $[r_0,\infty)\times[0,T]$. Moreover, as we shall see, the function c belongs to $H^{1+\alpha,(1+\alpha)/2}$, $\alpha \in (0,1)$. The functions $u(r,t)$ and $\rho_N(t)$ satisfy $u(\rho_N(t), t-\theta) - \dot\rho_N(t) > -u_{tol}/n$ (with $t-\theta$ set to zero for $t<\theta$), so they will not necessarily satisfy inequality (1.13). Also, it may happen that $\sigma < \sigma_N$ and $\nu > 0$ at the same (r,t) point; the approximating solutions may therefore be "nonphysical."

Our aim is now to show that the sequence of approximating solutions so generated defines sets of functions that, when restricted to suitable compact domains, are compact in the sup-norm. First, we establish the following properties.

LEMMA 3.3. *In the family of approximating solutions, the functions ν for $r \in [r_0, \rho_N(t)]$ and $t \in [0,T]$ satisfy the inequalities*

$$0 < N_1 \le \nu(r,t) \le N_2\,, \tag{3.24}$$

where

$$N_1 = \inf_{r\in[r_0,\rho_{N0}]} \nu_0(r)\mathrm{e}^{-\|H\|T}, \qquad N_2 = \sup_{r\in[r_0,\rho_{N0}]} \nu_0(r)\mathrm{e}^{\|H\|T} \tag{3.25}$$

and $\|H\|$ denotes the sup of $|H|$.

Proof. Having defined the characteristic line $r = \eta(\hat r, t)$ in the whole interval $[0,T]$, we can define the function $\hat r = \zeta(r,t)$ for $r \in [r_0, B(t)]$ and $t \in [0,T]$. Thus, we can extend (3.4) for $r \in [r_0, \rho_N(t)]$, $t \in [0,T]$, namely,

$$\nu(r,t) = \nu_0(\zeta(r,t)) \exp\left(-\int_0^t \mathcal{H}(\zeta(r,t),\tau)\, d\tau \right), \tag{3.26}$$

where $\zeta(r,t) \in [r_0, \rho_{N0}]$. From (3.26), the inequalities (3.24) follow immediately considering that $\|H\| \le \max[\max\mu + \max\mu_R + \max\mu_C, \chi_0 + \mu_N]$. □

LEMMA 3.4. *In the family of approximating solutions, $\rho_N(t)$ satisfies the inequalities*

$$r_0 < R_1 < \rho_N(t) < R_2\,, \tag{3.27}$$

where

$$R_1 = \left[r_0^2 + (\hat R_1^2 - r_0^2)\mathrm{e}^{-\mu_{Nmax}T} \right]^{1/2} \tag{3.28}$$

with $\mu_{Nmax} = \max[\mu_N, \tilde\mu_N]$, with $\hat R_1$ being the unique solution larger than r_0 of

$$x^2 \log\frac{x}{r_0} - \frac{1}{2}(x^2 - r_0^2) = 2\frac{\sigma^\star - \sigma_N}{f(\sigma^\star)N_2} \tag{3.29}$$

and R_2 the unique solution larger than r_0 of

$$x^2 \log \frac{x}{r_0} - \frac{1}{2}(x^2 - r_0^2) = 2\frac{\sigma^\star}{f(\sigma_N)N_1}\,. \tag{3.30}$$

Moreover,

$$B(t) > \rho_N(t)\,. \tag{3.31}$$

Proof. Let $\mathcal{T}_1$ be the set of values of $t \in [0,T]$ such that $\sigma(\rho_N(t),t) = \sigma_N$. For $t \in \mathcal{T}_1$, we have from (3.12) that

$$\sigma^\star - \sigma_N \le f(\sigma^\star)\frac{N_2}{2}\left[\rho_N^2 \log \frac{\rho_N}{r_0} - \frac{1}{2}(\rho_N^2 - r_0^2)\right], \tag{3.32}$$

so that $\rho_N(t) \ge \hat{R}_1 > R_1 > r_0$. Recalling the construction of ρ_N, at the time points of the set $\mathcal{T}_2 = [0,T] - \mathcal{T}_1$ (if not empty) the curve $r = \rho_N(t)$ is tangent to a characteristic line. Let us now consider a generic characteristic line $r = \eta(t)$ passing through (t', r'). For $t \ge t'$, $\eta(t)$ satisfies

$$\dot{\eta} = u(\eta(t), t-\theta)\,, \quad \eta(t') = r' \tag{3.33}$$

($t-\theta$ set to zero when $t < \theta$). From (3.16) we have

$$ru(r, t-\theta) > -\frac{\mu_{Nmax}}{2}(r^2 - r_0^2) \tag{3.34}$$

and thus

$$\eta\dot{\eta} > -\frac{\mu_{Nmax}}{2}(\eta^2 - r_0^2)\,, \tag{3.35}$$

which implies

$$\eta(t)^2 - r_0^2 > (r'^2 - r_0^2)\mathrm{e}^{-\mu_{Nmax}(t-t')}\,. \tag{3.36}$$

If $\mathcal{T}_2$ is not empty, for $t \in \mathcal{T}_2$ let $\mathcal{T}_{1t}$ be the subset of $\mathcal{T}_1$ such that $t > \tau$ for each $\tau \in \mathcal{T}_{1t}$, and let $s_t = \sup \mathcal{T}_{1t}$. The characteristic line to which $(t, \rho_N(t))$ belongs will pass through $(s_t, \rho_N(s_t))$, so from (3.36) we have

$$\rho_N(t)^2 - r_0^2 > (\rho_N(s_t)^2 - r_0^2)\mathrm{e}^{-\mu_{Nmax}(t-s_t)}\,. \tag{3.37}$$

Since $\rho_N(s_t) \ge \hat{R}_1$, it follows that $\rho_N(t) > R_1$ also in $\mathcal{T}_2$. Turning now to the upper bound, we have from (3.15) for $t \in [0,T]$

$$\sigma^\star - \sigma(\rho_N(t),t) \ge f(\sigma_N)\frac{N_1}{2}\left[\rho_N^2 \log \frac{\rho_N}{r_0} - \frac{1}{2}(\rho_N^2 - r_0^2)\right], \tag{3.38}$$

so that $\rho_N(t)$ is smaller than the solution larger than r_0 of

$$\frac{1}{2}x^2 \log \frac{x}{r_0} - \frac{1}{4}(x^2 - r_0^2) = \max_{t\in[0,T]} \frac{\sigma^\star - \sigma(\rho_N(t),t)}{f(\sigma_N)N_1}\,, \tag{3.39}$$

which is smaller than the solution R_2 of (3.30).

To prove $B(t) > \rho_N(t)$, it is enough to recognize that for each $t \in [0,T]$ there exists $\hat{r}_t \in (r_0, \rho_{N0}]$ such that $\rho_N(t) = \eta(\hat{r}_t, t)$. Taking into account that $B_0 > r_0$ and that the characteristic lines do not intersect each other, the property (3.31) follows. □

Moreover, we have the following lemma.

LEMMA 3.5. *In the family of approximating solutions, the functions ρ_N, B, and $\dot{B}$ are uniformly bounded and uniformly Lipschitz continuous. The functions σ, u are uniformly bounded and uniformly Lipschitz continuous in $[r_0, M_B] \times [0,T]$, M_B denoting a uniform upper bound of B. The function ν has the same property in $r_0 \le r \le \rho_N(t)$, $t \in [0,T]$. The function c is estimated uniformly in $H^{1+\alpha,(1+\alpha)/2}$, $\alpha \in (0,1)$, for $(r,t) \in [r_0, M_B] \times [0,T]$. In addition, in any domain whose closure has a positive distance from the boundary $r = r_0$ and from the interface $r = \rho_N(t)$, we have uniform estimates of the norm of c in the space $H^{2+\alpha,1+\alpha/2}$.*

Proof. The uniform boundedness of ν and ρ_N is given by Lemmas 3.1 and 3.2, respectively. Moreover, $c(r,t)$ takes values between 0 and $\sup_{[0,T]} c^\star(t)$, owing to the maximum principle. Again from the maximum principle, we can say that $0 < \sigma \le \sigma^\star$. Recalling (3.16) and taking into account the uniform boundedness of ν and ρ_N, the uniform boundedness of u, B, and $\dot{B}$ easily follows. Since H is uniformly bounded, and because of (3.24), $D_u \nu$ is also uniformly bounded.

In order to prove that ν is uniformly Lipschitz in $r_0 \le r \le \rho_N(t)$, $t \in [0,T]$, we note that (3.5)–(3.6) are also valid in the whole domain, with σ^θ, $\bar{\nu}^\theta$ suitably replacing σ_0, $\bar{\nu}^0$ in the expression of $\mathcal{H}$. First we use (3.5), noting that $\partial \mathcal{H}/\partial \hat{r}$ involves the derivatives $\partial \sigma^\theta/\partial r$, $\partial c/\partial r$, and $\partial \bar{\nu}^\theta/\partial r$ multiplied by $\partial \eta/\partial \hat{r}$, which can be easily estimated for each t as done in the first step. The derivatives $\partial \zeta/\partial r$, $\partial \zeta/\partial t$ also are bounded, as explained in the first step. Thus, from (3.5), we can derive a Gronwall-type inequality for $\sup_{[0,t]} |\partial \nu/\partial r|_{\gamma(\hat{r})}$, leading to an estimate of $\sup |\partial \nu/\partial r|$ in terms of ν_0, ν_0' and the sup of $|\partial \sigma^\theta/\partial r|$ and $|\partial c/\partial r|$. Uniform bounds on σ^θ, $|\partial \sigma^\theta/\partial r|$, $|\partial^2 \sigma^\theta/\partial r^2|$ are trivial. To find a bound on $|\partial c/\partial r|$, take the transformation

$$\tilde{r} - r_0 = \frac{R - r_0}{B(t) - r_0}(r - r_0) \tag{3.40}$$

that carries the domain $r_0 < r < B(t)$, $0 < t < T$ into a fixed domain $r_0 < \tilde{r} < R$, $0 < t < T$. Defining $\tilde{c}(\tilde{r}, t) = c(r(\tilde{r}), t)$, the operator $\mathcal{L}c = \partial c/\partial t - D_C \Delta c$ becomes

$$\begin{aligned} \tilde{\mathcal{L}}\tilde{c} = \frac{\partial \tilde{c}}{\partial t} - \frac{\partial \tilde{c}}{\partial \tilde{r}}\left[\frac{\dot{B}(\tilde{r} - r_0)}{B - r_0} + D_C \frac{R - r_0}{B - r_0}\left((\tilde{r} - r_0)\frac{B - r_0}{R - r_0} + r_0\right)^{-1}\right] \\ - D_C \left(\frac{R - r_0}{B - r_0}\right)^2 \frac{\partial^2 \tilde{c}}{\partial \tilde{r}^2}, \end{aligned} \tag{3.41}$$

and the problem (3.19)–(3.22) for c can be rewritten for $\tilde{c}$. Since c, σ, ν are bounded and $\dot{B}/(B - r_0)$, $(R - r_0)/(B - r_0)$, $(B - r_0)/(R - r_0)$ are a priori bounded ($B - r_0$ has indeed a positive lower bound; see Lemma 3.2), we can apply well-known results (see Theorem 9.1 and the remark at the end of section 9 in [17, Chap. 4]) guaranteeing uniform estimates for the norms of $\tilde{c}$ (and hence of c) at least in the spaces $W_q^{2,1}$ (for any $q > 1$) and $H^{1+\alpha,(1+\alpha)/2}$ (for any $\alpha \in (0,1)$). In particular we now have the uniform bound for $|\partial c/\partial r|$, needed to obtain a uniform bound for $|\partial \nu/\partial r|$. Thus, we get a uniform bound for $|\partial \nu/\partial t|$ using just (3.6).

The less trivial step is to establish uniform bounds on $|\partial \sigma/\partial t|$ and $|\dot{\rho}_N|$, needed to complete the proof of compactness. Let $z(r,t) = \partial \sigma/\partial t$. Differentiating $\Delta \sigma = f(\sigma)\nu$

w.r.t. time, we obtain

$$\frac{\partial^2 z}{\partial r^2} + \frac{1}{r}\frac{\partial z}{\partial r} = f'(\sigma) z \nu + f(\sigma)\frac{\partial \nu}{\partial t}\,. \tag{3.42}$$

At a possible maximum of z in (r_0, ρ_N), it must be $\partial z/\partial r = 0$, $\partial^2 z/\partial r^2 \le 0$, so that from (3.42) it follows $f'(\sigma) z\nu \le -f(\sigma)(\partial \nu/\partial t)$. Denoting by $\bar{z}_{max}$ the value of z at such local maximum, and being that $f_{max} = \max_\sigma f(\sigma)$ and $f'_{min} = \min_\sigma f'(\sigma)$, the previous inequality gives

$$\bar{z}_{max} \le -\frac{f(\sigma)}{f'(\sigma)}\frac{1}{\nu}\frac{\partial \nu}{\partial t} \le \frac{f_{max}}{f'_{min}}\frac{1}{N_1}\left\|\frac{\partial \nu}{\partial t}\right\|. \tag{3.43}$$

At a possible minimum of z in (r_0, ρ_N), it must be that

$$\bar{z}_{min} \ge -\frac{f(\sigma)}{f'(\sigma)}\frac{1}{\nu}\frac{\partial \nu}{\partial t} \ge -\frac{f_{max}}{f'_{min}}\frac{1}{N_1}\left\|\frac{\partial \nu}{\partial t}\right\|, \tag{3.44}$$

where $\bar{z}_{min}$ is the value of z at such local minimum.

When $\sigma(\rho_N(t), t) = \sigma_N$, since $z = 0$ for $r = r_0$ and $r = \rho_N(t)$, we can conclude that

$$\left|\frac{\partial \sigma}{\partial t}\right| \le \frac{f_{max}}{f'_{min}}\frac{1}{N_1}\left\|\frac{\partial \nu}{\partial t}\right\|. \tag{3.45}$$

Thus, from (3.13) written for $\rho_N(t)$, we get a uniform bound for $|\dot{\rho}_N|$.

When ρ_N is a material surface, i.e., $\dot{\rho}_N = u(\rho_N, t-\theta)$, the desired estimate for $|\dot{\rho}_N|$ is provided by the uniform boundedness of u. Moreover, since $\partial\sigma/\partial r|_{r=\rho_N(t)} = 0$, differentiating w.r.t. time we obtain

$$\left.\frac{\partial^2 \sigma}{\partial r^2}\right|_{r=\rho_N(t)} \dot{\rho}_N + \left.\frac{\partial^2 \sigma}{\partial r \partial t}\right|_{r=\rho_N(t)} = 0\,. \tag{3.46}$$

Also, we know that

$$\left.\frac{\partial^2 \sigma}{\partial r^2}\right|_{r=\rho_N(t)} = [f(\sigma)\nu]\Big|_{r=\rho_N(t)} \equiv g(t)\,, \tag{3.47}$$

which is positive and bounded. Therefore, at each step, we can construct the solution of the problem

$$\Delta z = \frac{\partial}{\partial t}[f(\sigma)\nu]\,, \qquad z(r_0, t) = 0\,, \quad z_r(\rho_N, t) = -g\dot{\rho}_N\,. \tag{3.48}$$

The solution will satisfy

$$z = -\dot{\rho}_N \rho_N g \log\frac{r}{r_0} - \int_{r_0}^{r} \frac{1}{r'}\left(\int_{r'}^{\rho_N} r''\frac{\partial}{\partial t}[f(\sigma)\nu]\,dr''\right)dr'$$

$$= -\dot{\rho}_N \rho_N g \log\frac{r}{r_0} - \int_{r_0}^{\rho_N} r' \log\frac{\min[r', r]}{r_0}\frac{\partial}{\partial t}[f(\sigma)\nu]\,dr'\,. \tag{3.49}$$

Let us suppose that $\dot\rho_N<0$. Then, from (3.47) and (3.48), $z_r(\rho_N,t)>0$ and z may have an absolute maximum at $r=\rho_N$. If $z_{max}=z(\rho_N)$, from (3.49) and taking into account (3.44) we obtain

$$z_{max}\le -\dot\rho_N\rho_N g\log\frac{\rho_N}{r_0}+f_{max}\left\|\frac{\partial\nu}{\partial t}\right\|\int_{r_0}^{\rho_N} r'\log\frac{r'}{r_0}\,dr'\left(1+\frac{f'_{max}}{f'_{min}}\frac{N_2}{N_1}\right), \tag{3.50}$$

which, together with (3.43)–(3.44), guarantees the uniform boundedness of $|\partial\sigma/\partial t|$. If $\dot\rho_N>0$, $z_r(\rho_N,t)<0$ and z may have an absolute minimum at $r=\rho_N$. If $z_{min}=z(\rho_N)$, we can obtain similarly

$$z_{min}\ge -\dot\rho_N\rho_N g\log\frac{\rho_N}{r_0}-f_{max}\left\|\frac{\partial\nu}{\partial t}\right\|\int_{r_0}^{\rho_N} r'\log\frac{r'}{r_0}\,dr'\left(1+\frac{f'_{max}}{f'_{min}}\frac{N_2}{N_1}\right), \tag{3.51}$$

leading to the parallel conclusion about the lower bound. If $\dot\rho_N=0$, z may have either a maximum or a minimum at $r=\rho_N$. In such cases (3.50) or (3.51) applies, still confirming the boundedness of $|\partial\sigma/\partial t|$.

From the above estimates, it also follows that $u(r,t)$ is uniformly Lipschitz continuous. Finally, concerning $\tilde c$, we can say that we have uniform inner Schauder estimates on both sides of the discontinuity curve of the consumption term (see [17, Chap. 4]). As a matter of fact, the Lipschitz continuity of $\dot B(t)$ allows us to extend such an estimate to the outer boundary $r=B(t)$. Therefore, uniform estimates for c in the norm $H^{2+\alpha,1+\alpha/2}$ are available in all the domains whose closure does not touch $r=r_0$ or the interface $r=\rho_N(t)$. □

Now we can prove Theorem 3.1.

Proof of Theorem 3.1. Let us indicate here by the subscript "n" the approximation of order n. The existence can be established simply thanks to Lemma 3.3, which provides enough compactness of the family of approximating solutions. Indeed, we can select a subsequence of indices, say, $\{n_k\}$, for which we have uniform convergence of $\rho_{N_{n_k}}, B_{n_k}$ to ρ_N, B, and of the functions ν_{n_k} to ν in $r_0\le r\le\rho_N(t)$ and of σ_{n_k}, c_{n_k} to σ, c in $r_0\le r\le B(t)$, $t\in[0,T]$. We notice that the convergence of the approximations ν_{n_k} has to be intended as

$$\lim_{k\to\infty}\sup_{(r,t)\in D_{n_k}\cap D}|\nu_{n_k}(r,t)-\nu(r,t)|=0\,,$$

where $D_n=\{(r,t):\ r\in[r_0,\rho_{N_n}(t)],t\in[0,T]\}$ and $D=\{(r,t):\ r\in[r_0,\rho_N(t)],t\in[0,T]\}$. In turn, through (3.16), this implies the uniform convergence of the corresponding sequence u_{n_k} (and $\partial u_{n_k}/\partial r$). Although the constraints (1.13) and (1.22) in the approximating scheme are not used as written, but rather with a time shift, and moreover they can be applied with the respective tolerances u_{tol}/n and σ_{tol}/n, it is clear that the correct inequalities are obtained in the limit. Similarly we have the uniform convergence of the characteristic lines and we can pass to the limit in equation (3.26). At the same time we can pass to the limit in (3.11)–(3.12) (with $\tilde\rho_N(t)=\rho_N(t)$) or (3.14)–(3.15), showing that the limit functions satisfy the same equations, so that in particular the limit ρ_N preserves the properties characterizing the boundary of the necrotic region. Thus we see that the limit functions ν, σ, u satisfy the governing equations of the model in their integral form. From the integral form we can go back to the original differential statement of the problem, just performing the derivatives and checking that all the governing differential equations, as well as the initial and boundary conditions, are satisfied. Concerning c, the Schauder estimates allow us to

pass to the limit directly in the parabolic differential equation, separately in P∪T∪Q and in N, while the differential equation is satisfied in the whole domain in the sense of $W_q^{2,1}$ for any $q>1$, thus guaranteeing the Hölder continuity of $\partial c/\partial r$. Therefore, any convergent subsequence in the family of approximating solutions provides a solution to the original problem. □

The approximating procedure previously described becomes constructive if we can say that the whole sequence is convergent. In turn, this is guaranteed if we prove uniqueness. First we prove the following property of the solution.

LEMMA 3.6. *The solution ν, for $r\in[r_0,\rho_N(t)]$, satisfies the inequalities*

$$0 < N_1 \le \nu(r,t) \le 1\,, \quad t\in[0,T]\,, \tag{3.52}$$

where N_1 is defined by (3.25).

Proof. We observe preliminarily that, as we did for the approximating solutions, we can also define for the actual solution the function $\hat r=\zeta(r,t)$ for $r\in[r_0,\rho_N(t)]$ and $t\in[0,T]$, such that $r=\eta(\zeta(r,t),t)$, with $\eta(\hat r,t)$ being the characteristic line starting from $\hat r$. In particular, we can interpret (1.11) as $D_u\nu=-\nu H(t,c,\sigma,\nu)$ on the characteristic lines (including $r=r_0$), and we can see that $D_u\nu<0$ at all points where $\nu>1$. Since $\nu_0\le 1$ everywhere, this implies that ν cannot take values greater than 1. Since the lower bound N_1 holds for all the approximations of ν, it will also hold for their limit. □

Let us pass to the following proof of uniqueness.

Proof of Theorem 3.2 We notice that, for a certain time interval starting from $t=0$, the difference $u(\rho_N,t)-\dot\rho_N(t)$ is positive for all possible solutions (the argument is the same one we applied for the approximate solutions during the first time step). Therefore we start comparing two solutions of this type.

Let us consider two possible solutions of type (1.18)–(1.19) in a time interval $(0,\hat t\,)$, $\hat t\le T$, and let us denote by $\delta\nu$, δu, $\delta\sigma$, $\delta\rho_N$, δc, δB the differences of the respective quantities. Using the labels 1 and 2 for the two solutions (so that $\delta\nu(r,t)=\nu_1(r,t)-\nu_2(r,t)$, and so on) and setting $\rho_{min}(t)=\min[\rho_{N_1}(t),\rho_{N_2}(t)]$, $\rho_{max}(t)=\max[\rho_{N_1}(t),\rho_{N_2}(t)]$, we have the following equation (where μ_1, μ_{R_1}, μ_{C_1}, χ_1 mean that the quantities are evaluated for solution 1, and the overbar means that the derivative is computed at a suitable point between the values of the independent variables for solutions 1 and 2):

$$\begin{aligned} &D_{u_1}\delta\nu+\big[\mu_1+\mu_{R_1}+\mu_{C_1}-(\chi_1+\mu_N)+(\chi_1+\mu_N)(\nu_1+\nu_2)\big]\delta\nu+\frac{\partial\nu_2}{\partial r}\delta u\\ &+\nu_2\left[\overline{\frac{\partial\mu}{\partial\sigma}}+\overline{\frac{\partial\mu_R}{\partial\sigma}}+\overline{\frac{\partial\mu_C}{\partial\sigma}}-\overline{\chi}'(1-\nu_2)\right]\delta\sigma+\nu_2\overline{\frac{\partial\mu_C}{\partial c}}\delta c=0 \end{aligned} \tag{3.53}$$

with zero initial condition and with $\delta\nu$ continued in (ρ_{min},ρ_{max}) as $(-1)^{j+1}\nu_j$, with $j=1$ if $\rho_{N_1}\ge\rho_{N_2}$ and $j=2$ otherwise. Of course $\delta\nu=0$ in the intersection of the two regions N_1, N_2. Moreover we have

$$r\delta u=\begin{cases}\displaystyle\int_{r_0}^{r} r'\big[\delta\nu(\chi_1+\mu_N)+\nu_2\overline{\chi}'\delta\sigma\big]\,dr'\,, & r\in[r_0,\rho_{min}],\\ \displaystyle\int_{r_0}^{\rho_{min}} r\big[\delta\nu(\chi_1+\mu_N)+\nu_2\overline{\chi}'\delta\sigma\big]\,dr+(-1)^{j+1}\int_{\rho_{min}}^{r} r'\big[\nu_j(\chi_j+\mu_N)-\mu_N\big]\,dr'\\ \qquad +(-1)^{j+1}\frac{\tilde\mu_N}{2}(r^2-\rho_{min}^2)\,, & r\in(\rho_{min},\rho_{max}]\,.\end{cases} \tag{3.54}$$

Taking into account that for σ_i, $i=1,2$, we have

$$\sigma_i(r,t)-\sigma_N=\int_r^{\rho_{N_i}(t)} r'\log\frac{r'}{r}f(\sigma_i)\nu_i\,dr'\,, \tag{3.55}$$

$$\sigma^\star-\sigma_N=\int_{r_0}^{\rho_{N_i}(t)} r\log\frac{r}{r_0}f(\sigma_i)\nu_i\,dr\,, \tag{3.56}$$

we obtain from (3.56)

$$(-1)^j\int_{\rho_{min}}^{\rho_{max}} r\log\frac{r}{r_0}f(\sigma_j)\nu_j\,dr=\int_{r_0}^{\rho_{min}} r\log\frac{r}{r_0}[f(\sigma_1)\nu_1-f(\sigma_2)\nu_2]\,dr\,. \tag{3.57}$$

From (3.55), for $r\in(r_0,\rho_{min}]$ we have

$$\begin{aligned}\delta\sigma=(-1)^{j+1}&\int_{\rho_{min}}^{\rho_{max}} r'\log\frac{r'}{r}f(\sigma_j)\nu_j\,dr'\\ &+\int_r^{\rho_{min}} r'\log\frac{r'}{r}[f(\sigma_1)\nu_1-f(\sigma_2)\nu_2]\,dr'\,,\end{aligned} \tag{3.58}$$

whereas, for $r\in(\rho_{min},\rho_{max}]$,

$$\delta\sigma=(-1)^{j+1}\int_r^{\rho_{max}} r'\log\frac{r'}{r}f(\sigma_j)\nu_j\,dr'\,. \tag{3.59}$$

Because there exists a positive lower bound of the product $f\nu$, a lower estimate of the left-hand side of (3.57) can be written as $(f\nu)_{min}\,\rho_{min}\log(\rho_{min}/r_0)|\delta\rho_N|$. Since (3.56) gives a lower estimate for ρ_N (as in Lemma 3.2), from (3.57) we get

$$|\delta\rho_N|\le K_1\int_{r_0}^{\rho_{min}} r\big(\|f\|\,|\delta\nu|+\|f'\|\,|\delta\sigma|\big)\log\frac{r}{r_0}\,dr, \tag{3.60}$$

where K_1 is a known constant. From (3.58)–(3.59), we see that for $r\in(r_0,\rho_{max}]$

$$|\delta\sigma|\le\int_{\rho_{min}}^{\rho_{max}} r\log\frac{r}{r_0}f(\sigma_j)\nu_j\,dr+\int_r^{\rho_{min}} r'\log\frac{r'}{r}|f(\sigma_1)\nu_1-f(\sigma_2)\nu_2|\,dr'\,,$$

and, taking into account (3.57) and that $\nu\le1$ (see Lemma 3.4), we obtain

$$|\delta\sigma|\le2\|f'\|\int_{r_0}^{\rho_{min}} r\log\frac{r}{r_0}|\delta\sigma|\,dr+2\|f\|\int_{r_0}^{\rho_{min}} r\log\frac{r}{r_0}|\delta\nu|\,dr\,. \tag{3.61}$$

Provided that

$$2\|f'\|\int_{r_0}^{\rho_{min}} r\log\frac{r}{r_0}\,dr=\|f'\|\left[\rho_{min}^2\log\frac{\rho_{min}}{r_0}-\frac{1}{2}(\rho_{min}^2-r_0^2)\right]<1\,, \tag{3.62}$$

as guaranteed by hypothesis (3.1), from (3.61) we obtain

$$\|\delta\sigma\|_t\le K_2\|\delta\nu\|_t\,, \tag{3.63}$$

where $\|\cdot\|_t$ means the sup w.r.t. r and to time in $[0,t]$. Concerning the equation for δc, to be satisfied in $r_0<r<B_{min}(t)$, $0<t<\hat{t}$, B_{min} denoting $\min[B_1,B_2]$, we have

$$\frac{\partial \delta c}{\partial t} - D_C \Delta \delta c = -\varphi_{C_1} \nu^\star \delta\nu - \left(\overline{\frac{\partial \varphi_C}{\partial c}}\right)\nu^\star \nu_2 \delta c - \left(\overline{\frac{\partial \varphi_C}{\partial \sigma}}\right)\nu^\star \nu_2 \delta\sigma - \lambda \delta c\,, \tag{3.64}$$

$$\delta c(r_0,t) = 0\,, \tag{3.65}$$

$$\delta c(r,0) = 0\,. \tag{3.66}$$

If $\hat{t}$ is sufficiently small to guarantee that the differences $B_i - \rho_{N_j}$, $i=1,2$, $j=1,2$, remain strictly positive, we also have

$$\left.\frac{\partial \delta c}{\partial r}\right|_{r=B_{min}(t)} = (-1)^{k+1} \left.\frac{\partial^2 c_k}{\partial r^2}\right|_{r=\bar{r}(t)} |\delta B|\,, \tag{3.67}$$

where $\bar{r}(t)$ is a suitable point between the boundaries $B_1(t)$, $B_2(t)$, and $k=1$ if $B_1 \geq B_2$ and $k=2$ otherwise. The coefficient of $|\delta B|$ in (3.67) is a priori bounded since c_1,c_2 possess the same properties stated for the approximating functions in Lemma 3.3.

Now we are able to write the following chain of inequalities, starting with

$$\|\delta\nu\|_t \leq \int_0^t \left(k_1 \|\delta u\|_\tau + k_2 \|\delta\sigma\|_\tau + k_3 \|\delta c\|_\tau\right) d\tau, \tag{3.68}$$

that can be obtained from (3.53). From (3.54) we see that $\|\delta u\|_\tau$ can be estimated in terms of $\|\delta\rho_N\|_\tau$, $\|\delta\sigma\|_\tau$, $\|\delta\nu\|_\tau$ and, ultimately, because of (3.60) and (3.63), in terms of $\|\delta\nu\|_\tau$. So (3.68) implies

$$\|\delta\nu\|_t \leq \int_0^t \left(k_3 \|\delta c\|_\tau + k_4 \|\delta\nu\|_\tau\right) d\tau\,. \tag{3.69}$$

Going back to problem (3.64)–(3.67), for which an estimate of $\partial^2 c_k/\partial r^2$ is available in the region between B_{min} and $B_{max}=\max[B_1,B_2]$, taking into account that $|\delta B| \leq \int_0^t \|\delta u\|_\tau \, d\tau$, and exploiting the estimates already used for $\delta\sigma$ and δu, we obtain by classical means the inequality

$$\|\delta c\|_t \leq \int_0^t \left(k_5 \|\delta\nu\|_\tau + k_6 \|\delta c\|_\tau\right) d\tau, \tag{3.70}$$

which, together with (3.69), immediately yields $\|\delta c\|_t = \|\delta\nu\|_t = 0$. Thus we may conclude that the solution is unique in a suitably small time interval, and by extension up to a possible time point $\bar{t}$ such that in any right neighborhood (1.18)–(1.19) cannot hold.

Let us now suppose that after $\bar{t}$ we have in some interval two solutions for which the interface $r=\rho_N(t)$ is a material surface. Arguments similar to those seen above can be repeated, taking into account that in the present case we have ($i=1,2$)

$$\sigma_i(r,t) - \sigma_i(\rho_{N_i}(t),t) = \int_r^{\rho_{N_i}(t)} r' \log\frac{r'}{r} f(\sigma_i)\nu_i \, dr'\,, \tag{3.71}$$

$$\sigma^\star - \sigma_i(\rho_{N_i}(t),t) = \int_{r_0}^{\rho_{N_i}(t)} r \log \frac{r}{r_0} f(\sigma_i)\nu_i \, dr \,, \tag{3.72}$$

implying

$$\sigma_i - \sigma^\star = \int_{r}^{\rho_{N_i}} r' \log \frac{r'}{r} f(\sigma_i)\nu_i \, dr' - \int_{r_0}^{\rho_{N_i}} r \log \frac{r}{r_0} f(\sigma_i)\nu_i \, dr \,. \tag{3.73}$$

Moreover, since $\rho_{N_1}(\bar t)=\rho_{N_2}(\bar t)$, we have for $t>\bar t$

$$\delta\rho_N = \int_{\bar t}^{t} [u_1(\rho_{N_1}(\tau),\tau) - u_2(\rho_{N_2}(\tau),\tau)] \, d\tau \,; \tag{3.74}$$

thus

$$|\delta\rho_N| \le \int_{\bar t}^{t} \left(|\delta u|_{r=\rho_{min}(\tau)} + \left\| \frac{\partial u}{\partial r} \right\| |\delta\rho_N(\tau)| \right) d\tau \,. \tag{3.75}$$

Now, using easy estimates on δu, $\delta\sigma$, $\delta\nu$, and the boundedness of $|\partial u/\partial r|$, we can infer uniqueness until $r=\rho_N(t)$ is a material boundary. The procedures previously described can be applied after each switch to solutions having the interface ρ_N of the same type (both nonmaterial or both material).

We observe now that the evolutive problem in which constraint (1.13) is not imposed, and ρ_N and σ are defined by (1.15), (1.17), (1.19)–(1.20), cannot have more than one solution (the comparison technique is the same as the one used above). The same holds for the evolutive problem in which constraint (1.22) is not imposed, ρ_N is defined by (1.21), and σ is defined by (1.15), (1.17), and (1.20). Therefore, we can exclude that after $\bar t$, or any other switching point, there can be a time interval in which our problem has a solution of one type and another solution of different type. This, in fact, would imply that two different unconstrained solutions exist in a time interval after $\bar t$, since it is the behavior of such unconstrained solutions that governs the switch of ρ_N from one type to the other.

Hence, it remains only to examine the case of two solutions having infinitely many switching points in any right neighborhood of the time $\bar t$ (the reader can observe that the comparison between two solutions of different type after $\bar t$, which we avoided on the basis of the argument above, is de facto included in the analysis that follows). The argument proceeds in a similar way, the main difference occurring in the comparison of σ_1,σ_2 and ρ_{N_1},ρ_{N_2}. Let us consider an interval $(\bar t, \bar t+\varepsilon)$ and its partition in intervals in which the two solutions are both of the same type and in intervals in which they are of different type. We fix our attention on the second class of intervals and, to be specific, let us assume that ρ_N is nonmaterial for solution 1 and material for solution 2. Subtracting (3.72) with $i=2$ from (3.56) with $i=1$, we get

$$\sigma_2(\rho_{N_2}(t),t) - \sigma_N = \int_{r_0}^{\rho_{min}} r \log \frac{r}{r_0} [f(\sigma_1)\nu_1 - f(\sigma_2)\nu_2] \, dr$$

$$+(-1)^{j+1} \int_{\rho_{min}}^{\rho_{max}} r \log \frac{r}{r_0} f(\sigma_j)\nu_j \, dr, \tag{3.76}$$

which replaces (3.57). Thus we may derive an inequality similar to (3.60):

$$|\delta\rho_N| \leq \quad K_1 \int_{r_0}^{\rho_{min}} r\big(\|f\|\,|\delta\nu| + \|f'\|\,|\delta\sigma|\big)\log\frac{r}{r_0}\,dr$$
$$+ K_1|\sigma_2(\rho_{N_2}(t),t)-\sigma_N|\,. \tag{3.77}$$

Now we can write

$$|\sigma_2(\rho_{N_2},t)-\sigma_N| \leq |\delta\sigma(\rho_{N_2},t)| + \sigma_1(\rho_{N_2},t)-\sigma_N\,, \tag{3.78}$$

if, e.g., $\rho_{N_1} > \rho_{N_2}$. Moreover,

$$\sigma_1(\rho_{N_2},t)-\sigma_N = \left|\frac{\partial\sigma_1}{\partial r}(\bar r,t)\right| |\delta\rho_N| \tag{3.79}$$

with $\bar r$ between ρ_{N_2} and ρ_{N_1}. Since $\partial\sigma_1/\partial r$ vanishes for $r=\rho_{N_1}$, we can also say that

$$\left|\frac{\partial\sigma_1}{\partial r}(\bar r,t)\right| \leq \left\|\frac{\partial^2\sigma_1}{\partial r^2}\right\| |\delta\rho_N|\,.$$

If, instead, $\rho_{N_1} < \rho_{N_2}$, we have similar inequalities with the indices interchanged. Thus from (3.78) we deduce

$$|\sigma_2(\rho_{N_2},t)-\sigma_N| \leq |\delta\sigma(\rho_{N_2},t)| + C|\delta\rho_N|^2 \tag{3.80}$$

with $C>0$ known a priori. On the other hand, we may take ε so small that in $(\bar t, \bar t+\varepsilon)$ we have $K_1C|\delta\rho_N|<1/2$, thanks to the fact that for both solutions we know an upper bound for $|\dot\rho_{N_i}|$. Thus, we may rewrite (3.77) in the form

$$|\delta\rho_N| \leq 2K_1 \int_{r_0}^{\rho_{min}} r\big(\|f\|\,|\delta\nu| + \|f'\|\,|\delta\sigma|\big)\log\frac{r}{r_0}\,dr + 2K_1|\delta\sigma(\rho_{N_2}(t),t)|\,. \tag{3.81}$$

We must now replace (3.58) by

$$\delta\sigma + \sigma_2(\rho_{N_2}(t),t)-\sigma_N = (-1)^{j+1}\int_{\rho_{min}}^{\rho_{max}} r'\log\frac{r'}{r} f(\sigma_j)\nu_j\,dr'$$

$$+ \int_r^{\rho_{min}} r'\log\frac{r'}{r}[f(\sigma_1)\nu_1 - f(\sigma_2)\nu_2]\,dr'\,. \tag{3.82}$$

Combining (3.82) and (3.76) we obtain the same inequality (3.61), eventually implying (3.63) with the norm referring to the appropriate time interval.

For the first class of intervals, i.e., the intervals in which the two solutions are both of the same type, the estimates for $|\delta\sigma|$ and $|\delta\rho_N|$ are the same ones we have already used, except for a small change concerning the comparison of solutions which both have a material ρ_N interface. Indeed, in (3.74) the term $\delta\rho_N(\hat t\,)$ (where $\hat t$ now denotes the initial time of the interval we are considering) must be added on the right-hand side. Such a term is inherited from the previous interval in which the solutions are of different type and therefore it is expressed in the way we have just seen, that is, by means of (3.81). Therefore, we can conclude that also in this case the solution is unique. □

4. Concluding comments. As a final comment, we stress that some of the simplifying assumptions made in the development of the present model could be relaxed with some further refinements. For simplicity, we have taken $\sigma^\star$ constant here, but the whole theory can be extended to the case of $\sigma^\star$ variable with the time, provided it remains strictly above σ_P. We could also consider a flux condition instead of (1.17). Similarly, condition (1.28) has no crucial role in the treatment and could be replaced with a flux condition. Finally, a dependence of cell proliferation on treatment could be taken into account by representing the proliferation rate as a function $\chi(\sigma, c, t)$. The numerical solution of the evolutive problem under various therapeutic treatment modalities will be presented in a forthcoming paper [6]. The numerical results show that the switching from a nonmaterial ρ_N interface to a material interface, and vice versa, does occur under usual treatment modalities.

REFERENCES

[1] J. A. Adam and S. A. Maggelakis, *Diffusion regulated growth characteristics of a spherical prevascular carcinoma*, Bull. Math. Biol., 52 (1990), pp. 549–582.

[2] D. Ambrosi and L. Preziosi, *On the closure of mass balance models for tumor growth*, Math. Models Methods Appl. Sci., 12 (2002), pp. 737–754.

[3] A. Bertuzzi and A. Gandolfi, *Cell kinetics in a tumour cord*, J. Theoret. Biol., 204 (2000), pp. 587–599.

[4] A. Bertuzzi, A. Fasano, and A. Gandolfi, *A mathematical model for the growth of tumour cords incorporating the dynamics of a nutrient*, in Free Boundary Problems: Theory and Applications II, GAKUTO Internat. Ser. Math. Sci. Appl. 14, N. Kenmochi, ed., Gakkōtosho, Tokyo, 2000, pp. 31–46.

[5] A. Bertuzzi, A. Fasano, A. Gandolfi, and D. Marangi, *Cell kinetics in tumor cords studied by a model with variable cell cycle length*, Math. Biosci., 177/178 (2002), pp. 103–125.

[6] A. Bertuzzi, A. d'Onofrio, A. Fasano, and A. Gandolfi, *Regression and regrowth of tumour cords following single-dose anticancer treatment*, Bull. Math. Biol., 65 (2003), pp. 903–931.

[7] C. J. W. Breward, H. M. Byrne, and C. E. Lewis, *Modelling the interactions between tumour cells and a blood vessel in a microenvironment within a vascular tumour*, European J. Appl. Math., 12 (2001), pp. 529–556.

[8] T. P. Butler, F. H. Grantham, and P. M. Gullino, *Bulk transfer of fluid in the interstitial compartment of mammary tumors*, Cancer Res., 35 (1975), pp. 3084–3088.

[9] J. J. Casciari, S. V. Sotirchos, and R. M. Sutherland, *Variations in tumor cell growth rates and metabolism with oxygen concentration, glucose concentration, and extracellular pH*, J. Cell Physiol., 151 (1992), pp. 386–394.

[10] M. A. J. Chaplain, *From mutation to metastasis: The mathematical modelling of the stages of tumor development*, in A Survey of Models for Tumor-Immune System Dynamics, J. A. Adam and N. Bellomo, eds., Birkhäuser, Boston, 1996, pp. 187–236.

[11] S. Cui and A. Friedman, *Analysis of a mathematical model of the growth of necrotic tumors*, J. Math. Anal. Appl., 255 (2001), pp. 636–677.

[12] J. Dyson, R. Villella-Bressan, and G. Webb, *The steady state of a maturity structured tumor cord cell population*, Discrete Contin. Dynam. Systems B, 4 (2004), pp. 115–134.

[13] A. Friedman and F. Reitich, *Analysis of a mathematical model for the growth of tumors*, J. Math. Biol., 38 (1999), pp. 262–284.

[14] H. P. Greenspan, *Models for the growth of a solid tumor by diffusion*, Studies in Appl. Math., 52 (1972), pp. 317–340.

[15] D. G. Hirst and J. Denekamp, *Tumour cell proliferation in relation to the vasculature*, Cell Tissue Kinet., 12 (1979), pp. 31–42.

[16] A. H. Kyle and A. I. Minchinton, *Measurement of delivery and metabolism of tirapazamine to tumour tissue using the multilayered cell culture model*, Cancer Chemother. Pharmacol., 43 (1999), pp. 213–220.

[17] O. A. Ladyženskaja, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, Translations of Mathematical Monographs 23, American Mathematical Society, Providence, RI, 1967.

[18] J. V. Moore, P. S. Hasleton, and C. H. Buckley, *Tumour cords in* 52 *human bronchial and cervical squamous cell carcinomas: Inferences for their cellular kinetics and radiobiology*, Br. J. Cancer, 51 (1985), pp. 407–413.

[19] I. F. Tannock, *The relation between cell proliferation and the vascular system in a transplanted mouse mammary tumour*, Br. J. Cancer, 22 (1968), pp. 258–273.

[20] G. F. Webb, *The steady state of a tumor cord cell population*, J. Evol. Equ., 2 (2002), pp. 425–438.

SIAM J. MATH. ANAL.
Vol. 36, No. 3, pp. 916–966

HIGHER DIMENSIONAL SLEP EQUATION AND APPLICATIONS TO MORPHOLOGICAL STABILITY IN POLYMER PROBLEMS*

YASUMASA NISHIURA† AND HIROMASA SUZUKI‡

Dedicated to Professors Masayasu Mimura and Takaaki Nishida on their sixtieth birthdays

Abstract. Existence and stability of stationary internal layered solutions to a rescaled diblock copolymer equation are studied in higher dimensional space. Rescaling is necessary since the characteristic domain size of any stable pattern eventually vanishes in an appropriate singular limit. A general sufficient condition for the existence of singularly perturbed solutions and the associated stability criterion are given in the form of linear operators acting only on the limiting location of the interface. Applying the results to radially symmetric and planar patterns, we can show, for instance, stability of radially symmetric patterns when one of the components of diblock copolymer dominates the other, and that of the long-striped pattern in a long and narrow domain for the planar case. These results are consistent with the experimental ones. The above existence and stability criterion can be easily extended to a class of reaction diffusion systems of activator-inhibitor type.

Key words. singular perturbation, diblock copolymer, reaction diffusion system, pattern formation, matched asymptotic expansion, stability, critical eigenvalues

AMS subject classification. 35B25, 35B35, 35K57

DOI. 10.1137/S0036141002420157

1. Introduction.

In this paper, we shall discuss two basic problems of singularly perturbed solutions to the fourth order equation (1.1) arising in modelling the dynamics of diblock copolymer melts in higher dimensional space, namely *Under what conditions existence of stationary singularly perturbed solution is guaranteed?* and *How is the stability of it?*. The former is related to the possible morphological forms, and the latter is crucial for observability. We shall see that both problems are eventually reduced to solving the equations on the *interface* in the singular limit. Especially, the spectral problem on the interface called the SLEP equation in higher dimensional space (see (1.13) and (4.22)), is a generalization of a one-dimensional version for reaction diffusion systems discussed in [11] and [10]. The model equation (1.1) below looks very special; however, the method employed here is quite general and can be extended to the activator-inhibitor systems treated in [11] and [10], in fact the mechanisms causing pattern formation have common features in both models. Our model takes the following form:

*Received by the editors December 27, 2002; accepted for publication (in revised form) December 12, 2003; published electronically November 17, 2004. This research was partially supported by the Grant-in-Aid for Scientific Research 11740060, 12440026, 13440027, the Japanese Society of the Promotion of Science.

http://www.siam.org/journals/sima/36-3/42015.html

†Laboratory of Nonlinear Studies and Computations, Research Institute for Electronic Science, Hokkaido University, Kita-ku, Sapporo, 060-0812, Japan (nishiura@aurora.es.hokudai.ac.jp).

‡Faculty of Education, Shiga University, Hiratsu, Otsu, 520-0862, Japan (kodachi@sue.shiga-u.ac.jp).

$$
\begin{cases}
u_t = -\Delta\{d^2\Delta u + f(u) - \sigma(-\Delta_N)^{-1}(u-\overline{u})\} \\
\qquad = -\Delta\{d^2\Delta u + f(u)\} - \sigma(u-\overline{u}), & (X,t) \in \hat{\Omega}\times(0,\infty), \\
\dfrac{\partial u}{\partial n} = \dfrac{\partial(\Delta u)}{\partial n} = 0, & (X,t) \in \partial\hat{\Omega}\times(0,\infty), \\
\dfrac{1}{|\hat{\Omega}|}\displaystyle\int_{\hat{\Omega}} u dX = \overline{u},
\end{cases} \tag{1.1}
$$

where u is the order parameter which represents the ratio of two homopolymers, f is basically a bistable nonlinearity (typically $u-u^3$), d and σ are positive constants, Δ_N is the Laplace operator under the Neumann boundary condition, and $\hat{\Omega}$ is a smooth bounded domain in $\mathbf{R}^N$ ($N \geq 2$). Originally, the model system was given by the energy functional form in [16] (see also [1]), and then reformulated in the above form by [12], which triggered many interesting rigorous works such as [25] and [26]. The associated Euler–Lagrange equation with the functional is given by (1.1). Diblock copolymer is a chain where two different homopolymers are connected, and the connectivity causes a long range interaction. In fact, its effect is reflected by the nonlocal term $\sigma(-\Delta_N)^{-1}(u-\overline{u})$, where σ is proportional to the inverse of the polymerization index (i.e., length of the chain). Due to the nonlocal term $\sigma(-\Delta_N)^{-1}(u-\overline{u})$, (1.1) displays a variety of stationary *mesoscopic* patterns including lamellar, column, spherical, double gyroid morphologies and so on (see, for instance, [24] and the reference therein). Here the mesoscopic means an intermediate scale between micro and macro, and it becomes very fine as d tends to zero as in the following proposition. This makes a sharp contrast with the case $\sigma \equiv 0$, i.e., Cahn–Hilliard equation. For experimental observation, see [6, 7, 8]. Note also that the nonlocal effect is similar to the role of the inhibitor field for activator-inhibitor system, and hence our analysis basically encompasses such a system as we shall see in section 5.

In order to have a reasonable singular limit, some sort of scaling law is necessary and we adopt the following one, which was originally proposed in [12] in the above setting and was proved rigorously by [14] for one-dimensional case and by [2] for higher dimensional case.

PROPOSITION 1.1. *(Theorem* 3.2 in [12]). *In order to have a well-defined limiting stationary problem of* (1.1) *as* $d \downarrow 0$, *which is independent of parameters* d *and* σ, *there exists a unique scaling with respect to space and time given by:* $x := X/(d^{1/3}\sigma^{-1/3})$, $\tau := \sigma t$. *The characteristic domain size is proportional to* $(d/\sigma)^{1/3}$ *and the morphology of pattern is determined by solving* (1.7)–(1.9) (*see assumption* (A4)).

This scaling law is not only a sufficient condition in order to have a well-defined singular limit problem of (1.1) as $d \downarrow 0$ (see [12] for details), but also can be justified from a view point of statistical physics (see [3]). In terms of the new variables x and τ, the equations in (1.1) are recast as

$$
u_\tau = \Delta\left\{-\epsilon\Delta u - \frac{1}{\epsilon}f(u)\right\} - (u-\overline{u}), \quad (x,\tau) \in \Omega\times(0,\infty), \tag{1.2}
$$

where $\epsilon := d^{2/3}\sigma^{1/3}$ and Ω is a magnified unit domain. Here we implicitly assume that the patterns are periodic in space (in the original scaling), and we focus on a unit cell of this periodic structure.

On the other hand, a typical activator-inhibitor system is given by

$$\begin{cases} u_t = d^2\Delta u + f(u) - v, \\ \\ v_t = D\Delta v + g(u,v), \end{cases} \qquad (x,t) \in \hat{\Omega} \times (0,\infty), \tag{1.3}$$

where typically $(f,g) = (u - u^3, u - \gamma v)(\gamma > 0)$. It was proved in [13] that if (1.3) has a d-family of stationary matched asymptotic solutions whose interface is smooth up to $d = 0$, then it must become unstable for small d. It is observed numerically that stable patterns become finer and finer when d becomes small (see also a recent work [15]), which strongly suggests the necessity for rescaling in order to track the stable patterns. It was also shown in [13] that the characteristic size of stable patterns is of order $d^{1/3}$ by formal computation. Magnifying the system (1.3) with this scaling, the rescaled system is given by

$$\begin{cases} u_t = \epsilon^2\Delta u + f(u) - v, \\ \\ \epsilon v_t = D\Delta v + \epsilon g(u,v), \end{cases} \qquad (x,t) \in \Omega \times (0,\infty), \tag{1.4}$$

where $\epsilon = d^{2/3}$. (See [4, 18, 19, 20, 21, 23].)

At first sight there is no resemblance between (1.2) and (1.4); however, rewriting (1.2) by introducing new variable $v := \epsilon^2\Delta u + f(u)$ and new time $\tilde{t} := \tau/\epsilon$, the resulting system becomes

$$\begin{cases} 0 = \epsilon^2\Delta u + f(u) - v, \\ \\ 0 = \Delta v + \epsilon(u - \overline{u}) + u_{\tilde{t}}, \end{cases} \qquad (x,\tilde{t}) \in \Omega \times (0,\infty), \tag{1.5}$$

$$\frac{\partial u}{\partial n} = 0 = \frac{\partial v}{\partial n}, \quad (x,\tilde{t}) \in \partial\Omega \times (0,\infty). \tag{1.6}$$

Now it is clear that (1.2) is similar to (1.4) as far as stationary solutions are concerned. Similarity in fact can be extended to dynamical level (i.e., stability), which will be discussed in section 5.

As a consequence of the above discussion, (1.5) is expected to have a well-defined smooth interface in the singular limit. The next issue is to locate such an interface by solving, what is called, the *reduced problem* of (1.5) (see (1.7)–(1.9)); however, it is in general quite difficult to solve (1.7)–(1.9) and obtain its explicit profile. The results listed below will be, therefore, stated under the assumption that such reduced solutions exist. Nevertheless, this assumption can be checked explicitly for simple but basic cases when Ω has a spherical or planar geometry, which will be shown in later sections.

Now we shall state the assumptions and introduce several key notation and operators for later use.

(A1) f is a smooth bistable nonlinearity. The equation $f(u) - v = 0$ has three sub-branches of solutions

$$\mathcal{C}_+ = \{(u,v) | u = h^+(v), v \in I_+\}, \ \mathcal{C}_- = \{(u,v) | u = h^-(v), v \in I_-\} \quad \text{and}$$

$$\mathcal{C}_0 = \{(u,v) | u = h^0(v), v \in I_+ \cap I_-\}$$

such that $h^-(v) < h^0(v) < h^+(v)$ for $v \in I_+ \cap I_-$.

(A2) $f_u(h^{\pm}(v)) = \frac{d}{du} f(h^{\pm}(v)) < 0$ on $I_{\pm}$.

(A3) $J(v) := \int_{h^-(v)}^{h^+(v)} [f(s) - v] ds$, $v \in I_0$, has an isolated zero $v^* \in I_0$ such that $J'(v^*) < 0$.

(A1)–(A3) with (A4′) and (A7) assumed later are the sufficient conditions for the existence and stability of transition layer solutions to one-dimensional activator-inhibitor system (1.3) (see [10] and [11]).

We assume that the domain Ω is simply connected, and Γ is either one of the following two cases.

Case I. $\Omega \in \mathbf{R}^N$ $(N \geq 2)$ has smooth boundary and define a set $\mathcal{F}$ by

$$\mathcal{F} = \big\{\Gamma \subset \Omega \mid \Gamma \text{ is an } N-1 \text{ dimensional smooth compact connected manifold without boundary}\big\}.$$

Each $\Gamma \in \mathcal{F}$ divides Ω into two connected components. We denote by $\Omega_+(\Gamma)$ the component of $\Omega \backslash \Gamma$ which has $\partial\Omega$ as part of its boundary. The other component is defined by $\Omega_-(\Gamma)$. Therefore we have

$$\partial\Omega_+(\Gamma) = \partial\Omega \cup \Gamma \quad \text{and} \quad \partial\Omega_-(\Gamma) = \Gamma.$$

Case II. $\Omega \in \mathbf{R}^2$ is a rectangle $(0, X) \times (0, Y)$. Then we define $\mathcal{F}$ by

$$\mathcal{F} = \big\{\Gamma \subset \Omega \mid \Gamma \text{ is a line, which is parallel to the } y\text{-axis and touches transversely with the boundary } \partial\Omega\big\}.$$

Then Ω is divided into two connected components $\Omega_-(\Gamma) = (0, x_0) \times (0, Y)$ and $\Omega_+(\Gamma) = (x_0, X) \times (0, Y)$, where $\Gamma = \{(x_0, y) \in \mathbf{R}^2 \mid 0 \leq y \leq Y\}$ for some $x_0 \in (0, X)$. For later use we introduce the *aspect ratio* κ of Ω, i.e., $\kappa := X/Y$.

Let us introduce a local coordinate system in the neighborhood of Γ. By the implicit function theorem there exists a $d_0 > 0$ such that the map $p : [-d_0, d_0] \times \Gamma \to \Gamma_{d_0}$ defined by

$$p(r, y) = y + r\nu(y)$$

is a diffeomorphism, where $\Gamma_{d_0} = \{x \in \Omega | \text{dist}(x, \Gamma) < d_0\}$, $\nu = \nu(y)$ is the unit normal vector on Γ $(y \in \Gamma)$ which points the interior of $\Omega_+(\Gamma)$. Using this diffeomorphism, we identify $x \in \Gamma_{d_0}$ with $(r, y) \in [-d_0, d_0] \times \Gamma_{d_0}$ and write $u(r, y)$ or $u(y + r\nu(y))$ for $u(x)$ $(x \in \Gamma_{d_0})$. We denote by $H(r, y)$ the mean curvature of the manifold

$$\Gamma(r) := \{x \in \Omega | x = y + r\nu(y) \ y \in \Gamma\}$$

at $y \in \Gamma$. Then $H(0, y)$ stands for the mean curvature of the manifold Γ at $y \in \Gamma$.

Such an interface $\Gamma \in \mathcal{F}$ is determined by solving the following equations called the *rescaled reduced problem*:

$$\begin{cases} \Delta V^- + h^-(v^*) - \overline{u} = 0 & \text{in } \Omega_-(\Gamma), \\ V^- = b_1^*(y) & \text{on } \Gamma, \end{cases} \tag{1.7}$$

$$\begin{cases} \Delta V^+ + h^+(v^*) - \overline{u} = 0 & \text{in } \Omega_+(\Gamma), \\ V^+ = b_1^*(y) & \text{on } \Gamma, \\ \dfrac{\partial V^+}{\partial n} = 0 & \text{on } \partial\Omega, \end{cases} \tag{1.8}$$

$$\frac{\partial V^+}{\partial \nu} = \frac{\partial V^-}{\partial \nu} \quad \text{on } \Gamma, \tag{1.9}$$

where

$$b_1^*(y) := \frac{m^2}{[h]}(N-1)H(0,y), \;\; m^2 := \int_{-\infty}^{\infty} [u_\xi^*(\xi)]^2 d\xi$$

and $[h]$ denotes the jump of two branches $u = h^{\pm}(v)$ at $v = v^*$, i.e.,

$$[h] := h^+(v^*) - h^-(v^*).$$

Here $u = u^*(\xi)$ is the unique solution of

$$u_{\xi\xi} + f(u) - v^* = 0, \;\; \xi \in \mathbf{R}, \;\; u(0) = h^0(v^*), \;\; \lim_{\xi \to \pm\infty} u(\xi) = h^{\pm}(v^*).$$

When Γ is a rectangle, we add the boundary condition

$$\frac{\partial V^-}{\partial n} = 0 \quad \text{on } \partial\Omega \cap \partial\Omega_-(\Gamma)$$

to (1.7) and replace the boundary condition for $\frac{\partial V^+}{\partial n}$ in (1.8) by

$$\frac{\partial V^+}{\partial n} = 0 \quad \text{on } \partial\Omega \cap \partial\Omega_+(\Gamma).$$

Roughly speaking, the interface Γ is a set on which the limiting function of the stationary solution u^ϵ as $\epsilon \downarrow 0$ becomes discontinuous in the normal direction. For more discussions on (1.7)–(1.9), see [12]. We assume the existence of a solution (V^*, Γ^*) of (1.7)–(1.9).

(A4) There exists a solution $(V^*, \Gamma^*) \in C^1(\overline{\Omega}) \times \mathcal{F}$ of (1.7)–(1.9), where

$$V^*(x) = \begin{cases} V^-(x), & x \in \Omega_-(\Gamma^*), \\ V^+(x), & x \in \Omega_+(\Gamma^*). \end{cases}$$

$V^{\pm}$ are smooth solutions of (1.7) and (1.8), respectively, satisfying (1.9).

In order to state a sufficient condition for the existence and the SLEP equation, we need to define several operators. Let us consider the following linear boundary value problem:

$$\begin{cases} \Delta Z^- = 0 & \text{in } \Omega_-(\Gamma^*), \\ Z^- = q & \text{on } \Gamma^*, \end{cases} \tag{1.10}$$

$$\begin{cases} \Delta Z^+ = 0 & \text{in } \Omega_+(\Gamma^*), \\ \dfrac{\partial Z^+}{\partial n} = 0 & \text{on } \partial\Omega, \;\; Z^+ = q \quad \text{on } \Gamma^*, \end{cases} \tag{1.11}$$

where $\Gamma^* \in \mathcal{F}$ is a solution of (1.7)–(1.9). When Γ^* is a rectangle, we add the boundary condition

$$\frac{\partial Z^-}{\partial n} = 0 \quad \text{on } \partial\Omega \cap \partial\Omega_-(\Gamma^*)$$

to (1.10) and replace the boundary condition for $\frac{\partial Z^+}{\partial n}$ in (1.11) by

$$\frac{\partial Z^+}{\partial n} = 0 \quad \text{on } \partial\Omega \cap \partial\Omega_+(\Gamma^*).$$

For $q \in C^{2,\alpha}(\Gamma^*)$ $(0 < \alpha < 1)$, the problem (1.10) and (1.11) have a unique solution denoted by $\mathcal{P}^- q$ and $\mathcal{P}^+ q$, respectively, satisfying $\mathcal{P}^\pm q \in C^{2,\alpha}(\overline{\Omega})$, and define normal derivative operator $\Pi_\pm : C^{2,\alpha}(\Gamma^*) \to C^{1,\alpha}(\Gamma^*)$ by

$$\Pi_- q = \frac{\partial}{\partial \nu}(\mathcal{P}^- q)\bigg|_{\Gamma^*}, \quad \Pi_+ q = -\frac{\partial}{\partial \nu}(\mathcal{P}^+ q)\bigg|_{\Gamma^*}.$$

Note that $\Pi := \Pi_- + \Pi_+$ is self-adjoint and the null space $\mathcal{N}(\Pi)$ is one-dimensional spanned by a constant function. Then there is a bounded operator

$$\mathcal{T} : (I - P)C^{1,\alpha}(\Gamma^*) \to (I - P)C^{2,\alpha}(\Gamma^*)$$

satisfying $\Pi\mathcal{T} = I$ on $(I - P)C^{1,\alpha}(\Gamma^*)$ and $\mathcal{T}\Pi = I - P$ on $C^{2,\alpha}(\Gamma^*)$, where P is a projection defined by $P\theta := \frac{1}{|\Gamma^*|}\int_{\Gamma^*} \theta \, dS$ for $\theta \in C^\alpha(\Gamma^*)$ and $|\Gamma^*|$ is surface area of Γ^*.

Concerning the existence of the stationary solution, we have the following theorem.

THEOREM 1.2 (existence). *Assume that* (A1)–(A4) *are satisfied. Moreover, we assume that*

(A5) *There is a bounded linear operator* $L^\dagger : (I - P)C^\alpha(\Gamma^*) \to (I - P)C^{2,\alpha}(\Gamma^*)$ *which is the inverse of the operator*

$$L := \Delta^{\Gamma^*} + H^*(y) - \frac{1}{m^2} J'(v^*) V_r^*(0, y) + \frac{1}{m^2}[h] J'(v^*)\mathcal{T}(\cdot)$$

such that $LL^\dagger = I$ *on* $(I - P)C^\alpha(\Gamma^*)$ *and* $L^\dagger L = I - P$ *on* $(I - P)C^{2,\alpha}(\Gamma^*)$. *Here,* Δ^{Γ^*} *is the Laplace–Beltrami operator on* Γ^*, $H^*(y)$ *the sum of the square of the principal curvature of* Γ^*, $V_r^*(0, y) = \frac{\partial V^*}{\partial \nu}$, *and* $[h]$ *denotes the aforementioned jump at* $v = v^*$.

Then, there is an $\epsilon_0 > 0$ *such that* (1.5)–(1.6) *have an* ϵ*-family of stationary solutions* (u^ϵ, v^ϵ) *for* $\epsilon \in (0, \epsilon_0]$ *satisfying*

(i) $\lim_{\epsilon \to 0} v^\epsilon(x) = v^*$ *uniformly on* $\overline{\Omega}$,

(ii) *for each* $\delta > 0$,

$$\lim_{\epsilon \to 0} u^\epsilon(x) = \begin{cases} h^-(v^*), & x \in \overline{\Omega}_-(\Gamma^*) \backslash \Gamma_\delta^* \\ \\ h^+(v^*), & x \in \overline{\Omega}_+(\Gamma^*) \backslash \Gamma_\delta^* \end{cases} \quad \textit{uniformly,}$$

where Γ_δ^* *is a tubular neighborhood of* Γ^*,

(iii) *for each* $K > 0$,

$$\lim_{\epsilon \to 0} u^\epsilon(y + \epsilon\xi\nu(y)) = u^*(\xi + s^*(y)) \quad \textit{in } C^2[-K, K]$$

uniformly in $y \in \Gamma^*$ *for some* $s^* \in C^{2,\alpha}(\Gamma^*)$.

The stability property of the stationary solutions is determined by the spectrum of the following associated linearized problem:

$$\begin{cases} 0 = \epsilon^2 \Delta w + f_u^\epsilon w - z, \\ \qquad\qquad\qquad\qquad\qquad \text{in } \ \Omega, \\ 0 = \Delta z + \epsilon w + \lambda^\epsilon w, \\ \dfrac{\partial w}{\partial n} = 0 = \dfrac{\partial z}{\partial n} \quad \text{on } \ \partial\Omega, \end{cases} \tag{1.12}$$

where $f_u^\epsilon := \frac{d}{du} f(u^\epsilon)$. We assume the following for (1.12).

(A6) Each eigenvalue λ^ϵ and the associated eigenfunctions (w^ϵ, z^ϵ) of (1.12) have the following asymptotic forms:

$$\begin{cases} \lambda^\epsilon = \epsilon\lambda^* + o(\epsilon), \\ w^\epsilon(x) = \displaystyle\sum_{i=0}^{2} \epsilon^i W^{\pm,i}(x) + \omega(r) \cdot \sum_{i=0}^{2} \epsilon^i w^{\pm,i}(r/\epsilon, y) + o(\epsilon^2), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{in } \ \Omega_\pm(\Gamma^*), \\ z^\epsilon(x) = \displaystyle\sum_{i=0}^{2} \epsilon^i Z^{\pm,i}(x) + \omega(r) \cdot \sum_{j=3}^{5} \epsilon^j z^{\pm,j}(r/\epsilon, y) + o(\epsilon^2), \end{cases}$$

where $W^{\pm,i}(x) \in C^{2,\alpha}(\overline{\Omega}_\pm(\Gamma^*))$ and $Z^{\pm,i}(x) \in C^{2,\alpha}(\overline{\Omega}_\pm(\Gamma^*))$ are outer expansions, $w^{\pm,i}(\xi, y)$ and $z^{\pm,j}(\xi, y)$ are inner expansions bounded for $\pm\xi \in (0, \infty)$, and $y \in \Gamma^*$, $x = (r, y)$ is a local coordinate system in the neighborhood of Γ^*, and $\omega(r)$ is a smooth cutoff function such that

$$\omega(r) = 1, \ \ |r| \le \frac{d_0}{2} \ \ \omega(r) = 0, \ \ |r| \ge d_0.$$

For more details, see section 3.

Here we are interested in only the critical case, i.e., the expansion starts from $\epsilon\lambda^*$, which is justified by Lemma 3.1.

Concerning (1.12), we have the following theorem.

THEOREM 1.3 (SLEP equation). *Suppose the conditions* (A1)–(A6) *are valid. Then the principal part of the critical eigenvalues is given by* $\epsilon\lambda^*$, *where* λ^* *is the eigenvalue of the following problem:*

$$L\theta^* + \frac{1}{m^2}\lambda^*[h]J'(v^*)\mathcal{T}(\theta^*) = 0 \tag{1.13}$$

for $\theta^* \in (I - P)C^{2,\alpha}(\Gamma^*)$. $[h]$ *denotes the aforementioned jump at* $v = v^*$.

Theorem 1.3 gives us a general form which characterizes the asymptotic form of critical eigenvalues. When Γ^* has a special geometry such as spherical or planar shape with $f(u) = u - u^3$, then we can rigorously prove that the critical eigenvalues λ^ϵ given by the form $\lambda^\epsilon = \epsilon\lambda^* + o(1)$, and determine the stability properties. More precisely we have the following results (see section 4 for details).

THEOREM 1.4 (stability of radially symmetric patterns). *Let* Ω *be a ball of radius* R. *Then the following hold:*

(i) *For any fixed* $R \in (0, \infty)$, *there exists* $\overline{u}_0 = \overline{u}_0(R) \in (-1, 1)$ *such that the radially symmetric solution is stable for* $\overline{u} \in (\overline{u}_0, 1)$.

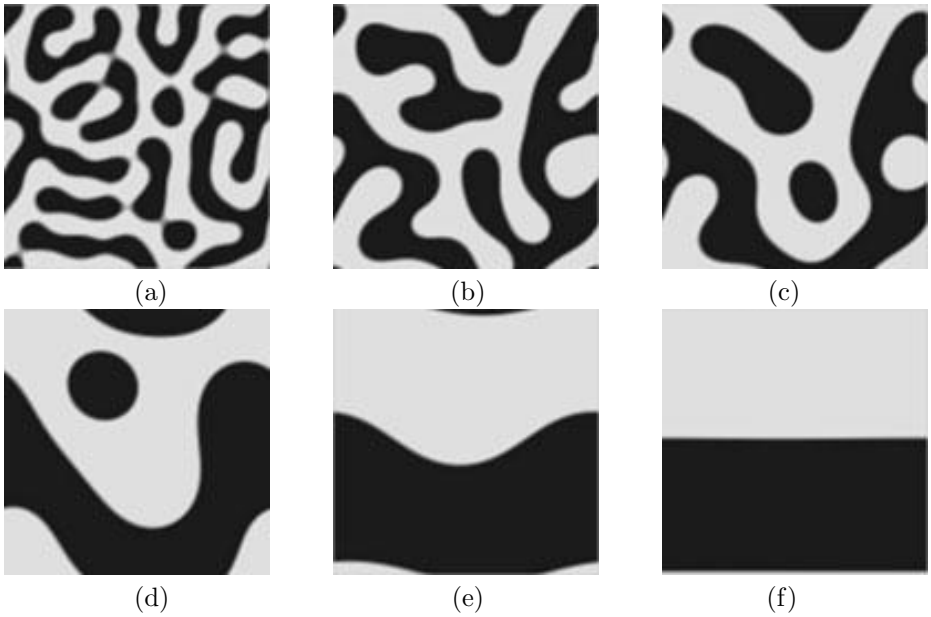

FIG. 1. *Evolution of the interfaces slightly perturbed from the homogeneous state* $u = \overline{u}$*, where* $X = Y = 2.0$*,* $\epsilon = 0.03, \overline{u} = 7.10992 \times 10^{-5}$*. Since* $\overline{u}$ *is very close to 0, lamellar shape is preferable:* (a) $\tilde{t} = 0.05$*,* (b) $\tilde{t} = 0.15$*,* (c) $\tilde{t} = 0.3$*,* (d) $\tilde{t} = 1.5$*,* (e) $\tilde{t} = 6.0$*,* (f) $\tilde{t} = 15.0$*.*

(ii) *For any fixed* $\overline{u} \in (-1, 1)$*, there exists* $R_0 = R_0(\overline{u}) > 0$ *such that the radially symmetric solution is unstable for* $R > R_0$*.*

Note that here the location of interface is given by $\Gamma^* = \{x \in \mathbf{R}^N \mid |x| = r_0\}$, where $r_0 = r_0(\overline{u})$ is a monotone decreasing function of $\overline{u}$, and $\Omega_-(\Gamma^*) = \{x \in \mathbf{R}^N \mid |x| < r_0\}$. See section 4 for details.

THEOREM 1.5 (stability of planar patterns). *Let* Ω *be a rectangle* $(0, X) \times (0, Y)$ *and* $\kappa := X/Y$ *be the aspect ratio defined before. Then the following hold:*

(i) *For any* $\overline{u} \in (-1, 1)$*, there exists* $\underline{X} = \underline{X}(\overline{u}) > 0$ *such that the planar solution is stable for any* $X < \underline{X}$ *and* $\kappa > 0$*.*

(ii) *For any fixed* $\kappa > 0$ *and* $\overline{u} \in (-1, 1)$*, there exists* $\overline{X} = \overline{X}(\kappa, \overline{u}) > 0$ *such that the planar solution is unstable for* $X > \overline{X}$*.*

The results of Theorems 1.4 and 1.5 are useful to understand how the morphology depends on the ratio $\overline{u}$. For instance, Theorem 1.4 (i) implies that the system prefers spherical patterns when either one of the homopolymers dominates the system, which is consistent with the numerics as in Figures 1 and 2. The result of Theorem 1.5 (i) seems against our intuition at first sight; in fact, recalling that the interface is parallel to y-axis, it indicates that very long interface in a slender domain with any small aspect ratio can be stabilized, which makes a sharp contrast to the lamellar patterns arising in activator-inhibitor systems (see [22]). This is, however, consistent with the experimental results as well as numerics, because very fine lamellar structure in the original scaling becomes a long strip after rescaling.

The article is organized as follows. The construction of stationary solution of (1.5) and (1.6) is discussed in section 2 by using the matched asymptotic expansion method. In section 3, the linearized eigenvalue problem (1.12) is reduced to an eigenvalue problem on the interface Γ^*. By using these results, we study the stability properties

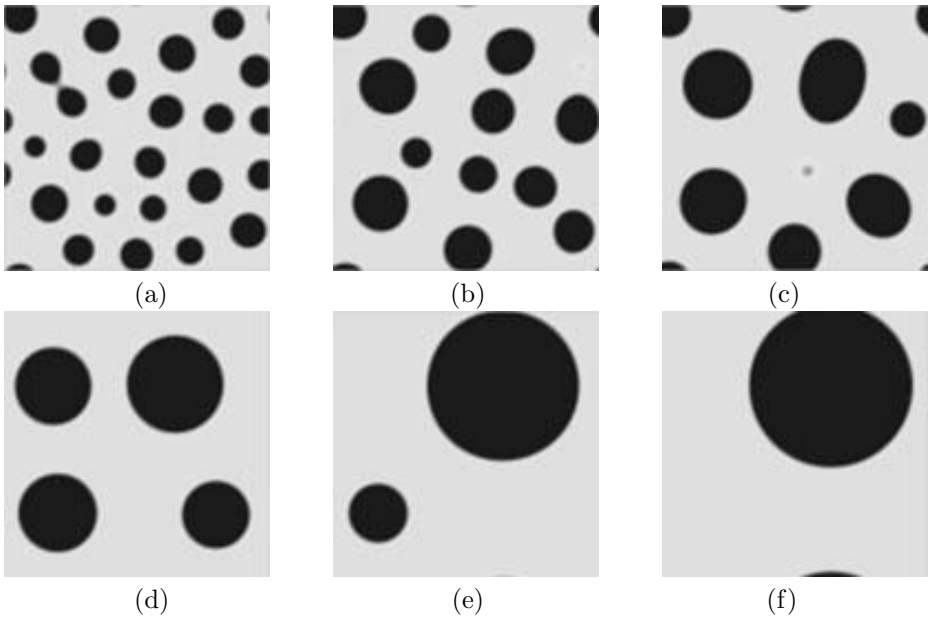

FIG. 2. *Evolution of the interfaces slightly perturbed from the homogeneous state* $u = \overline{u}$*, where* $X = Y = 2.0$*,* $\epsilon = 0.03, \overline{u} = -0.399929$*. Since* $\overline{u}$ *is away from 0, spherical shape is preferable:* (a) $\tilde{t} = 0.15$*,* (b) $\tilde{t} = 0.5$*,* (c) $\tilde{t} = 1.0$*,* (d) $\tilde{t} = 4.0$*,* (e) $\tilde{t} = 12.0$*,* (f) $\tilde{t} = 15.0$*.*

of the lamellar and radially symmetric patterns in section 4. Finally, in section 5, we derive similar results for the activator-inhibitor system (1.4).

2. Construction of stationary solutions by matched asymptotic expansion. In this section, we prove Theorem 1.2. The strategy is as follows. First, we divide the stationary problem for (1.5) and (1.6) into two problems. That is, for $a_0(y), b_0(y), b_1(y), b_2(y) \in C^{2,\alpha}(\Gamma^*)$, consider the following two problems:

$$(2.1)_- \qquad \begin{cases} \epsilon^2 \Delta u^- + f(u^-) - v^- = 0, & \\ \Delta v^- + \epsilon(u^- - \overline{u}) = 0, & \text{in } \Omega_-(\Gamma^*), \\ u^- = a_0(y), \quad v^- = b_0(y) + \epsilon b_1(y) + \epsilon^2 b_2(y) & \text{on } \Gamma^*, \end{cases}$$

$$(2.1)_+ \qquad \begin{cases} \epsilon^2 \Delta u^+ + f(u^+) - v^+ = 0, & \\ \Delta v^+ + \epsilon(u^+ - \overline{u}) = 0, & \text{in } \Omega_+(\Gamma^*), \\ u^+ = a_0(y), \quad v^+ = b_0(y) + \epsilon b_1(y) + \epsilon^2 b_2(y) & \text{on } \Gamma^*, \\ \dfrac{\partial u^+}{\partial n} = 0 = \dfrac{\partial v^+}{\partial n} & \text{on } \partial\Omega. \end{cases}$$

Here, the interface is regarded as the boundary layer at Γ^*.

Let $(u^{\pm,\epsilon}(x), v^{\pm,\epsilon}(x))$ be the solution of $(2.1)_\pm$ and define $(u^\epsilon(x), v^\epsilon(x))$ as

$$(u^\epsilon(x), v^\epsilon(x)) = \begin{cases} (u^{-,\epsilon}(x), v^{-,\epsilon}(x)), & x \in \overline{\Omega}_-(\Gamma^*), \\ (u^{+,\epsilon}(x), v^{+,\epsilon}(x)), & x \in \overline{\Omega}_+(\Gamma^*). \end{cases}$$

Since they are continuous on Γ^* and satisfy $(2.1)_\pm$ in each domain $\overline{\Omega}_\pm(\Gamma^*)$, they become a stationary solution of (1.5) and (1.6) if and only if their normal derivatives are continuous on Γ^*. So we determine $a_0(s)$, $b_0(s)$, $b_1(s)$, and $b_2(s)$ in order that the C^1-matching condition is satisfied in subsection 3.3. In fact, by taking account of the C^1-matching condition of order $O(1)$ and $O(\epsilon)$, we can see that $b_0(y) = v^*$ and $b_1(y) = b_1^*(y)$.

In Theorem 2.1 below, the symbol $C^{2,\alpha}_\epsilon(\overline{\Omega})$ stands for the function space $C^{2,\alpha}(\overline{\Omega})$ endowed with the weighted norm defined by

$$\|u\|_{C^{2,\alpha}_\epsilon(\overline{\Omega})} := \sum_{j=0}^{2} \epsilon^j |u|_{j,\overline{\Omega}} + \epsilon^{2+\alpha} |u|_{2+\alpha,\overline{\Omega}},$$

where

$$|u|_{j,\overline{\Omega}} = \max_{|\sigma|=j} \sup_{x\in\overline{\Omega}} |\partial^\sigma u(x)|, \quad |u|_{k+\alpha,\overline{\Omega}} = \max_{|\sigma|=k} \sup_{x,y\in\overline{\Omega}} \frac{|\partial^\sigma u(x) - \partial^\sigma u(y)|}{|x-y|^\alpha}$$

and σ denote the usual multi-indices.

THEOREM 2.1 (Ikeda [9]). *Suppose the conditions* (A1)–(A4) *are valid and set* $b_0(y) = v^*$ *and* $b_1(y) = b_1^*(y)$. *Then for* $\epsilon \in (0, \epsilon_0]$ *and* $a_0, b_2 \in C^{2,\alpha}(\Gamma^*)$ *satisfying* $a_0(y) \in (h^-(v^*), h^+(v^*))$, *there exist two families of solutions,*

$$(u^{-,\epsilon}, v^{-,\epsilon}) \in C^{2,\alpha}_\epsilon(\overline{\Omega}_-(\Gamma^*)) \times C^{2,\alpha}(\overline{\Omega}_-(\Gamma^*))$$

and

$$(u^{+,\epsilon}, v^{+,\epsilon}) \in C^{2,\alpha}_\epsilon(\overline{\Omega}_+(\Gamma^*)) \times C^{2,\alpha}(\overline{\Omega}_+(\Gamma^*))$$

of $(2.1)_-$ *and* $(2.1)_+$, *respectively, satisfying the following properties:*

(1) $\lim_{\epsilon\to 0} v^{\pm,\epsilon}(x) = v^*$ *uniformly on* $\overline{\Omega}_\pm(\Gamma^*)$.

(2) *For an arbitrary* $\delta > 0$, $\lim_{\epsilon\to 0} u^{\pm,\epsilon}(x) = h^\pm(v^*)$ *uniformly on* $\overline{\Omega}_\pm(\Gamma^*)\backslash\Gamma^*_\delta$.

(3) *The solutions* $(u^{\pm,\epsilon}, v^{\pm,\epsilon})$ *have an asymptotic characterization as follows: There exists a constant such that the estimates below are valid uniformly in* $\epsilon \in (0, \epsilon_0]$:

$$\|u^{\pm,\epsilon} - \mathcal{U}^{\pm,\epsilon}\|_{C^{2,\alpha}_\epsilon(\overline{\Omega}_\pm(\Gamma^*))} \le C\epsilon^{3-\alpha},$$

$$\|v^{\pm,\epsilon} - \mathcal{V}^{\pm,\epsilon}\|_{C^{2,\alpha}(\overline{\Omega}_\pm(\Gamma^*))} \le C\epsilon^{3-\alpha},$$

where $(\mathcal{U}^{\pm,\epsilon}, \mathcal{V}^{\pm,\epsilon})$ *are approximate solutions* (*for more precise information, see* (2.7) *and the subsequent subsections*).

In the next two subsections, we show the procedures of the asymptotic expansion.

2.1. Outer expansion. Substituting the expansion

$$U^{\pm,\epsilon}(x) = U^{\pm,0}(x) + \epsilon U^{\pm,1}(x) + \epsilon^2 U^{\pm,2}(x),$$

$$V^{\pm,\epsilon}(x) = V^{\pm,0}(x) + \epsilon V^{\pm,1}(x) + \epsilon^2 V^{\pm,2}(x)$$

into $(2.1)_\pm$ and equating like powers of ϵ^k, we have the following problem for $(U^{\pm,k}(x), V^{\pm,k}(x))$ $(k = 0, 1, 2)$:

$$\begin{cases} f(U^{\pm,0}) - V^{\pm,0} = 0, \\ \Delta V^{\pm,0} = 0, \end{cases}$$

$$\begin{cases} f_u^{\pm,0} U^{\pm,1} - V^{\pm,1} = 0, \\ \Delta V^{\pm,1} + U^{\pm,0} - \overline{u} = 0, \end{cases}$$

$$\begin{cases} f_u^{\pm,0} U^{\pm,2} - V^{\pm,2} + \dfrac{1}{2} f_{uu}^{\pm,0} (U^{\pm,1})^2 = 0, \\ \Delta V^{\pm,2} + U^{\pm,1} = 0, \end{cases}$$

where $f_u^{\pm,0} = \frac{d}{du} f(U^{\pm,0}(x))$ and others are similarly defined. $V^{-,0}$ and $V^{+,0}$ are uniquely determined under the boundary conditions

$$V^{\pm,0} = b_0 \quad \text{on } \Gamma^*, \quad \frac{\partial V^{+,0}}{\partial n} = 0 \quad \text{on } \partial\Omega,$$

respectively. They are represented as $V^{\pm,0} = \mathcal{P}^\pm b_0$. Then we define $U^{\pm,0}$ as

$$U^{\pm,0} = h^\pm(V^{\pm,0}),$$

where $h^\pm(v)$ is sub-branch of $f(u) - v = 0$ defined in section 1.

The equation for $V^{\pm,1}$ with the boundary conditions is recast as

$$(2.2)_\pm \qquad \begin{cases} \Delta V^{\pm,1} + h^\pm(V^{\pm,0}) - \overline{u} = 0 & \text{in } \Omega_\pm(\Gamma^*), \\ V^{\pm,1} = b_1(y) & \text{on } \Gamma^*, \\ \dfrac{\partial V^{+,1}}{\partial n} = 0 & \text{on } \partial\Omega, \end{cases}$$

respectively. It is not difficult to show that $(2.2)_\pm$ has a unique solution, $V^{\pm,1}$. Then, by using $V^{\pm,1}$, $U^{\pm,1}$ are determined as

$$U^{\pm,1} = h_v^\pm(V^{\pm,0}) V^{\pm,1},$$

where $h_v^\pm(v) = \frac{d}{dv} h^\pm(v)$. As for the equations for $(U^{\pm,2}, V^{\pm,2})$, we obtain the following:

$$U^{\pm,2}(x) = h_v^\pm(V^{\pm,0}) V^{\pm,2} + \frac{1}{2} h_{vv}^\pm(V^{\pm,0})(V^{\pm,1})^2$$

and

$$\begin{cases} 0 = \Delta V^{\pm,2} + h_v^\pm(V^{\pm,0}) V^{\pm,1}(x) & \text{in } \Omega_\pm(\Gamma^*), \\ \dfrac{\partial V^{+,2}}{\partial n} = 0 \quad \text{on } \partial\Omega, \quad V^{\pm,2} = b_2(y) & \text{on } \Gamma^*. \end{cases}$$

The boundary value problems for $V^{\pm,2}$ are uniquely solvable for arbitrary $b_2 \in C^{2,\alpha}(\Gamma^*)$. They are expressed as

$$V^{\pm,2} = \mathcal{P}^\pm b_2 + W^\pm,$$

where $W^{\pm} \in C^{2,\alpha}(\overline{\Omega}_{\pm}(\Gamma^*))$ are solutions of

$$\begin{cases} \Delta W^{\pm} = -h_v^{\pm}(V^{\pm,0}(x))V^{\pm,1}(x) & \text{in } \Omega_{\pm}(\Gamma^*), \\ W^{\pm} = 0 \ \text{ on } \Gamma^*, \quad \dfrac{\partial W^+}{\partial n} = 0 & \text{on } \partial\Omega, \end{cases}$$

respectively. Once $V^{\pm,2}$ is known, $U^{\pm,2}$ is uniquely determined by the expression above.

In this way, we have obtained the following outer expansion:

$$\begin{cases} V^{\pm,\epsilon}(x) = V^{\pm,0}(x) + \epsilon V^{\pm,1}(x) + \epsilon^2(\mathcal{P}^{\pm}b_2 + W^{\pm}), \\ U^{\pm,\epsilon}(x) = h^{\pm}(V^{\pm,0}(x)) + \epsilon h_v^{\pm}(V^{\pm,0}(x))V^{\pm,1}(x) + \epsilon^2 U^{\pm,2}(x), \end{cases} \qquad x \in \overline{\Omega}_{\pm}(\Gamma^*).$$

This expansion is due to the lack of the layer part, in fact, $U^{+,\epsilon}$ and $U^{-,\epsilon}$ are not continuous on Γ^*. So we need a new variable that stretches a neighborhood of the interface.

2.2. Inner expansion. Since the outer expansion $U^{+,\epsilon}$ and $U^{-,\epsilon}$ are not continuous on Γ^*, we need to introduce the stretched variable and make another expansion in the neighborhood of Γ^*.

Let us introduce a local coordinate system in the neighborhood of Γ^*. By the implicit function theorem there exists a $d_0 > 0$ such that the map $p : [-d_0, d_0] \times \Gamma^* \to \Gamma^*_{d_0}$ defined by

$$p(r, y) = y + r\nu(y)$$

is a diffeomorphism, where $\Gamma^*_{d_0} = \{x \in \Omega \mid \operatorname{dist}(x, \Gamma^*) < d_0\}$. Using this diffeomorphism, we identify $x \in \Gamma^*_{d_0}$ with $(r, y) \in [-d_0, d_0] \times \Gamma^*_{d_0}$ and write $u(r, y)$ for $u(x)$ $(x \in \Gamma^*_{d_0})$. With this representation, the suitable magnification of $\Gamma^*_{d_0}$ corresponds to the scaling: $r = \epsilon\xi$. In terms of the variables ξ and $y \in \Gamma^*$, the equations in (1.5) are recast as

$$\begin{cases} 0 = u_{\xi\xi} + \epsilon(N-1)H(\epsilon\xi, y)u_{\xi} + \epsilon^2\Delta(\epsilon\xi)u + f(u) - v, \\ 0 = v_{\xi\xi} + \epsilon(N-1)H(\epsilon\xi, y)v_{\xi} + \epsilon^2\Delta(\epsilon\xi)v + \epsilon^3(u - \overline{u}). \end{cases} \tag{2.3}$$

Here, $H(r, y)$ stands for the mean curvature of the manifold

$$\Gamma^*(r) := \{x \in \Omega | x = y + r\nu(y) \ y \in \Gamma^*\},$$

and $\Delta(r)$ for the Laplace–Beltrami operator on $\Gamma^*(r)$ (for more details, see Sakamoto [17]).

We now determine the functions $u^{\pm,i}$ $(i = 0, 1, 2)$, $v^{\pm,j}$ $(j = 3, 4)$ in the following expressions:

$$\begin{cases} u = U^{\pm,\epsilon}(\epsilon\xi, y) + u^{\pm,0}(\xi, y) + \epsilon u^{\pm,1}(\xi, y) + \epsilon^2 u^{\pm,2}(\xi, y), \\ v = V^{\pm,\epsilon}(\epsilon\xi, y) + \epsilon^3 v^{\pm,3}(\xi, y) + \epsilon^4 v^{\pm,4}(\xi, y). \end{cases} \tag{2.4}$$

The reason why we go to third order $v^{\pm,3}(\xi, y)$ in ϵ is that the inhomogeneous term of the equation for v appears in $O(\epsilon)$-term. We expand the mean curvature $H(\epsilon\xi, y)$

and the Laplace–Beltrami operator $\Delta(\epsilon\xi)$ of manifold $\Gamma^*(\epsilon\xi)$ as

$$H(\epsilon\xi, y) = \sum_{i\geq 0} \epsilon^i \quad H^i(\xi, y), \quad H^i(\xi, y) := \frac{1}{i!}\frac{d^i}{d\epsilon^i} H(\epsilon\xi, y)\Big|_{\epsilon=0},$$

$$\Delta(\epsilon\xi) = \sum_{i\geq 0} \epsilon^i \quad \Delta^i(\xi), \qquad \Delta^i(\xi) := \frac{1}{i!}\frac{d^i}{d\epsilon^i} \Delta(\epsilon\xi)\Big|_{\epsilon=0}.$$

Here we take account of $O(1)$-term of the C^1-matching condition for v. It is given by

$$\frac{\partial V^{-,0}}{\partial \nu} = \frac{\partial V^{+,0}}{\partial \nu} \quad \text{on } \Gamma^*.$$

Combining the result of subsection 2.1, we can see that $V^{\pm,0}$ is a constant function, that is,

$$V^{-,0}(x) = V^{+,0}(x) = \overline{b}_0$$

and $b_0(y) = \overline{b}_0$ for some $\overline{b}_0 \in \mathbf{R}$. In the following, we will use this fact.

Substituting (2.4) and the expansions of $H(\epsilon\xi, y)$ and $\Delta(\epsilon\xi)$ into (2.3), equating like powers of ϵ, we have equations for $u^{\pm,i}$ $(i = 0, 1, 2)$, $v^{\pm,j}$ $(j = 3, 4)$. We present them only for $u^{-,i}$, $v^{-,j}$ below, and omit the superscript "-" of $u^{-,i}$ and $v^{-,j}$.

The equation for u^0 is

$$u^0_{\xi\xi} + f(h^-(\overline{b}_0) + u^0) - \overline{b}_0 = 0, \ \ \xi \in (-\infty, 0).$$

In view of the boundary conditions in $(2.1)_-$, we impose boundary conditions

$$u^0(0, y) + h^-(\overline{b}_0) = a_0(y), \ \ \lim_{\xi\to-\infty} u^0(\xi, y) = 0.$$

Let $u^{\pm,*}(\xi; v)$ be the unique solution of

$$u^{\pm,*}_{\xi\xi} + f(u^{\pm,*}) - v = 0, \ \ \xi \in \pm(0, \infty), \ \ u^{\pm,*}(0; v) = h^0(v), \ \ \lim_{\xi\to\pm\infty} u^{\pm,*}(\xi; v) = h^{\pm}(v).$$

By using these functions, $u^{-,0}(\xi, y)$ is given by

$$u^0(\xi, y) = u^{-,*}(s^-_0(y) + \xi; \overline{b}_0) - h^-(\overline{b}_0),$$

where $s^-_0 \in C^{2,\alpha}(\Gamma^*)$ is related to $a_0 \in C^{2,\alpha}(\Gamma^*)$ via the equation

$$a_0(y) = u^{-,*}(s^-_0(y); \overline{b}_0), \ \ y \in \Gamma^*.$$

The equation for $(u^{-,1}, v^{-,3})$ is

$$\begin{cases} 0 = u^1_{\xi\xi} + \tilde{f}_u u^1 + p_1(\xi, y), \\ 0 = v^3_{\xi\xi} + u^0(\xi, y), \end{cases} \qquad \xi \in (-\infty, 0), \tag{2.5}$$

where

$$p_1(\xi, y) = (N-1)H(0, y)u^0_\xi(\xi, y) + \tilde{f}_u U^1(0, y) - V^1(0, y),$$

$\tilde{f}_u = \frac{d}{du} f(u^{-,*}(s_0(y)+\xi;\bar{b}_0))$ and others are similarly defined. We emphasize the fact $p_1(\xi, y)$ decays exponentially to zero as $\xi \to -\infty$ uniformly in $y \in \Gamma^*$. If we impose the boundary conditions

$$u^1(0,y) = -U^{-,1}(0,y) = -h_v^-(v^*)b_1^*(y), \quad \lim_{\xi\to-\infty} u^1(\xi,y) = 0,$$

then the first equation of (2.5) has a unique solution given by

$$u^1(\xi,y) = -U^1(0,y)\frac{u_\xi^0(\xi,y)}{u_\xi^0(0,y)} - u_\xi^0(\xi,y)\int_0^\xi \frac{1}{[u_\xi^0(\tau,y)]^2}\int_{-\infty}^\tau p_1(s,y)u_\xi^0(s,y)dsd\tau.$$

The boundary conditions for v^3 are

$$v^3(0,y) = 0, \quad \lim_{\xi\to-\infty} v^3(\xi,y) = 0,$$

and v^3 is uniquely determined as

$$v^3(\xi,y) = -\int_{-\infty}^\xi \int_{-\infty}^\tau u^0(s,y)dsd\tau.$$

Finally, we treat the equation for $(u^{-,2}, v^{-,4})$

$$\begin{cases} 0 = u_{\xi\xi}^2 + \tilde{f}_u u^2 + p_2(\xi,y), \\ 0 = v_{\xi\xi}^4 + q_4(\xi,y), \end{cases} \qquad \xi \in (-\infty, 0), \tag{2.6}$$

where

$$\begin{aligned} p_2(\xi,y) &= (N-1)H(0,y)u_\xi^1 + \Delta(0)u^0 + (N-1)H_r(0,y)\xi u_\xi^0 \\ &\quad + \frac{1}{2}\tilde{f}_{uu}[U^1(0,y)+u^1]^2 + \tilde{f}_u[\xi U_r^1(0,y) + U^2(0,y) + u^2] \\ &\quad - [\xi V_r^1(0,y) + V^2(0,y)], \\ q_4(\xi,y) &= (N-1)H(0,y)v_\xi^1 + u^1, \end{aligned}$$

and $U_r^1(0,y) = \frac{\partial}{\partial r}U^1(0,y)$ and others are similarly defined. We note the fact that $p_2(\xi,y)$ and $q_4(\xi,y)$ decay exponentially to zero as $\xi \to -\infty$ uniformly in $y \in \Gamma^*$. If we impose the boundary conditions

$$\begin{aligned} &u^2(0,y) = -U^{-,2}(0,y) = -h_v^-(v^*)b_2(y) - \frac{1}{2}h_{vv}^-(v^*)(b_1(y))^2, \quad \lim_{\xi\to-\infty} u^2(\xi,y) = 0, \\ &v^4(0,y) = 0, \qquad\qquad \lim_{\xi\to-\infty} v^4(\xi,y) = 0, \end{aligned}$$

then the equations in (2.6) have unique solutions given by

$$u^2(\xi,y) = -U^2(0,y)\frac{u_\xi^0(\xi,y)}{u_\xi^0(0,y)} - u_\xi^0(\xi,y)\int_0^\xi \frac{1}{[u_\xi^0(\tau,y)]^2}\int_{-\infty}^\tau p_2(s,y)u_\xi^0(s,y)dsd\tau$$

and

$$v^4(\xi,y) = -\int_{-\infty}^\xi \int_{-\infty}^\tau q_4(s,y)dsd\tau.$$

The same type of arguments can be applied to $u^{+,i}$ $(i=0,1,2)$ and $v^{+,j}$ $(j=3,4)$, which leads to the following approximation:

$$\begin{cases} \mathcal{U}^{\pm,\epsilon}(x) = U^{\pm,\epsilon}(x) + \{u^{\pm,0}(r/\epsilon,y) + \epsilon u^{\pm,1}(r/\epsilon,y) + \epsilon^2 u^{\pm,2}(r/\epsilon,y)\}\cdot\omega(r), \\ \mathcal{V}^{\pm,\epsilon}(x) = V^{\pm,\epsilon}(x) + \{\epsilon^3 v^{\pm,3}(r/\epsilon,y) + \epsilon^4 v^{\pm,4}(r/\epsilon,y)\}\cdot\omega(r), \end{cases} \tag{2.7}$$

where $\omega(r)$ is a smooth cutoff function such that

$$\omega(r)=1, \ \ |r|\le \frac{d_0}{2} \ \ \omega(r)=0, \ \ |r|\ge d_0.$$

2.3. C^1-matching of normal derivatives on Γ^*. In this subsection, we make stationary solutions with the internal transition layer on a whole domain Ω by matching the normal derivatives of $(u^{\pm,\epsilon}(x), v^{\pm,\epsilon}(x))$ on Γ^*.

The normal derivatives $\mathcal{U}^{\pm,\epsilon}$ and $\mathcal{V}^{\pm,\epsilon}$ on Γ^* are computed as

$$\frac{\partial}{\partial\nu}\mathcal{U}^{\pm,\epsilon}\Big|_{\Gamma^*} = \epsilon U_r^{\pm,1}(0,y) + \epsilon^2 U_r^{\pm,2}(0,y) + \frac{1}{\epsilon}u_\xi^{\pm,0}(0,y) + u_\xi^{\pm,1}(0,y) + \epsilon u_\xi^{\pm,2}(0,y),$$

$$\frac{\partial}{\partial\nu}\mathcal{V}^{\pm,\epsilon}\Big|_{\Gamma^*} = \epsilon V_r^{*}(0,y) + \epsilon^2\frac{\partial}{\partial\nu}(\mathcal{P}^{\pm}b_2) + \epsilon^2 W_r^{\pm}(0,y) + \epsilon^2 v_\xi^{\pm,3}(0,y) + \epsilon^3 v_\xi^{\pm,4}(0,y).$$

Then $O(\frac{1}{\epsilon})$ and $O(1)$-term of $u_r^{-,\epsilon}(0,y) - u_r^{+,\epsilon}(0,y)$ are given by

$$u_\xi^{-,0}(0,y) - u_\xi^{+,0}(0,y) \quad \text{and} \quad u_\xi^{-,1}(0,y) - u_\xi^{+,1}(0,y).$$

Multiplying $u_\xi^{\pm,0}$ by the equation of $u^{\pm,0}$ and integrating it from $\pm\infty$ to 0, we have

$$\frac{1}{2}[u_\xi^{\pm,0}(0,y)]^2 + \int_{h^{\pm}(\bar b_0)}^{a_0(y)} [f(u) - \bar b_0]du = 0.$$

Noting that $u^{\pm,0}(\xi,y)$ is a monotonically increasing function with respect to ξ, we can see that $u_\xi^{-,0}(0,y) - u_\xi^{+,0}(0,y) = 0$ is equivalent to

$$J(\bar b_0) = \int_{h^-(\bar b_0)}^{h^+(\bar b_0)} [f(u) - \bar b_0]du = 0.$$

We can conclude from (A3) that

$$b_0(y) \equiv \bar b_0 = v^* = 0.$$

In the following, we set $\bar b_0 = v^*$. Then we can see that $u^{\pm,0}(\xi,y)$ is represented as

$$u^{\pm,0}(\xi,y) = u^*(\xi + s_0(y)) - h^{\pm}(v^*),$$

where $u^*(\xi)$ is the function defined in section 1 and $s_0 \in C^{2,\alpha}(\Gamma^*)$ is a function related to $a_0 \in C^{2,\alpha}(\Gamma^*)$ via the equation

$$a_0(y) = u^*(s_0(y)), \ \ y\in\Gamma^*.$$

Also we obtain $s_0(y) = s_0^-(y) = s_0^+(y)$.

Since $u_\xi^{\pm,1}(0,y)$ is computed as

$$u_\xi^{\pm,1}(0,y) = -\frac{1}{u_\xi^*(s_0)} \int_{\pm\infty}^0 \left[(N-1)H(0,y)u_\xi^*(\xi+s_0(y)) - b_1(y)\right] u_\xi^*(s+s_0)ds,$$

we can see that $u_\xi^{-,1}(0,y) - u_\xi^{+,1}(0,y) = 0$ is equivalent to

$$0 = \frac{1}{u_\xi^*(s_0)} \left[(N-1)H(0,y) \int_{-\infty}^{\infty} [u_\xi^*(\xi)]^2 d\xi - b_1(y)[h]\right].$$

Thus we obtain

$$b_1(y) = b_1^*(y) = \frac{m^2}{[h]}(N-1)H(0,y).$$

Concerning the normal derivatives of $u^{-,\epsilon}(x)$ and $v^{-,\epsilon}(x)$, we have the following proposition.

PROPOSITION 2.2. *Set $b_0(y) = v^*$ and $b_1(y) = b_1^*(y)$. For the derivatives of $(u^{\pm,\epsilon}(x), v^{\pm,\epsilon}(x))$ on Γ^*, we have the following relations:*

$$v_r^{-,\epsilon}(0,y) - v_r^{+,\epsilon}(0,y) = \epsilon^2 \left[\Pi_- b_2 + \Pi_+ b_2 - s_0[h] - \Psi_0^*\right] + \epsilon^{3-\alpha} R_1^\epsilon(s_0,b_2), \tag{2.8}$$

$$\begin{aligned} &u_\xi^*(s_0)[u_r^{-,\epsilon}(0,y) - u_r^{+,\epsilon}(0,y)] \\ &\quad = \epsilon \left[(-m^2\Delta(0) - m^2 H^*(y) + J'(v^*)V_r(0,y))s_0 - J'(v^*)b_2 - \Phi_0^*\right] \\ &\qquad + \epsilon^{2-\alpha} R_2^\epsilon(s_0,b_2), \end{aligned} \tag{2.9}$$

where

$$m^2 = \int_{-\infty}^{\infty} [u_\xi^*(\xi)]^2 d\xi, \quad H^*(y) := -(N-1)\frac{\partial}{\partial r} H(0,y) = \sum_{j=1}^{N-1} \kappa_j^2,$$

$$\Psi_0^* = W_r^+(0,y) - W_r^-(0,y) + \int_{-\infty}^0 [u^*(\xi) - h^-(v^*)]d\xi - \int_{\infty}^0 [u^*(\xi) - h^+(v^*)]d\xi,$$

$$\begin{aligned} \Phi_0^* = &-H^*(y) \int_{-\infty}^{\infty} \xi [u_\xi^*(\xi)]^2 d\xi - V_r^1(0,y) \int_{-\infty}^{\infty} \xi u_\xi^*(\xi) d\xi \\ &+ \int_{-\infty}^{\infty} \xi [(N-1)H(0,y)u_\xi^*(\xi) - b_1^*(y)]u_\xi^*(\xi)d\xi - \frac{1}{2}[b_1^*(y)]^2[h^+(v^*) - h^-(v^*)], \end{aligned}$$

κ_j $(j = 1,2,\ldots,N-1)$ are the principal curvatures of Γ^*. Moreover, $R_1^\epsilon(s_0,b_2)$ and $R_2^\epsilon(s_0,b_2)$ satisfy

$$\|R_1^\epsilon(s_0,b_2)\|_{C^{1,\alpha}(\Gamma^*)} = O(1), \quad \|R_2^\epsilon(s_0,b_2)\|_{C^\alpha(\Gamma^*)} = O(1) \quad as \ \epsilon \to 0.$$

Proof. See Appendix A. □

LEMMA 2.3. *The operators $R_1^\epsilon(s_0,b_2)$ and $R_2^\epsilon(s_0,b_2)$ are Lipschitz continuous in (s_0,b_2). More precisely, there exists a $C > 0$, independent of $\epsilon \in (0,\epsilon_0]$, such that*

$$\|R_1^\epsilon(s^1,b^1) - R_1^\epsilon(s^2,b^2)\|_{C^{1,\alpha}(\Gamma^*)} \le C[\|s^1 - s^2\|_{C^{2,\alpha}(\Gamma^*)} + \|b^1 - b^2\|_{C^{2,\alpha}(\Gamma^*)}],$$

$$\|R_2^\epsilon(s^1,b^1) - R_2^\epsilon(s^2,b^2)\|_{C^{\alpha}(\Gamma^*)} \le C[\|s^1 - s^2\|_{C^{2,\alpha}(\Gamma^*)} + \|b^1 - b^2\|_{C^{2,\alpha}(\Gamma^*)}].$$

Proof of Lemma 2.3. We can prove this lemma in the same manner as the proof of Theorem 3.1 in [17].

Let us solve the following equations for s_0 and b_2 under the conditions $b_0(y) = v^*$ and $b_1(y) = b_1^*(y)$:

$$\begin{cases} \Phi(s_0, b_2, \epsilon) := -\dfrac{1}{m^2\epsilon} u_\xi^*(s_0)[u_r^{-,\epsilon}(0,y) - u_r^{+,\epsilon}(0,y)] = 0, \\ \Psi(s_0, b_2, \epsilon) := \dfrac{1}{\epsilon^2}[v_r^{-,\epsilon}(0,y) - v_r^{+,\epsilon}(0,y)] = 0. \end{cases} \tag{2.10}$$

The next proposition guarantees that the normal derivatives of $u^{\pm,\epsilon}$ and $v^{\pm,\epsilon}$ on Γ^* are matched continuously for an appropriate pair $(s_0^\epsilon, b_2^\epsilon)$.

PROPOSITION 2.4. *Suppose the conditions* (A1)–(A5) *are valid. There exists a pair* $(s_0^\epsilon, b_2^\epsilon) \in C^{2,\alpha}(\Gamma^*) \times C^{2,\alpha}(\Gamma^*)$ *satisfying* (2.10) *for small* $\epsilon > 0$.

Proof. When $\epsilon = 0$, (2.10) can be rewritten as

$$\begin{cases} Ms + \dfrac{1}{m^2} J'(v^*)b + \dfrac{1}{m^2}\Phi_0^* = 0, \\ \Pi b - [h]s - \Psi_0^* = 0, \end{cases} \tag{2.11}$$

where

$$M := \Delta^{\Gamma^*} + H^*(y) - \frac{1}{m^2} J'(v^*) V_r(0,y).$$

Let P be a projection onto the null space $\mathcal{N}(\Pi)$, i.e.,

$$Pa := \frac{1}{|\Gamma^*|} \int_{\Gamma^*} a \, dS \quad \text{for } a \in C^{2,\alpha}(\Gamma^*).$$

Then (2.11) is equivalent to the following system:

$$\begin{cases} PM(s_{\mathcal{N}} + s_\dagger) + \dfrac{1}{m^2} J'(v^*) b_{\mathcal{N}} + \dfrac{1}{m^2} P\Phi_0^* = 0, \\ (I-P)M(s_{\mathcal{N}} + s_\dagger) + \dfrac{1}{m^2} J'(v^*) b_\dagger + \dfrac{1}{m^2}(I-P)\Phi_0^* = 0, \\ -[h]s_{\mathcal{N}} - P\Psi_0^* = 0, \\ \Pi b_\dagger - [h]s_\dagger - (I-P)\Psi_0^* = 0, \end{cases} \tag{2.12}$$

where $s = s_{\mathcal{N}} + s_\dagger$, $b = b_{\mathcal{N}} + b_\dagger$, $s_{\mathcal{N}}, b_{\mathcal{N}} \in PC^{2,\alpha}(\Gamma^*)$, and $s_\dagger, b_\dagger \in (I-P)C^{2,\alpha}(\Gamma^*)$. Solving the third and fourth equations in (2.12) with respect to $s_{\mathcal{N}}$ and $b_\dagger$, respectively, we have

$$\begin{cases} s_{\mathcal{N}}^* = -\dfrac{1}{[h]} P\Psi_0^*, \\ b_\dagger = [h]\mathcal{T} s_\dagger + \mathcal{T}(I-P)\Psi_0^*. \end{cases} \tag{2.13}$$

Substituting (2.13) into the first and second equations in (2.12), we obtain

$$\begin{cases} PM(s_{\mathcal{N}}^* + s_\dagger) + \dfrac{1}{m^2} J'(v^*) b_{\mathcal{N}} + \dfrac{1}{m^2} P\Phi_0^* = 0, \\ (I-P)Ms_\dagger + \dfrac{1}{m^2}[h]J'(v^*)\mathcal{T}s_\dagger + (I-P)Ms_{\mathcal{N}}^* \\ \quad + \dfrac{1}{m^2} J'(v^*)\mathcal{T}(I-P)\Psi_0^* + \dfrac{1}{m^2}(I-P)\Phi_0^* = 0. \end{cases} \tag{2.14}$$

By using the assumption (A5), we can solve the second equation in (2.14) with respect to $s_\dagger$.

$$s_\dagger^* = -L^\dagger \left[(I-P)Ms_{\mathcal{N}}^* + \frac{1}{m^2} J'(v^*)\mathcal{T}(I-P)\Psi_0^* + \frac{1}{m^2}(I-P)\Phi_0^* \right].$$

Then $b_{\mathcal{N}}$ and $b_\dagger$ are determined as

$$b_{\mathcal{N}}^* = -\frac{1}{J'(v^*)}[m^2 PM(s_{\mathcal{N}}^* + s_\dagger^*) - P\Phi_0^*], \;\; b_\dagger^* = [h]\mathcal{T}s_\dagger^* + \mathcal{T}(I-P)\Psi_0^*.$$

Now we find solutions of (2.10). Substituting

$$s_0 = s_{\mathcal{N}}^* + s_\dagger^* + \sigma_{\mathcal{N}} + \sigma_\dagger \quad \text{and} \quad b_2 = b_{\mathcal{N}}^* + b_\dagger^* + \beta_{\mathcal{N}} + \beta_\dagger$$

into (2.10) and operating P and $I-P$ to the equations, we have

$$PM(\sigma_{\mathcal{N}} + \sigma_\dagger) + \frac{1}{m^2} J'(v^*)\beta_{\mathcal{N}} - \epsilon^{1-\alpha} P\hat{R}_2^\epsilon(\sigma_{\mathcal{N}}, \sigma_\dagger, \beta_{\mathcal{N}}, \beta_\dagger) = 0, \tag{2.15}$$

$$(I-P)M(\sigma_{\mathcal{N}} + \sigma_\dagger) + \frac{1}{m^2} J'(v^*)\beta_\dagger - \epsilon^{1-\alpha}(I-P)\hat{R}_2^\epsilon(\sigma_{\mathcal{N}}, \sigma_\dagger, \beta_{\mathcal{N}}, \beta_\dagger) = 0, \tag{2.16}$$

$$-[h]\sigma_{\mathcal{N}} + \epsilon^{1-\alpha} P\hat{R}_1^\epsilon(\sigma_{\mathcal{N}}, \sigma_\dagger, \beta_{\mathcal{N}}, \beta_\dagger) = 0, \tag{2.17}$$

$$\Pi\beta_\dagger - [h]\sigma_\dagger + \epsilon^{1-\alpha}(I-P)\hat{R}_1^\epsilon(\sigma_{\mathcal{N}}, \sigma_\dagger, \beta_{\mathcal{N}}, \beta_\dagger) = 0, \tag{2.18}$$

where

$$\hat{R}_1^\epsilon(\sigma_{\mathcal{N}}, \sigma_\dagger, \beta_{\mathcal{N}}, \beta_\dagger) = R_1^\epsilon(s_{\mathcal{N}}^* + s_\dagger^* + \sigma_{\mathcal{N}} + \sigma_\dagger, b_{\mathcal{N}}^* + b_\dagger^* + \beta_{\mathcal{N}} + \beta_\dagger),$$

$$\hat{R}_2^\epsilon(\sigma_{\mathcal{N}}, \sigma_\dagger, \beta_{\mathcal{N}}, \beta_\dagger) = \frac{1}{m^2} R_2^\epsilon(s_{\mathcal{N}}^* + s_\dagger^* + \sigma_{\mathcal{N}} + \sigma_\dagger, b_{\mathcal{N}}^* + b_\dagger^* + \beta_{\mathcal{N}} + \beta_\dagger).$$

We can solve (2.17) with respect to $\sigma_{\mathcal{N}}$ when $\epsilon > 0$ is small.

$$\sigma_{\mathcal{N}} = \tilde{\sigma}_{\mathcal{N}}^\epsilon(\sigma_\dagger, \beta_{\mathcal{N}}, \beta_\dagger) = \frac{\epsilon^{1-\alpha}}{[h]} P\hat{R}_1^\epsilon(\tilde{\sigma}_{\mathcal{N}}^\epsilon(\sigma_\dagger, \beta_{\mathcal{N}}, \beta_\dagger), \sigma_\dagger, \beta_{\mathcal{N}}, \beta_\dagger). \tag{2.19}$$

Substituting (2.19) into (2.18), we have

$$\Pi\beta_\dagger - [h]\sigma_\dagger + \epsilon^{1-\alpha}(I-P)\hat{R}_1^\epsilon(\tilde{\sigma}_{\mathcal{N}}^\epsilon(\sigma_\dagger, \beta_{\mathcal{N}}, \beta_\dagger), \sigma_\dagger, \beta_{\mathcal{N}}, \beta_\dagger) = 0. \tag{2.20}$$

When $\epsilon > 0$ is small, (2.20) is solvable in $\beta_\dagger$ as

$$\begin{aligned} \beta_\dagger &= \beta_\dagger^\epsilon(\sigma_\dagger, \beta_{\mathcal{N}}) \\ &= [h]\mathcal{T}\sigma_\dagger - \epsilon^{1-\alpha}\mathcal{T}(I-P)\hat{R}_1^\epsilon(\tilde{\sigma}_{\mathcal{N}}(\sigma_\dagger, \beta_{\mathcal{N}}, \beta_\dagger^\epsilon(\sigma_\dagger, \beta_{\mathcal{N}})), \sigma_\dagger, \beta_{\mathcal{N}}, \beta_\dagger^\epsilon(\sigma_\dagger, \beta_{\mathcal{N}})). \end{aligned} \tag{2.21}$$

By using (2.21), $\sigma_{\mathcal{N}}$ is represented as

$$\sigma_{\mathcal{N}} = \sigma_{\mathcal{N}}^\epsilon(\sigma_\dagger, \beta_{\mathcal{N}}) := \tilde{\sigma}_{\mathcal{N}}^\epsilon(\sigma_\dagger, \beta_{\mathcal{N}}, \beta_\dagger^\epsilon(\sigma_\dagger, \beta_{\mathcal{N}})). \tag{2.22}$$

Here note that $\sigma^\epsilon_{\mathcal{N}}(\sigma_\dagger, \beta_{\mathcal{N}})$ and $\beta^\epsilon_\dagger(\sigma_\dagger, \beta_{\mathcal{N}})$ are Lipschitz continuous in $(\sigma_\dagger, \beta_{\mathcal{N}})$ since $R^\epsilon_j(s_0, b_2)$ are Lipschitz continuous in (s_0, b_2). Moreover, they have the following properties:

(2.23)
$$\begin{cases} |\sigma^\epsilon_{\mathcal{N}}(\sigma^1_\dagger, \beta^1_{\mathcal{N}}) - \sigma^\epsilon_{\mathcal{N}}(\sigma^2_\dagger, \beta^2_{\mathcal{N}})| \le \epsilon^{1-\alpha} C[\|\sigma^1_\dagger - \sigma^2_\dagger\|_{C^{2,\alpha}(\Gamma^*)} + \epsilon^{1-\alpha}|\beta^1_{\mathcal{N}} - \beta^2_{\mathcal{N}}|], \\ \|\beta^\epsilon_\dagger(\sigma^1_\dagger, \beta^1_{\mathcal{N}}) - \beta^\epsilon_\dagger(\sigma^2_\dagger, \beta^2_{\mathcal{N}})\|_{C^{2,\alpha}(\Gamma^*)} \le C[\|\sigma^1_\dagger - \sigma^2_\dagger\|_{C^{2,\alpha}(\Gamma^*)} + \epsilon^{1-\alpha}|\beta^1_{\mathcal{N}} - \beta^2_{\mathcal{N}}|], \\ \|B^\epsilon_\dagger(\sigma^1_\dagger, \beta^1_{\mathcal{N}}) - B^\epsilon_\dagger(\sigma^2_\dagger, \beta^2_{\mathcal{N}})\|_{C^{2,\alpha}(\Gamma^*)} \le \epsilon^{1-\alpha} C[\|\sigma^1_\dagger - \sigma^2_\dagger\|_{C^{2,\alpha}(\Gamma^*)} + \epsilon^{1-\alpha}|\beta^1_{\mathcal{N}} - \beta^2_{\mathcal{N}}|], \end{cases}$$

where

$$B^\epsilon_\dagger(\sigma_\dagger, \beta_{\mathcal{N}}) := \beta^\epsilon_\dagger(\sigma_\dagger, \beta_{\mathcal{N}}) - [h]\mathcal{T}\sigma_\dagger,$$
$$|\sigma^\epsilon_{\mathcal{N}}| = O(\epsilon^{1-\alpha}), \quad \|B^\epsilon_\dagger\|_{C^{2,\alpha}(\Gamma^*)} = O(\epsilon^{1-\alpha}) \ \text{ as } \ \epsilon \to 0.$$

Substituting (2.21) and (2.22) into (2.16), we obtain

$$\begin{aligned} (I-P)L\sigma_\dagger = &-(I-P)M\sigma^\epsilon_{\mathcal{N}}(\sigma_\dagger, \beta_{\mathcal{N}}) - \frac{1}{m^2}J'(v^*)B^\epsilon_\dagger(\sigma_\dagger, \beta_{\mathcal{N}}) \\ &+ \epsilon^{1-\alpha}(I-P)\hat{R}^\epsilon_2(\sigma^\epsilon_{\mathcal{N}}(\sigma_\dagger, \beta_{\mathcal{N}}), \sigma_\dagger, \beta_{\mathcal{N}}, \beta^\epsilon_\dagger(\sigma_\dagger, \beta_{\mathcal{N}})). \end{aligned} \tag{2.24}$$

From the assumption (A5), there exists a constant $C_0 > 0$ such that

$$\|L^\dagger\|_{C^\alpha(\Gamma^*) \longrightarrow C^{2,\alpha}(\Gamma^*)} \le C_0, \tag{2.25}$$

and (2.24) is recast as

$$\begin{aligned} \sigma_\dagger = &-L^\dagger(I-P)M\sigma^\epsilon_{\mathcal{N}}(\sigma_\dagger, \beta_{\mathcal{N}}) - \frac{1}{m^2}J'(v^*)L^\dagger B^\epsilon_\dagger(\sigma_\dagger, \beta_{\mathcal{N}}) \\ &+ \epsilon^{1-\alpha}L^\dagger(I-P)\hat{R}^\epsilon_2(\sigma^\epsilon_{\mathcal{N}}(\sigma_\dagger, \beta_{\mathcal{N}}), \sigma_\dagger, \beta_{\mathcal{N}}, \beta^\epsilon_\dagger(\sigma_\dagger, \beta_{\mathcal{N}})). \end{aligned} \tag{2.26}$$

It follows from (2.23) and (2.25) that the right-hand side of (2.26) is a contraction on $\mathbf{B} := \{\sigma_\dagger \in (I-P)C^{2,\alpha}(\Gamma^*) \mid \|\sigma_\dagger\|_{C^{2,\alpha}(\Gamma^*)} \le 1\}$ with Lipschitz constant $O(\epsilon^{1-\alpha})$. Therefore, (2.26) has a unique solution $\sigma_\dagger = \sigma^\epsilon_\dagger(\beta_{\mathcal{N}}) \in \mathbf{B}$ with the property,

$$\|\sigma^\epsilon_\dagger\|_{C^{2,\alpha}(\Gamma^*)} = O(\epsilon^{1-\alpha}) \ \text{ as } \ \epsilon \to 0.$$

Finally, substituting (2.21), (2.22), and $\sigma_\dagger = \sigma^\epsilon_\dagger(\beta_{\mathcal{N}})$ into (2.15), we obtain an equation of $\beta_{\mathcal{N}}$, which is solvable in $\beta_{\mathcal{N}}$. □

Proof of Theorem 1.2. By virtue of Theorem 2.1 and Proposition 2.4, we obtain the statements of Theorem 1.2.

3. Linearized eigenvalue problem. In this section, we study the linearized eigenvalue problem around a stationary solution $(u^\epsilon(x), v^\epsilon(x))$,

$$\begin{cases} 0 = \epsilon^2\Delta w + f^\epsilon_u w - z, & \\ & \text{in } \Omega, \\ 0 = \Delta z + \epsilon w + \lambda^\epsilon w, & \\ \dfrac{\partial w}{\partial n} = 0 = \dfrac{\partial z}{\partial n} & \text{on } \partial\Omega. \end{cases} \tag{3.1}$$

Our goal in this section is to prove Theorem 1.3. Roughly speaking, the SLEP equation (1.13) in Theorem 1.3 becomes a sufficient condition for C^1-matching conditions of the concerning eigenfunctions.

We first divide (3.1) into the following two parts:

$$(3.2)_- \qquad \begin{cases} \epsilon^2 \Delta w^{-,\epsilon} + f_u^\epsilon w^{-,\epsilon} - z^{-,\epsilon} = 0, & \\ & \text{in } \Omega_-(\Gamma^*), \\ \Delta z^{-,\epsilon} + \epsilon w^{-,\epsilon} + \lambda^\epsilon w^{-,\epsilon} = 0, & \\ w^{-,\epsilon}(y) = \Theta^\epsilon(y), \quad z^{-,\epsilon}(y) = q^\epsilon(y), & y \in \Gamma^*, \end{cases}$$

$$(3.2)_+ \qquad \begin{cases} \epsilon^2 \Delta w^{+,\epsilon} + f_u^\epsilon w^{+,\epsilon} - z^{+,\epsilon} = 0, & \\ & \text{in } \Omega_+(\Gamma^*), \\ \Delta z^{+,\epsilon} + \epsilon w^{+,\epsilon} + \lambda^\epsilon w^{+,\epsilon} = 0, & \\ \dfrac{\partial w^{+,\epsilon}}{\partial n} = 0 = \dfrac{\partial z^{+,\epsilon}}{\partial n} & \text{on } \partial\Omega, \\ w^{+,\epsilon}(y) = \Theta^\epsilon(y), \quad z^{+,\epsilon}(y) = q^\epsilon(y), & y \in \Gamma^*, \end{cases}$$

where

$$q^\epsilon(y) = q_0(y) + \epsilon q_1(y) + \epsilon^2 q_2(y), \quad \lambda^\epsilon = \epsilon \lambda_1.$$

$\Theta^\epsilon, q_0, q_1, q_2 \in C^{2,\alpha}(\Gamma^*)$ are unknown boundary data and $\lambda_1 \in \mathbf{C}$ are regarded as parameters. Here we choose $\epsilon\lambda_1$ as the principal term in order that we can successively expand $(3.2)_\pm$ and determine the parameters one after another. In fact, we have the following lemma:

LEMMA 3.1. *Suppose that $\lambda^\epsilon = \lambda_0 + \epsilon\lambda_1$. Then, if we can expand $(3.2)_\pm$ and determine the parameters $\Theta^\epsilon(y)$, $q^\epsilon(y)$ and λ^ϵ one after another, we have either* Re $\lambda_0 < 0$ *or* $\lambda_0 = 0$.

Proof. This result is proved by using the formal matched asymptotic expansions starting from $O(1)$. See Appendix B. ∎

Let us first construct the solutions $(w^{\pm,\epsilon}, z^{\pm,\epsilon})$ of $(3.2)_\pm$, namely, the following.

THEOREM 3.2 (see [9]). *Suppose the conditions* (A1)–(A5) *are valid. Then, for $\epsilon \in (0, \epsilon_0]$, $\lambda_1 \in \mathbf{C}$, $\Theta^\epsilon = \Theta \in C^{2,\alpha}(\Gamma^*)$, $q_0, q_1, q_2 \in C^{2,\alpha}(\Gamma^*)$, there exist two families of solutions*

$$(w^{-,\epsilon}, z^{-,\epsilon}) \in C_\epsilon^{2,\alpha}(\overline{\Omega}_-(\Gamma^*)) \times C^{2,\alpha}(\overline{\Omega}_-(\Gamma^*))$$

and

$$(w^{+,\epsilon}, z^{+,\epsilon}) \in C_\epsilon^{2,\alpha}(\overline{\Omega}_+(\Gamma^*)) \times C^{2,\alpha}(\overline{\Omega}_+(\Gamma^*))$$

of $(3.2)_-$ and $(3.2)_+$, respectively, which have the following asymptotic characterization: There exists a constant $C > 0$ such that the estimates below are valid uniformly in $\epsilon \in (0, \epsilon_0]$:

$$\left\| w^{\pm,\epsilon} - \mathcal{W}_2^{\pm,\epsilon} \right\|_{C_\epsilon^{2,\alpha}(\overline{\Omega}_\pm(\Gamma^*))} \le C\epsilon^{3-\alpha},$$

$$\left\| z^{\pm,\epsilon} - \mathcal{Z}_2^{\pm,\epsilon} \right\|_{C^{2,\alpha}(\overline{\Omega}_\pm(\Gamma^*))} \le C\epsilon^{3-\alpha},$$

where $(\mathcal{W}_2^{\pm,\epsilon}, \mathcal{Z}_2^{\pm,\epsilon})$ are approximate solutions (see (3.18) *for the details). Here $C_\epsilon^{2,\alpha}$ is the same Banach space defined in section* 2.

In the next two subsections, we only construct the approximate solutions of $(3.2)_-$ and omit the superscript (or subscript) + (or −).

3.1. Outer expansion. Let us substitute

$$w(x) = W^0(x) + \epsilon W^1(x) + \epsilon^2 W^2(x), \;\; z(x) = Z^0(x) + \epsilon Z^1(x) + \epsilon^2 Z^2(x)$$

into $(3.2)_-$ and equate like powers of ϵ. Then we have the following problem for $W^{-,i}(x)$ and $Z^{-,j}(x)$ $(i = 0, 1, 2)$:

$$\begin{cases} f_u^0 W^0 - Z^0 = 0, \\ \Delta Z^0 = 0, \end{cases}$$

$$\begin{cases} f_u^0 W^1 + f_u^1 W^0 - Z^1 = 0, \\ \Delta Z^1 + (1+\lambda_1) W^0 = 0, \end{cases} \tag{3.3}$$

$$\begin{cases} f_u^0 W^2 + f_u^1 W^j + f_u^2 W^0 + \Delta W^0 - Z^2 = 0, \\ \Delta Z^2 = -(1+\lambda_1) W^1, \end{cases} \tag{3.4}$$

where

$$f_u^i := \frac{1}{i!} \frac{d^i}{d\epsilon^i} f_u \left(\sum_{i=0}^{2} \epsilon^i U^i(x) \right) \Bigg|_{\epsilon=0}.$$

Z^0 is uniquely determined under the boundary condition

$$Z^0 = q_0 \quad \text{on } \Gamma^*.$$

That is represented as $Z^0 = \mathcal{P}^- q_0$. By using Z^0, W^0 is determined as

$$W^0 = \frac{1}{f_u^0} Z^0 = h_v^-(v^*) Z^0.$$

Here we used the fact that $f(h^-(v)) - v = 0$ and $f_u^0 = f_u(h^-(v^*)) \neq 0$ (see (A1)). Then, (3.3) can be rewritten as

$$\begin{cases} W^1 = \dfrac{1}{f_u^0}[-f_u^1 W^0 + Z^1], \\ \\ \Delta Z^1 = -(1+\lambda_1) h_v^-(v^*) Z^0, \end{cases} \quad \text{in } \Omega_-(\Gamma^*). \tag{3.5}$$

Once W^0 and Z^0 are known, the second equation of (3.5) is Poisson equation associated with Z^1. Therefore, Z^1 is uniquely determined under the boundary conditions

$$Z^1 = q_1 \quad \text{on } \Gamma^*.$$

That is represented as

$$Z^1 = \mathcal{P}^- q_1 + \hat{Z}^1,$$

where $\hat{Z}^1$ is the unique solution of

$$\begin{cases} \Delta \hat{Z}^1 = -(1+\lambda_1) h_v^-(v^*) Z^0 & \text{in } \Omega_-(\Gamma^*), \\ \hat{Z}^1 = 0 & \text{on } \Gamma^*. \end{cases}$$

Once Z^1 is known, W^1 is uniquely determined by the first equation of (3.5).

Noting that the boundary conditions for Z^2 is given by $Z^2 = q_2$ on Γ^*, we can solve Z^2 as

$$Z^2 = \mathcal{P}^- q_2 + \hat{Z}^2,$$

where $\hat{Z}^2$ is the unique solutions of

$$\begin{cases} \Delta \hat{Z}^2 = -(1+\lambda_1) \dfrac{Z^1 - f_u^1 W^0}{f_u^0} & \text{in } \Omega_-(\Gamma^*), \\ \hat{Z}^2 = 0 & \text{on } \Gamma^*. \end{cases}$$

Once Z^2 is known, W^2 is uniquely determined by the first equation of (3.4).

In this way, we have obtained the following outer expansion of order $O(\epsilon^2)$:

$$\begin{cases} Z_2^{-,\epsilon}(x) = \mathcal{P}^- q_0 + \displaystyle\sum_{i=1}^{2} \epsilon^i (\mathcal{P}^- q_i + \hat{Z}^i). \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad x \in \overline{\Omega}_-(\Gamma^*). \\ W_2^{-,\epsilon}(x) = \displaystyle\sum_{j=0}^{2} \epsilon^j W^j(x), \end{cases}$$

In the same way as above, we can obtain an outer expansion also for $(3.2)_+$. Since $W_2^{+,\epsilon}$ and $W_2^{-,\epsilon}$ are not continuous on Γ^*, we introduce a new variable $\xi = r/\epsilon$ and construct the inner part.

3.2. Inner expansion. In terms of the variables ξ and $y \in \Gamma^*$, the equations in (3.1) are recast as

$$\begin{cases} w_{\xi\xi} + \epsilon(N-1)H(\epsilon\xi, y) w_\xi + \epsilon^2 \Delta(\epsilon\xi) w + f_u^\epsilon w - z = 0, \\ z_{\xi\xi} + \epsilon(N-1)H(\epsilon\xi, y) z_\xi + \epsilon^2 \Delta(\epsilon\xi) z + \epsilon^3 w + \epsilon^2 \lambda^\epsilon w = 0. \end{cases} \tag{3.6}$$

We now determine the functions $w^{-,i}$ $(i = 0, 1, 2)$, $z^{-,j}$ $(j = 0, 1, \ldots, 5)$ in the following expressions:

$$\begin{aligned} w &= \sum_{i=0}^{2} \epsilon^i W^i(\epsilon\xi, y) + \sum_{i=0}^{2} \epsilon^i w^i(\xi, y) = \sum_{i=0}^{2} \epsilon^i \tilde{W}^i(\xi, y) + \sum_{i=0}^{2} \epsilon^i w^i(\xi, y), \\ z &= \sum_{i=0}^{2} \epsilon^i Z^i(\epsilon\xi, y) + \sum_{i=0}^{5} \epsilon^i z^i(\xi, y) = \sum_{i=0}^{2} \epsilon^i \tilde{Z}^i(\xi, y) + \sum_{i=0}^{5} \epsilon^i z^i(\xi, y), \end{aligned} \tag{3.7}$$

where

$$\tilde{W}^i := \frac{1}{i!} \frac{d^i}{d\epsilon^i} \left(\sum_{k=0}^{2} \epsilon^k W^k(\epsilon\xi, y) \right) \Bigg|_{\epsilon=0}, \quad \tilde{Z}^i := \frac{1}{i!} \frac{d^i}{d\epsilon^i} \left(\sum_{k=0}^{2} \epsilon^k Z^k(\epsilon\xi, y) \right) \Bigg|_{\epsilon=0}.$$

Also we expand the mean curvature $H(\epsilon\xi, y)$ and the Laplace–Beltrami operator $\Delta(\epsilon\xi)$ of manifold $\Gamma^*(\epsilon\xi)$ as in subsection 2.2. Substituting (3.7) into (3.6) and equating like powers of ϵ, we have equations for $w^{-,i}$ $(i = 0, 1, 2)$ and $z^{-,j}$ $(j = 0, 1, \ldots, 5)$ as follows:

$$\begin{cases} w^0_{\xi\xi} + \tilde{f}^0_u w^0 + \tilde{f}^0_u \tilde{W}^0 - (\tilde{Z}^0 + z^0) = 0, & \xi \in (-\infty, 0), \\ w^0(0, y) = \Theta(y) - W^0(0, y), & \lim_{\xi\to-\infty} w^0(\xi, y) = 0, \end{cases}$$

$$\begin{cases} w^1_{\xi\xi} + \tilde{f}^0_u w^1 + (N-1)H(0,y)w^0_\xi + \tilde{f}^0_u \tilde{W}^1 + \tilde{f}^1_u(\tilde{W}^0 + w^0) - (\tilde{Z}^1 + z^1) = 0, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \xi \in (-\infty, 0), \\ w^1(0, y) = -W^1(0, y), \quad \lim_{\xi\to-\infty} w^1(\xi, y) = 0, \end{cases}$$

(3.8)
$$\begin{cases} w^2_{\xi\xi} + \tilde{f}^0_u w^2 + (N-1)\displaystyle\sum_{i+j=1} H^i w^j_\xi + \Delta^0 w^0 + (\text{value of } \Delta W^0 \text{ on } \Gamma^*) \\ \quad + \tilde{f}^0_u \tilde{W}^2 + \displaystyle\sum_{i+j=2, i\geq 1} \tilde{f}^i_u(\tilde{W}^j + w^j) - (\tilde{Z}^2 + z^2) = 0, \quad \xi \in (-\infty, 0), \\ w^2(0, y) = -W^2(0, y), \qquad\qquad\qquad\qquad\qquad\qquad \lim_{\xi\to-\infty} w^2(\xi, y) = 0, \end{cases}$$

where

$$\tilde{f}^i_u := \frac{1}{i!}\frac{d^i}{d\epsilon^i} f_u \left(U^{\pm,\epsilon}(\epsilon\xi, y) + \sum_{j=0}^{2} \epsilon^j u^{\pm,j}(\xi, y) \right)\Bigg|_{\epsilon=0} \qquad (i = 0, 1, 2),$$

(3.9)
$$\begin{cases} z^0_{\xi\xi} = 0, & \xi \in (-\infty, 0), \\ z^0(0, y) = 0, & \lim_{\xi\to-\infty} z^0(\xi, y) = 0, \end{cases}$$

(3.10)
$$\begin{cases} z^1_{\xi\xi} + (N-1)H(0,y)z^0_\xi = 0, & \xi \in (-\infty, 0), \\ z^1(0, y) = 0, & \lim_{\xi\to-\infty} z^1(\xi, y) = 0, \end{cases}$$

(3.11)
$$\begin{cases} z^2_{\xi\xi} + (N-1)(H^0 z^1_\xi + H^1 z^0_\xi) + \Delta^0 z^0 = 0, & \xi \in (-\infty, 0), \\ z^2(0, y) = 0, & \lim_{\xi\to-\infty} z^2(\xi, y) = 0, \end{cases}$$

(3.12)
$$\begin{cases} z^n_{\xi\xi} + (N-1)\displaystyle\sum_{i+j=n-1} H^i z^j_\xi + \sum_{i+j=n-2} \Delta^i z^j + (1+\lambda_1) w^{n-3} = 0, \ \xi \in (-\infty, 0), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (n = 3, 4, 5), \\ z^n(0, y) = 0, \quad \lim_{\xi\to-\infty} z^n(\xi, y) = 0. \end{cases}$$

The equations (3.9), (3.10), and (3.11) imply that $z^j(\xi, y) \equiv 0$ $(j = 0, 1, 2)$.

The equation for w^0 is recast as

$$w^0_{\xi\xi} + \tilde{f}^0_u w^0 + P_0(\xi, y) = 0, \quad \xi \in (-\infty, 0), \tag{3.13}$$

where

$$P_0(\xi, y) = \tilde{f}^0_u W^0(0, y) - q_0(y), \qquad \tilde{f}^0_u = f_u(u^*(\xi + s_0(y))).$$

In view of the boundary conditions in $(3.2)_-$, we impose boundary conditions

$$w^0(0, y) = \Theta(y) - W^0(0, y), \quad \lim_{\xi \to -\infty} w^0(\xi, y) = 0.$$

By using the fact that $u^*_\xi(\xi + s_0(y))$ is a fundamental solution of (3.13), w^0 is uniquely determined as

$$\begin{aligned} w^0(\xi, y) = {} & [\Theta(y) - W^0(0, y)] \frac{u^*_\xi(\xi + s_0)}{u^*_\xi(s_0)} \\ & - u^*_\xi(\xi + s_0) \int_0^\xi \frac{1}{[u^*_\xi(\tau + s_0)]^2} \int_{-\infty}^\tau P_0(s, y) u^*_\xi(s + s_0) ds d\tau \end{aligned} \tag{3.14}$$

with $s_0 = s_0(y)$. Noting that w^0 decays exponentially to zero as $\xi \to -\infty$, we can solve z^3 as

$$z^3(\xi, y) = -(1 + \lambda_1) \int_{-\infty}^\xi \int_{-\infty}^\tau w^0(s, y) ds d\tau. \tag{3.15}$$

The equations for (w^1, z^4) are

$$\begin{cases} 0 = w^1_{\xi\xi} + \tilde{f}_u w^1 + P_1(\xi, y), \\ \\ 0 = z^4_{\xi\xi} + Q_1(\xi, y), \end{cases} \qquad \xi \in (-\infty, 0), \tag{3.16}$$

where

$$\begin{aligned} P_1(\xi, y) &= (N - 1) H(0, y) w^0_\xi + \tilde{f}^0_u \tilde{W}^1 + \tilde{f}^1_u (\tilde{W}^0 + w^0) - \tilde{Z}^1, \\ Q_1(\xi, y) &= (N - 1) H(0, y) z^3_\xi + (1 + \lambda_1) w^1. \end{aligned}$$

We emphasize the fact $P_1(\xi, y)$ decays exponentially to zero as $\xi \to -\infty$ uniformly in $y \in \Gamma^*$. Therefore, the first equation in (3.16) has a unique solution given by

$$\begin{aligned} w^1(\xi, y) = {} & -W^1(0, y) \frac{u^*_\xi(\xi + s_0)}{u^*_\xi(s_0)} \\ & - u^*_\xi(\xi + s_0) \int_0^\xi \frac{1}{[u^*_\xi(\tau + s_0)]^2} \int_{-\infty}^\tau P_1(s, y) u^*_\xi(s + s_0) ds d\tau \end{aligned}$$

with $s_0 = s_0(y)$. Once z^3 and w^1 are determined, z^4 is uniquely determined as

$$z^4(\xi, y) = - \int_{-\infty}^\xi \int_{-\infty}^\tau Q_1(s, y) ds d\tau.$$

Here we used the fact that $Q_1(\xi, y)$ decays exponentially to zero as $\xi \to -\infty$ uniformly in $y \in \Gamma^*$.

By using the above results, w^2 and z^5 are solved as

$$
\begin{aligned}
w^2(\xi, y) &= -W^2(0, y)\frac{u^*_\xi(\xi + s_0)}{u^*_\xi(s_0)} \\
&\quad - u^*_\xi(\xi + s_0)\int_0^\xi \frac{1}{[u^*_\xi(\tau + s_0)]^2}\int_{-\infty}^\tau P_2(s, y)u^*_\xi(s + s_0)dsd\tau, \\
z^5(\xi, y) &= -\int_{-\infty}^\xi \int_{-\infty}^\tau Q_2(s, y)dsd\tau.
\end{aligned} \tag{3.17}
$$

with $s_0 = s_0(y)$. Here we used the fact $P_2(\xi, y)$ and $Q_5(\xi, y)$ decays exponentially to zero as $\xi \to -\infty$ uniformly in $y \in \Gamma^*$.

The same type of arguments as above apply to $w^{+,i}$ $(i = 0, 1, 2)$ and $z^{+,j}$ $(j = 0, 1, \ldots, 5)$. Now we have obtained the approximation of order $O(\epsilon^2)$,

$$
\begin{cases}
\mathcal{W}_2^{\pm,\epsilon}(x) = W_2^{\pm,\epsilon}(x) + \omega(r)\cdot\displaystyle\sum_{i=0}^{2}\epsilon^i w^{\pm,i}(r/\epsilon, y), \\
\mathcal{Z}_2^{\pm,\epsilon}(x) = Z_2^{\pm,\epsilon}(x) + \omega(r)\cdot\displaystyle\sum_{j=3}^{5}\epsilon^j z^{\pm,j}(r/\epsilon, y),
\end{cases} \tag{3.18}
$$

where $\omega(r)$ is a smooth cutoff function such that

$$
\omega(r) = 1, \ \ |r| \le \frac{d_0}{2}\ \omega(r) = 0, \ \ |r| \ge d_0.
$$

3.3. Matching of normal derivatives on Γ^*. Now we are ready to make the eigenfunctions on a whole domain by matching the normal derivatives of $(w^{\pm,\epsilon}, z^{\pm,\epsilon})$. That is, $\theta(y) := \Theta(y)/u^*_\xi(s_0(y))$, $q_i(y)$ $(i = 0, 1, 2) \in C^{2,\alpha}(\Gamma^*)$, and $\lambda_1 \in \mathbf{C}$ must satisfy the following C^1-matching conditions:

$$
\Phi(\theta, q_0, q_1, q_2, \lambda_1, \epsilon) = 0, \quad \Psi(\theta, q_0, q_1, q_2, \lambda_1, \epsilon) = 0 \quad \text{on } \Gamma^*, \tag{3.19}
$$

where

$$
\begin{aligned}
\Phi(\theta, q_0, q_1, q_2, \lambda_1, \epsilon)(y) &= \epsilon u^*_\xi(s_0(y))[w_r^{-,\epsilon}(0, y) - w_r^{+,\epsilon}(0, y)] \\
&= u^*_\xi(s_0(y))[w_\xi^{-,0}(0, y) - w_\xi^{+,0}(0, y)] \\
&\quad + u^*_\xi(s_0(y))\sum_{i=0}^{1}\epsilon^{i+1}[W_r^{-,i}(0, y) + w_\xi^{-,i+1}(0, y) \\
&\quad - W_r^{+,i}(0, y) - w_\xi^{+,i+1}(0, y)] + \epsilon^{3-\alpha}R_2^\epsilon(\theta, q_0, q_1, q_2, \lambda_1),
\end{aligned}
$$

$$
\begin{aligned}
\Psi(\theta, q_0, q_1, q_2, \lambda_1, \epsilon)(y) &= z_r^{-,\epsilon}(0, y) - z_r^{+,\epsilon}(0, y) \\
&= Z_r^{-,0}(0, y) - Z_r^{+,0}(0, y) + \epsilon[Z_r^{-,1}(0, y) - Z_r^{+,1}(0, y)] \\
&\quad + \epsilon^2[Z_r^{-,2}(0, y) + z_\xi^{-,3}(0, y) - Z_r^{+,2}(0, y) - z_\xi^{+,3}(0, y)] \\
&\quad + \epsilon^{3-\alpha}R_1^\epsilon(\theta, q_0, q_1, q_2, \lambda_1).
\end{aligned}
$$

The crucial part of (3.19) will turn out to be $O(\epsilon^2)$, which leads to the conclusion (1.13). We first compute $O(1)$ and $O(\epsilon)$ terms of (3.19).

LEMMA 3.3. *When* $\epsilon = 0$, (3.19) *is equivalent to*

$$u_\xi^*(s_0(y))[w_\xi^{-,0}(0,y) - w_\xi^{+,0}(0,y)] = q_0(y)[h] = 0, \tag{3.20}$$

$$Z_r^{-,0}(0,y) - Z_r^{+,0}(0,y) = (\Pi_- + \Pi_+)q_0 = 0, \tag{3.21}$$

where $[h] = h^+(v^*) - h^-(v^*)$.

Proof. Differentiating the representation of $w^{\pm,1}(\xi, y)$ at $\xi = 0$,

$$\begin{aligned} u_\xi^*(s_0)w_\xi^{\pm,0}(0,y) &= [\theta(y)u_\xi^*(s_0) - W^{\pm,0}(0,y)]u_{\xi\xi}^*(s_0) \\ &\quad - \int_{\pm\infty}^0 [f_u(u^*(\xi + s_0))W^{\pm,0}(0,y) - q_0(y)]u_\xi^*(\xi + s_0)d\xi \\ &= \theta(y)u_\xi^*(s_0)u_{\xi\xi}^*(s_0) + q_0(y)(u^*(s_0) - h^\pm(v^*)). \end{aligned}$$

Here we used the fact that $u_{\xi\xi\xi}^*(s_0) + f_u(u^*(\xi + s_0))u_\xi^*(\xi + s_0) = 0$. Thus we obtain (3.20). (3.21) is obvious. □

Lemma 3.3 implies that $q_0(y) \equiv 0$. Then we see that

$$W^{\pm,0}(x) \equiv 0 \equiv Z^{\pm,0}(x)$$

and $w^{\pm,0}$ is represented as

$$w^{\pm,0}(\xi, y) = u_\xi^*(\xi + s_0(y))\theta(y). \tag{3.22}$$

In the following, we omit the superscript $\pm$ of $w^{\pm,0}(\xi, y)$.

Next we consider the $O(\epsilon)$-term of (3.19).

LEMMA 3.4. $O(\epsilon)$*-terms of* (3.19) *are equivalent to*

$$u_\xi^*(s_0(y))[W_r^{-,0}(0,y) + w_\xi^{-,1}(0,y) - W_r^{+,0}(0,y) - w_\xi^{+,1}(0,y)] = q_1(y)[h] = 0, \tag{3.23}$$

$$Z_r^{-,1}(0,y) - Z_r^{+,1}(0,y) = (\Pi_- + \Pi_+)q_1 = 0. \tag{3.24}$$

Proof. In view of (3.22), the equation for $w^{\pm,1}$ is recast as

$$\begin{aligned} &w_{\xi\xi}^{\pm,1} + f_u(u^*(\xi + s_0))w^{\pm,1} \\ &\quad + (N-1)H(0,y)w_\xi^0 + f_u(u^*(\xi + s_0))W^{\pm,1}(0,y) + \tilde{f}_u^1 w_\xi^0 - q_1(y) = 0. \end{aligned}$$

Differentiating the representation of $w^{\pm,1}(\xi, y)$ at $\xi = 0$,

$$\begin{aligned} -u_\xi^*(s_0)w_\xi^{\pm,1}(0,y) &= \theta(y)\int_{\pm\infty}^0 [(N-1)H(0,y)u_{\xi\xi}^*(\xi + s_0) + \tilde{f}_u^1 u_\xi^*(\xi + s_0)] \\ &\quad \times u_\xi^*(\xi + s_0)d\xi - q_1(y)\int_{\pm\infty}^0 u_\xi^*(\xi + s_0)d\xi \\ &= \theta(y)[u_\xi^{\pm,1}(0,y)u_{\xi\xi}^*(s_0) - u_\xi^*(s_0)u_{\xi\xi}^{\pm,1}(0,y)] \\ &\quad - q_1(y)\int_{\pm\infty}^0 u_\xi^*(\xi + s_0)d\xi. \end{aligned}$$

Here we used the fact proved in Appendix D that

$$\begin{aligned}(3.25)\qquad &\int_{\pm\infty}^{0}[(N-1)H(0,y)u^*_{\xi\xi}(\xi+s_0)+\tilde{f}^1_u u^*_\xi(\xi+s_0)]u^*_\xi(\xi+s_0)d\xi\\ &\quad=u^{\pm,1}_\xi(0,y)u^*_{\xi\xi}(s_0)-u^*_\xi(s_0)u^{\pm,1}_{\xi\xi}(0,y).\end{aligned}$$

By using the facts that $u^{-,1}_\xi(0,y)=u^{+,1}_\xi(0,y)$ and $u^{-,1}_{\xi\xi}(0,y)=u^{+,1}_{\xi\xi}(0,y)$, we have

$$u^*_\xi(s_0(y))[w^{-,1}_\xi(0,y)-w^{+,1}_\xi(0,y)]=q_1(y)\int_{-\infty}^{\infty}u^*_\xi(\xi)d\xi=q_1(y)[h].$$

Thus we obtain (3.23). Equation (3.24) is obvious. □

Lemma 3.4 implies that $q_1(y)\equiv 0$, and then we see that

$$W^{\pm,1}(x)\equiv 0\equiv Z^{\pm,1}(x).$$

Let us define new functions $\tilde{\Phi}$ and $\tilde{\Psi}$ as follows:

$$\tilde{\Phi}(\theta,q_2,\lambda_1,\epsilon):=\frac{1}{\epsilon^2}\Phi(\theta,0,0,q_2,\lambda_1,\epsilon),\quad \tilde{\Psi}(\theta,q_2,\lambda_1,\epsilon):=\frac{1}{\epsilon^2}\Psi(\theta,0,0,q_2,\lambda_1,\epsilon).$$

LEMMA 3.5. *$\tilde{\Phi}(\theta,q_2,\lambda_1,0)=0$ and $\tilde{\Psi}(\theta,q_2,\lambda_1,0)=0$ are equivalent to*

$$(3.26)\qquad [m^2\Delta^{\Gamma^*}+m^2H^*(y)-V^1_r(0,y)J'(v^*)]\theta+J'(v^*)q_2(y)=0,$$

$$(3.27)\qquad (\Pi_-+\Pi_+)q_2-(1+\lambda_1)[h]\theta=0.$$

Proof. See Appendix C. □

Proof of Theorem 1.3. Using Lemma 3.5, we obtain (1.13).

4. Applications. In this section, we apply the results of the previous sections to the case where the domain Ω is a ball or a rectangle. In the following, we assume that the nonlinearity takes the form

$$f(u)=u-u^3.$$

First we note that the constants appeared in the previous section are computed as follows.

LEMMA 4.1.

$$v^*=0,\quad h^\pm(v^*)=\pm 1,\quad J'(v^*)=-2,\quad [h]=2,\quad m^2=\frac{2\sqrt{2}}{3}=\frac{4}{3\sqrt{2}}.$$

Proof. This is proved by straightforward computation. □

4.1. Stability of radially symmetric solutions. In this subsection, we study the radially symmetric solution when Ω is a ball of radius R. It is convenient to introduce new coordinate system $x=(r,y)$, where $x=ry$ for $r\in[0,R]$ and $y\in\partial\Omega=S^{N-1}$.

Then the rescaled reduced problem (1.7)–(1.9) can be rewritten as

$$(4.1)_+\qquad \begin{cases} V^+_{rr}+\dfrac{N-1}{r}V^+_r+G^+=0, & r_0<r<R,\\ V^+(r_0)=\dfrac{m^2(N-1)}{J'(v^*)r_0}, & V^+_r(1)=0,\end{cases}$$

$$(4.1)_- \qquad \begin{cases} V^-_{rr} + \dfrac{N-1}{r} V^-_r + G^- = 0, & 0 < r < r_0, \\ V^-_r(0) = 0, & V^-(r_0) = \dfrac{m^2(N-1)}{J'(v^*) r_0}, \end{cases}$$

and

$$V^-_r(r_0) = V^+_r(r_0), \tag{4.2}$$

where $G^{\pm} := \pm 1 - \overline{u}$. Here $V^-(r,y)$, $V^+(r,y)$ and r_0 are unknown functions and a parameter, respectively. The solutions of $(4.1)_{\pm}$ have the following expressions:

$$V^+(r) = \frac{m^2(N-1)}{J'(v^*) r_0} + \frac{G^+}{N} \int_{r_0}^{r} (R^N t^{1-N} - t) dt,$$
$$V^-(r) = \frac{m^2(N-1)}{J'(v^*) r_0} + \frac{G^-}{2N} (r_0^2 - r^2).$$

Then, by using the condition (4.2), r_0 is uniquely determined as

$$r_0 = \left(\frac{G^+}{[h]} \right)^{1/N} R,$$

and the interface Γ^* is defined by

$$\Gamma^* = \{ x \in \mathbf{R}^N \mid |x| = r_0 \}.$$

The following existence theorem can be proved in a similar way as in [5], so we omit it (see also [19]).

THEOREM 4.2. *There exists a constant $\epsilon_0 > 0$ such that* (1.5) *and* (1.6) *have an ϵ-family of radially symmetric layer solutions $(u^\epsilon(r), v^\epsilon(r))$ for $\epsilon \in (0, \epsilon_0]$ satisfying the following:*

(i) $\lim_{\epsilon \to 0} v^\epsilon = v^*$ *uniformly on* Ω.

(ii) *For each $\delta > 0$,*

$$\lim_{\epsilon \to 0} u^\epsilon(r) = \begin{cases} -1, & 0 \le r \le r_0 - \delta, \\ 1, & r_0 + \delta \le r \le R. \end{cases}$$

The asymptotic form of $(U^\epsilon(r), v^\epsilon(r))$ is given in section 2.

Using Lemma 4.1 and the above results, we can find that the operator L is recast as

$$L = \frac{1}{r_0^2} \Delta^S + \frac{N-1}{r_0^2} + \frac{3\sqrt{2}}{2} \frac{1+\overline{u}}{N} r_0 - 3\sqrt{2} \mathcal{T}(\cdot).$$

We prepare two lemmas before we state the key proposition.

LEMMA 4.3. *All the eigenvalues of $\mathcal{T}$ and L are real. More precisely,*

(i) *the jth eigenvalue Λ_j of $\mathcal{T}$ is given by*

$$\Lambda_j = \alpha R \hat{\Lambda}(j, \alpha), \tag{4.3}$$

where

$$\hat{\Lambda}(z,\alpha) = \frac{(z+N-2)\alpha^{2z+N-2} + z}{z(2z+N-2)}, \quad r_0 = \alpha R, \quad \alpha = \alpha(\overline{u}) = \left(\frac{1-\overline{u}}{2} \right)^{1/N}. \tag{4.4}$$

Then the associated eigenfunction $\beta_j^m(y)$ is the harmonic function of degree m.

(ii) *The jth eigenvalue L is real and given by $\Sigma_j = \Sigma(j, \alpha, R)$, where*

$$\Sigma(z,\alpha,R) = \frac{1}{\alpha^2 R^2}[-z^2-(N-1)z+N-1] + 3\sqrt{2}\left[\frac{1-\alpha^N}{N} - \hat{\Lambda}(z,\alpha)\right]\alpha R. \tag{4.5}$$

Then the associated eigenfunction is the same as that of $\mathcal{T}$.

Proof. See Appendix E. □

LEMMA 4.4 (properties of $\hat{\Lambda}(z,\alpha)$ and $\Sigma(z,\alpha,R)$).

(i) *For $z \geq 1$ and $\alpha \in (0,1)$,*

$$\frac{\partial}{\partial z}\hat{\Lambda}(z,\alpha) < 0, \quad \frac{\partial^2}{\partial z^2}\hat{\Lambda}(z,\alpha) > 0.$$

(ii)

$$\begin{aligned}\frac{\partial}{\partial z}\Sigma(1,\alpha,R) =& -\frac{N}{\alpha^2R^2}\\ &-\frac{3\sqrt{2}\alpha R}{N^2}[-\alpha^N\{2+4(N-2)+(N-2)^2\}+2(N-1)N\alpha^N\log\alpha-2],\\ \Sigma(1,\alpha,R) =& -3\sqrt{2}\alpha^{N+1}R<0 \quad \textit{for } \alpha\in(0,1),\\ &\frac{\partial^2}{\partial z^2}\Sigma(z,\alpha,R)<0 \quad \textit{for } z\geq 1 \quad \textit{and} \quad \alpha\in(0,1).\end{aligned}$$

Proof. These results can be obtained by direct calculations. □

The following proposition is a key to prove Theorem 1.4.

PROPOSITION 4.5.

(i) *The jth principal eigenvalue λ_j^* of (1.13) is given by the following form.*

$$\lambda_j^* = \lambda_j^*(\alpha,\overline{u}) = \frac{1}{3\sqrt{2}}\cdot\frac{\Sigma(j,\alpha(\overline{u}),R)}{\alpha R\hat{\Lambda}(j,\alpha(\overline{u}),R)}. \tag{4.6}$$

Moreover, it holds that

(ii) *for any fixed $R\in(0,\infty)$, there exists $\overline{u}_0 = \overline{u}_0(R)\in(-1,1)$ such that*

$$\Sigma(z,\alpha(\overline{u}),R)<0 \quad \textit{for} \quad \overline{u}\in(0,\overline{u}_0) \quad \textit{and} \quad z\geq 1.$$

(iii) *For any fixed $\overline{u}\in(-1,1)$, there exist $R_0 = R_0(\overline{u}) > 0$ and integer $z_0 \geq 1$ such that*

$$\Sigma(z_0,\alpha(\overline{u}),R_0) > 0.$$

Proof. (i) Noting that L and $\mathcal{T}$ have the same eigenfunctions $\{\beta_j(y)\}_{j=1}^{\infty}$, we can rewrite (1.13) as

$$L\beta_j = 3\sqrt{2}\lambda^*\mathcal{T}(\beta_j),$$

which leads to (4.6).

(ii) Note that $\alpha(\overline{u})$ is a monotone decreasing function of $\overline{u}$. Since $\frac{\partial}{\partial z}\Sigma(1,\alpha,R)<0$ and $\frac{\partial^2}{\partial z^2}\Sigma(z,\alpha,R)<0$ for sufficiently small α, we obtain (ii).

(iii) For fixed α, we choose z_0 satisfying $\frac{1-\alpha^N}{N} - \hat{\Lambda}(z_0,\alpha) > 0$. Then the sign of the first term of (4.5) is negative and that of the second one is positive. Therefore, for sufficiently large R, it holds that $\Sigma(z_0,\alpha,R) > 0$. □

Proof of Theorem 1.4. It is a direct consequence of Proposition 4.5.

4.2. Stability of planar solutions. Let Ω be a rectangle in (x,y)-plane $\Omega := (0,X)\times(0,Y)$. Then the rescaled reduced problem (1.7)–(1.9) is recast as

$$(4.7)_+ \qquad \begin{cases} V^+_{xx} + G^+ = 0, & x_0 < x < X, \\ V^+(x_0) = 0, & V^+_x(X) = 0, \end{cases}$$

$$(4.7)_- \qquad \begin{cases} V^-_{xx} + G^- = 0, & 0 < x < x_0, \\ V^-_x(0) = 0, & V^-(x_0) = 0, \end{cases}$$

and

$$V^-_x(x_0) = V^+_x(x_0), \tag{4.8}$$

where $G^\pm := \pm 1 - \overline{u}$, $\Omega_+(\Gamma^*) = \{(x,y)\in\mathbf{R}^2 \mid x_0 < x < X,\ 0 < y < Y\}$, $\Omega_-(\Gamma^*) = \{(x,y)\in\mathbf{R}^2 \mid 0 < x < x_0, 0 < y < Y\}$, and $\Gamma^* = \{(x_0,y)\in\mathbf{R}^2 \mid 0 < y < Y\}$. Here we used the fact that $H_1(y) = 0$. We can easily show that $(4.7)_-$ and $(4.7)_+$ have unique solutions given by

$$\begin{aligned} V^-(x) &= -\frac{1}{2}G^-[x^2 - x_0^2], \\ V^+(x) &= -G^+\left[\frac{1}{2}(x^2 - x_0^2) - X(x - x_0)\right]. \end{aligned}$$

Then, by using the C^1-matching condition (4.8), we can uniquely determine x_0 as

$$x_0 = \frac{G^+ X}{[h]},$$

so the derivative $V_x(x_0)$ is given by

$$V_x(x_0) = -\frac{G^+G^-X}{[h]} = \frac{1}{2}X(1-\overline{u})(1+\overline{u}).$$

The existence results of the planar solution to (1.5) and (1.6) is given by Taniguchi and Nishiura [23].

THEOREM 4.6 (see [23]). *There exists a constant $\epsilon_0 > 0$ such that* (1.5) *and* (1.6) *have an ϵ-family of stationary planar solutions $(u^\epsilon(x), v^\epsilon(x))$ independent of $y\in[0,Y]$ for $\epsilon\in(0,\epsilon_0]$ satisfying the following:*

(i) $\lim_{\epsilon\to 0} v^\epsilon = v^*$ *uniformly on* Ω.

(ii) *For each $\delta > 0$,*

$$\lim_{\epsilon\to 0} u^\epsilon(x) = \begin{cases} -1, & 0 \le x \le x_0 - \delta, \\ 1, & x_0 + \delta \le x \le X. \end{cases}$$

Using Lemma 4.1 and the above results, we can find that the operator L is recast as

$$L = \frac{d^2}{dy^2} - \frac{3\sqrt{2}}{2}\left[\frac{1}{2}X(\overline{u}-1)(\overline{u}+1) + 2\mathcal{T}(\cdot)\right].$$

The spectral properties of L and $\mathcal{T}$ are given in the following two lemmas.

LEMMA 4.7. *All the eigenvalues of $\mathcal{T}$ and L are real. More precisely,*

(i) *the jth eigenvalue Λ_j of $\mathcal{T}$ is given by*

$$\Lambda_j = X\Lambda(\kappa\pi j, \overline{u}), \tag{4.9}$$

where

$$\Lambda(z, \overline{u}) := \frac{\cosh z + \cosh \overline{u} z}{2z \sinh z} \tag{4.10}$$

and $\kappa := X/Y$. Then the associated eigenfunction is $\beta_j(y) = \cos(\tau_j y)$, where $\tau_j = \pi j/Y$.

(ii) *The jth eigenvalue of L is real and given by $\Sigma_j = \Sigma(\kappa\pi j, X, \overline{u})$, where*

$$\Sigma(z, X, \overline{u}) := -\left(\frac{z}{X}\right)^2 + \frac{3\sqrt{2}}{2} X \left[\frac{1}{2}(1-\overline{u})(1+\overline{u}) - 2\Lambda(z, \overline{u})\right].$$

Then the associated eigenfunction is the same as that of $\mathcal{T}$.

Proof. See Appendix F. □

LEMMA 4.8 (properties of $\Lambda(z, \overline{u})$). *For $\overline{u} \in (-1, 1)$ and $z > 0$,*

$$\Lambda(z, \overline{u}) > 0, \quad \frac{\partial}{\partial z}\Lambda(z, \overline{u}) < 0, \quad \frac{\partial^2}{\partial z^2}\Lambda(z, \overline{u}) > 0,$$
$$\lim_{z \to +0} \Lambda(z, \overline{u}) = \infty, \quad \lim_{z \to \infty} z\Lambda(z, \overline{u}) = \frac{1}{2}.$$

Proof. These results can be obtained by direct calculations. □

Using the above two lemmas, we can prove the following proposition.

PROPOSITION 4.9. (i) *The jth principal eigenvalue λ_j^* of (1.13) is given by the following form:*

$$\lambda_j^* = \lambda_j^*(\kappa, X, \overline{u}) = \frac{1}{3\sqrt{2}} \cdot \frac{\Sigma(\kappa\pi j, X, \overline{u})}{X\Lambda(\kappa\pi j, \overline{u})}. \tag{4.11}$$

(ii) *The nullcline of $\Sigma(z, X, \overline{u})$ as a function of $z > 0$ and $X > 0$ is given by $\{(z, X(z, \overline{u})) \mid z > z_0\}$, where z_0 is a unique zero of $\frac{1}{2}(1-\overline{u}^2) - 2\Lambda(z, \overline{u}) = 0$ and*

$$X(z; \overline{u}) = \left(\frac{\sqrt{2}}{3}\right)^{1/3} z^{2/3} \left[\frac{1}{2}(1-\overline{u}^2) - 2\Lambda(z, \overline{u})\right]^{-1/3}. \tag{4.12}$$

Moreover, $X(z; \overline{u})$ has the following properties:

$$\lim_{z \to z_0 + 0} X(z; \overline{u}) = \infty, \tag{4.13}$$

$$\lim_{z \to \infty} \left[X(z; \overline{u}) - \left(\frac{2\sqrt{2}}{3}\right)^{1/3} (1-\overline{u}^2)^{-1/3} z^{2/3}\right] = 0, \tag{4.14}$$

$$\frac{dX}{dz} \quad \begin{cases} < 0 & \text{for} \quad z_0 < z < z_1, \\ = 0 & \text{for} \quad z = z_1, \\ > 0 & \text{for} \quad z_1 < z, \end{cases} \tag{4.15}$$

for some $z_1 = z_1(\overline{u})$.

Proof. (i) Noting that L and $\mathcal{T}$ have the same eigenfunctions $\{\beta_j(y)\}_{j=1}^{\infty}$, we can rewrite (1.13) as

$$L\beta_j = 3\sqrt{2}\lambda^*\mathcal{T}(\beta_j),$$

which leads to (4.11).

(ii) We define a function g of z and X on $(0,\infty)\times(0,\infty)$ by

$$g(z,X,\overline{u}) := X^2\Sigma(z,X,\overline{u}) = -z^2 + \frac{3\sqrt{2}}{2}X^3\left[\frac{1}{2}(1-\overline{u}^2) - \Lambda(z,\overline{u})\right].$$

Using Lemma 4.8, we see that $p(z,\overline{u}) := \frac{1}{2}(1-\overline{u}^2)-\Lambda(z,\overline{u})$ has a unique zero $z = z_0(\overline{u})$ and $p(z,\overline{u}) > 0$ for $z > z_0$. Then we can solve $g(z,X,\overline{u}) = 0$ in X as (4.12). Noting the properties of $\Lambda(z,\overline{u})$, we obtain (4.13) and (4.14).

By using the implicit function theorem, we see that

$$\frac{dX}{dz} = -\frac{\partial g}{\partial z}\Big/\frac{\partial g}{\partial X}$$

and $\frac{\partial g}{\partial X} = \frac{9\sqrt{2}}{2}X^2[(1-\overline{u}^2)/2-\Lambda(z,\overline{u})] > 0$. Noting that $\frac{\partial g}{\partial z} = -2z-3\sqrt{2}X^3\Lambda_z$, (4.13) and (4.14), we have

$$\lim_{z\to z_0+0}\frac{\partial g}{\partial z}(z,X(z;\overline{u});\overline{u}) = \infty, \quad \lim_{z\to\infty}\frac{\partial g}{\partial z}(z,X(z;\overline{u});\overline{u}) = -\infty.$$

Combining the above facts and $\frac{\partial^2 g}{\partial z^2} = -2-3\sqrt{2}X^3\Lambda_{zz} < 0$, we can see that $\frac{\partial g}{\partial z}$ has a unique zero $z = z_1$ and $\frac{dX}{dz}$ satisfies (4.15). ∎

Proof of Theorem 1.5. The proof follows from Proposition 4.9. ∎

4.3. Justification of the SLEP equation for the planar case. In subsection 3.3, we derived the SLEP equation for the principal parts of the critical eigenvalues and eigenfunctions. Since our method is constructive, it is not a priori clear that there are no other dangerous eigenvalues. But, fortunately, when the domain Ω is a ball or a rectangle, we can justify our results. Precisely speaking, the principal parts of all dangerous eigenvalues are reduced to the solutions to (1.13).

In this subsection, we outline the justification of (1.13) when Ω is a rectangle $(0,X)\times(0,Y)$. When Ω is a ball, we can justify (1.13) in a parallel way. See [19] and [20] for the activator-inhibitor case.

By using a complete orthonormal system $\{\Phi_j(y)\}_{j=0}^{\infty}$ in $L^2(0,Y)$, where

$$\Phi_j(y) = \begin{cases} 1/\sqrt{Y}, & j = 0, \\ \sqrt{2/Y}\cos(\pi jy/Y), & j > 0, \end{cases}$$

(w,z) is expanded as

$$w(x,y) = \sum_{j=0}^{\infty} w_j(x)\Phi_j(y), \quad z(x,y) = \sum_{j=0}^{\infty} z_j(x)\Phi_j(y),$$

where

$$w_j(x) = \int_0^Y w(x,y)\Phi_j(y)dy, \quad z_j(x) = \int_0^Y z(x,y)\Phi_j(y)dy$$

in $L^2(\Omega)$. By using the above notation, the eigenvalue problem (3.1) is rewritten as

$$\begin{cases} 0 = L^{\epsilon,j} w_j - z_j, \\ 0 = M^j z_j + \epsilon w_j + \lambda w_j, \end{cases} \qquad x \in I := (0, X), \tag{4.16}$$

with the boundary condition

$$\frac{dw_j}{dx}(0) = 0 = \frac{dw_j}{dx}(X), \quad \frac{dz_j}{dx}(0) = 0 = \frac{dz_j}{dx}(X)$$

for $j = 0, 1, \ldots$, where

$$L^{\epsilon,j} := \epsilon^2 \frac{d^2}{dx^2} + f_u^\epsilon - \epsilon^2 \mu_j, \quad M^j := \frac{d^2}{dx^2} - \mu_j, \quad \mu_j = \left(\frac{\pi j}{Y}\right)^2.$$

Then the condition $\iint_\Omega w \, dx \, dy = 0$ is equivalent to either

$$\text{(i)} \qquad \int_0^X w_0(x) \, dx = 0 \quad (w_0(x) \not\equiv 0)$$

or

$$\text{(ii)} \qquad w_0(x) \equiv 0$$

since $\int_0^Y \Phi_j(y) dy = 0$ for $j = 1, 2, \ldots$.

First, we consider the former case. Let $\{\phi_k^{\epsilon,j}\}_{k \geq 0}$ be the complete orthonormal set in $L^2(I)$ consisting of the eigenfunctions of $L^{\epsilon,j}$, and $\{\zeta_k^{\epsilon,j}\}_{k \geq 0}$ the associated eigenvalues. They have the following properties.

LEMMA 4.10 (Nishiura [10] and [11]).

(i) *It holds that*

$$\zeta_0^{\epsilon,j} > 0 > -\delta > \zeta_1^{\epsilon,j} > \zeta_2^{\epsilon,j} > \cdots$$

for sufficiently small $\epsilon > 0$, where δ is a positive constant independent of $j \geq 0$ and $\epsilon > 0$.

(ii) $\lim_{\epsilon \downarrow 0} \frac{\zeta_0^{\epsilon,j}}{\epsilon^2} = \hat{\zeta}_0^* - \mu_j$, *where* $\hat{\zeta}_0^* = -\frac{1}{m^2} J'(v^*) V_x(x_0)$.

(iii) $\lim_{\epsilon \downarrow 0} \frac{\phi_0^{\epsilon,j}}{\sqrt{\epsilon}} = \frac{[h]}{m} \delta_{x_0}$ *in* $H^{-1}(I)$*-sense,*

where δ_{x_0} is a Dirac's δ-function at $x = x_0$.

We decompose w as

$$w = (L^{\epsilon,j})^{-1} z = \frac{\langle z, \ \phi_0^{\epsilon,j} \rangle}{\zeta_0^{\epsilon,j}} \phi_0^{\epsilon,j} + (L^{\epsilon,j})^\dagger(z),$$

where

$$(L^{\epsilon,j})^\dagger := \sum_{n \geq 1} \frac{\langle \ \cdot \ , \ \phi_n^{\epsilon,j} \rangle}{\zeta_n^{\epsilon,j}} \phi_n^{\epsilon,j}.$$

$(L^{\epsilon,j})^\dagger$ has the following properties.

LEMMA 4.11 (Nishiura [10] and [11]). *There exists a constant $\epsilon_0 > 0$ such that $(L^{\epsilon,j})^\dagger$ is a uniformly L^2-bounded operator for $\epsilon \in (0, \epsilon_0)$ and $j \geq 0$. Moreover, the following property holds:*

$$\lim_{\epsilon \to 0} (L^{\epsilon,j})^\dagger p = \frac{p}{f_u^*} \quad \textit{strongly in } L^2 - \textit{sense},$$

where $p \in L^2(I) \cap L^\infty(I)$ and $f_u^* := f_u(h^\pm(v^*))$.

The eigenvalue problem (4.16) and the condition $\iint_\Omega w\, dx\, dy = 0$ with $j = 0$ are recast as

$$z_{xx} + (\epsilon + \lambda) \left[\frac{\langle z,\ \phi_0^{\epsilon,0} \rangle}{\zeta_0^{\epsilon,0}} \phi_0^{\epsilon,0} + (L^{\epsilon,0})^\dagger(z) \right] = 0 \tag{4.17}$$

and

$$\frac{\langle z,\ \phi_0^{\epsilon,0}/\sqrt{\epsilon} \rangle}{\epsilon} \int_0^X \frac{1}{\zeta_0^\epsilon/\epsilon^2} \frac{\phi_0^{\epsilon,0}}{\sqrt{\epsilon}} dx + \int_0^X (L^{\epsilon,0})^\dagger(z) dx = 0.$$

In view of Lemmas 4.10 and 4.11, we see that

$$\lim_{\epsilon \to 0} \int_0^X \frac{1}{\zeta_0^\epsilon/\epsilon^2} \frac{\phi_0^{\epsilon,0}}{\sqrt{\epsilon}} dx = \frac{[h]}{m\hat{\zeta}_0^*}$$

and $\lim_{\epsilon \to 0} \int_0^X (L^{\epsilon,0})^\dagger(z) dx$ is bounded. Consequently, the following limits exist:

$$\hat{z}^* := \frac{m}{[h]} \lim_{\epsilon \downarrow 0} \frac{1}{\epsilon} \langle z,\ \phi_0^{\epsilon,0}/\sqrt{\epsilon} \rangle, \quad 0 = \lim_{\epsilon \downarrow 0} \langle z,\ \phi_0^{\epsilon,0}/\sqrt{\epsilon} \rangle.$$

Now we rewrite (4.17) in a weak form

$$- \langle z_x, \psi_x \rangle + (\epsilon + \lambda) \left[\frac{\langle z,\ \phi_0^{\epsilon,0}/\sqrt{\epsilon} \rangle/\epsilon}{\zeta_0^\epsilon/\epsilon^2} \langle \phi_0^{\epsilon,0}/\sqrt{\epsilon}, \psi \rangle + \langle (L^{\epsilon,0})^\dagger(z), \psi \rangle \right] = 0,$$

$$z \in H_N^1(I),\ \psi \in H^1(I),$$

where $H_N^1(I)$ is the space of closure of $\{\cos(n\pi x/X)\}_{n=0}^\infty$ in $H^1(I)$. Then the limit function $z^* = \lim_{\epsilon \to 0} z$ must exist and satisfy the following limit equation:

$$-\langle z_x^*, \psi_x \rangle + \lambda \left[\frac{[h]^2 \hat{z}^*}{m^2 \hat{\zeta}_0^*} \langle \delta_{x_0}, \psi \rangle + \frac{1}{f_u^*} \langle z^*, \psi \rangle \right] = 0, \quad z^*(x_0) = 0.$$

This is equivalent to the next system:

$$\begin{cases} Z_{xx}^- + \dfrac{\lambda}{f_u^*} Z^- = 0, & x \in (0, x_0), \\ Z_x^-(0) = 0, & Z^-(x_0) = 0, \end{cases} \tag{4.18}$$

$$\begin{cases} Z_{xx}^+ + \dfrac{\lambda}{f_u^*} Z^+ = 0, & x \in (x_0, X), \\ Z^+(x_0) = 0, & Z_x^+(X) = 0, \end{cases} \tag{4.19}$$

$$Z_x^+(x_0) - Z_x^-(x_0) = -\lambda \frac{[h]^2 \hat{z}^*}{m^2 \hat{\zeta}_0^*}. \tag{4.20}$$

LEMMA 4.12. *There exists a constant $\delta > 0$ such that* (4.18)–(4.20) *have no nontrivial solutions for $\lambda > -\delta$ for the case* (i)*, i.e., $\int_0^X w_0(x)\,dx = 0$ ($w_0(x) \not\equiv 0$).*

Proof. When $\lambda = 0$, (4.18)–(4.20) have a unique solution $Z^-(x) \equiv 0$ and $Z^+(x) \equiv 0$.

If $\lambda > 0$, the general solution $Z(x)$ of (4.18) is represented as

$$Z(x) = A \exp(\alpha x) + B \exp(-\alpha x),$$

where $\alpha^2 = -\lambda / f_u^*$, $\alpha > 0$. Then, by using the boundary conditions, we obtain $A = 0$ and $B = 0$. Similarly, we can prove that (4.19) has no nonhomogeneous solutions.

If $-\delta < \lambda < 0$, where $-\delta := \frac{\pi^2}{X^2} f_u^*$, the general solutions $Z^-(x) = C^- \cos \beta x$ and $Z^+(x) = C^+ \cos \beta(X - x)$, where $\beta^2 = \lambda / f_u^*$ and $\beta > 0$, cannot satisfy the boundary conditions $Z^-(x_0) = 0$ and $Z^+(x_0) = 0$ simultaneously. □

Next, we consider the case (ii) $w_0(x) \equiv 0$. The eigenvalue problem (4.16) with $j > 0$ is recast as

$$-\langle z_x, \psi_x \rangle - \mu_j \langle z,\ \psi \rangle + \frac{\epsilon + \lambda}{\epsilon} \left[\frac{\langle z,\ \phi_0^{\epsilon,j}/\sqrt{\epsilon} \rangle}{\zeta_0^{\epsilon,j}/\epsilon^2} \langle \phi_0^{\epsilon,j}/\sqrt{\epsilon}, \psi \rangle + \epsilon \langle (L^{\epsilon,j})^\dagger(z), \psi \rangle \right] = 0,$$

where $z \in H_N^1(I)$, $\psi \in H^1(I)$.

By Lemmas 4.10 and 4.11, we see that $\lambda = O(\epsilon)$ and the limit function $z^* = \lim_{\epsilon \to 0} z$ must exist. Then z^* satisfies the following limit equation:

$$-\langle z_x^*, \psi_x \rangle - \mu_j \langle z^*,\ \psi \rangle + (1 + \hat{\lambda}) \frac{[h]^2}{m^2} \frac{\langle z^*, \delta_{x_0} \rangle}{\hat{\zeta}_0^* - \mu_j} \langle \delta_{x_0}, \psi \rangle = 0, \tag{4.21}$$

where $\hat{\lambda} = \lim_{\epsilon \to 0} \lambda / \epsilon$. Hereafter, we normalize the limit eigenfunction z^* as $\langle z^*, \delta_0 \rangle = 1$. Then (4.21) is equivalent to the next system:

$$\begin{cases} z_{xx}^* - \mu_j z^* = 0, \quad x \in (0, x_0) \cup (x_0, X), \\ \lim\limits_{x \to x_0 + 0} z_x^*(x) - \lim\limits_{x \to x_0 - 0} z_x^*(x) = -(1 + \hat{\lambda}) \dfrac{[h]^2}{m^2 (\hat{\zeta}_0^* - \mu_j)}, \\ z^*(x_0) = 1, \quad z_x^*(0) = 0 = z_x^*(X). \end{cases} \tag{4.22}$$

PROPOSITION 4.13. *The eigenvalue $\hat{\lambda}$ of* (4.22) *is given by* (4.11) *for the case* (ii)*, i.e., $w_0(x) \not\equiv 0$.*

Proof. Note that z^* is determined by the first equation with the third and fourth condition of (4.22). In fact, after simple computations, we have

$$z^*(x) = \begin{cases} \dfrac{\cosh \tau_j x}{\cosh \tau_j x_0}, & 0 < x < x_0, \\ \dfrac{\cosh \tau_j (x - X)}{\cosh \tau_j (X - x_0)}, & x_0 < x < X. \end{cases}$$

Here we used the facts that $\mu_j = \tau_j^2 = (\pi j/Y)^2$. Then

$$\begin{aligned}\lim_{x\to x_0+0} z^*(x) - \lim_{x\to x_0-0} z^*(x) &= \frac{\tau \sinh \tau_j(x-X)}{\cosh \tau_j(X-x_0)} - \frac{\tau \sinh \tau_j x}{\cosh \tau_j x_0} \\ &= -\frac{1}{X}\frac{2\tau_j X \sinh \tau_j X}{\cosh \tau_j X + \cosh \overline{u}\tau_j X} \\ &= -\frac{1}{X\Lambda(\kappa\pi j, \overline{u})}.\end{aligned}$$

Solving the second equation of (4.22) in $\hat{\lambda}$, we have

$$\hat{\lambda} = \frac{m^2}{[h]^2 X\Lambda(\kappa\pi j, \overline{u})}\left[-\mu_j + \hat{\zeta}_0^* - \frac{[h]^2}{m^2} X\Lambda(\kappa\pi j, \overline{u})\right].$$

Noting the facts that

$$\kappa = \frac{X}{Y}, \quad \mu_j = \left(\frac{\kappa\pi j}{X}\right)^2, \quad \hat{\zeta}_0^* = \frac{3\sqrt{2}}{4} X(1-\overline{u})(1+\overline{u}), \quad \frac{[h]^2}{m^2} = 3\sqrt{2},$$

we obtain (4.11). Thus we complete the proof. □

In view of Lemma 4.12 and Proposition 4.13, it is clear that all dangerous eigenvalues to stability are controlled by the SLEP equation (1.13) (or, equivalently, (4.22)).

5. Concluding remarks—activator-inhibitor case. In the preceding sections, we have discussed the stability of mesoscopic patterns in diblock copolymers. Here we show that our approach also works well for the activator-inhibitor system (1.4) under the Neumann boundary condition. We assume the following for the nonlinearity $g(u, v)$:

$$\text{(A7)} \qquad \pm g(h^{\pm}(v^*), v^*) > 0, \quad g_u(h^{\pm}(v^*), v^*) > 0, \quad \frac{d}{dv} g(h^{\pm}(v), v)\Big|_{v=v^*} < 0.$$

In the following, we will use the same notation defined in the previous sections without explanation. The reduced problem to (1.4) is given by

$$\text{(5.1)} \qquad \begin{cases} D\Delta V^- = -g(h^-(v^*), v^*) & \text{in } \Omega_-(\Gamma), \\ V^- = -\dfrac{m^2(N-1)}{J'(v^*)} H(0, y) & \text{on } \Gamma, \\ \dfrac{\partial V^-}{\partial n} = 0 & \text{on } \partial\Omega, \end{cases}$$

$$\text{(5.2)} \qquad \begin{cases} D\Delta V^+ = -g(h^+(v^*), v^*) & \text{in } \Omega_+(\Gamma), \\ V^+ = -\dfrac{m^2(N-1)}{J'(v^*)} H(0, y) & \text{on } \Gamma, \end{cases}$$

$$\text{(5.3)} \qquad \frac{\partial V^+}{\partial \nu} = \frac{\partial V^-}{\partial \nu} \quad \text{on } \Gamma.$$

Concerning the existence of stationary solutions, we have the following corollary.

COROLLARY OF THEOREM 1.2. *Assume that* (A1)–(A3) *in section* 1, (A7)*, and the following* (A4′)–(A5′) *are satisfied:*

(A4′) *There exists a solution* $(V^*, \Gamma^*) \in C^1(\overline{\Omega}) \times \mathcal{F}$ *of* (5.1)–(5.3)*, where*

$$V^*(x) = \begin{cases} V^-(x), & x \in \Omega_-(\Gamma^*), \\ V^+(x), & x \in \Omega_+(\Gamma^*). \end{cases}$$

$V^{\pm}$ *are smooth solutions of* (5.1) *and* (5.2)*, respectively, satisfying* (5.3).

(A5′) *There is a bounded linear operator* $L^\dagger : (I-P)C^\alpha(\Gamma^*) \to (I-P)C^{2,\alpha}(\Gamma^*)$ *called the inverse of the operator*

$$L := \Delta^{\Gamma^*} + H^*(y) - \frac{1}{m^2} J'(v^*) V_r^*(0,y) + \frac{1}{Dm^2}[g] J'(v^*) \mathcal{T}(\cdot)$$

such that $LL^\dagger = I$ *on* $(I-P)C^\alpha(\Gamma^*)$ *and* $L^\dagger L = I - P$ *on* $(I-P)C^{2,\alpha}(\Gamma^*)$. *Here,* Δ^{Γ^*} *is the Laplace–Beltrami operator on* Γ^*, $H^*(y)$ *sum of the square of the principal curvature of* Γ^*, *and* $[g] := g(h^+(v^*), v^*) - g(h^-(v^*), v^*)$.

Then, there is an $\epsilon_0 > 0$ *such that* (1.4) *have an* ϵ*-family of stationary solutions* (u^ϵ, v^ϵ) *for* $\epsilon \in (0, \epsilon_0]$ *satisfying the following:*

(i) $\lim_{\epsilon \to 0} v^\epsilon(x) = v^*$ *uniformly on* $\overline{\Omega}$.

(ii) *For each* $\delta > 0$,

$$\lim_{\epsilon \to 0} u^\epsilon(x) = \begin{cases} h^-(v^*), & x \in \overline{\Omega}_-(\Gamma^*) \backslash \Gamma^*_\delta, \\ h^+(v^*), & x \in \overline{\Omega}_+(\Gamma^*) \backslash \Gamma^*_\delta, \end{cases} \quad \textit{uniformly},$$

where Γ^*_δ *is a tubular neighborhood of* Γ^*.

(iii) *For each* $K > 0$,

$$\lim_{\epsilon \to 0} u^\epsilon(y + \epsilon \xi \nu(y)) = u^*(\xi + s^*(y)) \quad \textit{in } C^2[-K, K],$$

uniformly in $y \in \Gamma^*$ *for some* $s^* \in C^{2,\alpha}(\Gamma^*)$.

The associated linearized problem of (1.4) is of the form

$$\begin{cases} \lambda^\epsilon w = \epsilon^2 \Delta w + f_u^\epsilon w - z, \\ \epsilon \lambda^\epsilon z = D \Delta z + \epsilon g_u^\epsilon w + \epsilon g_v^\epsilon z, & \text{in } \Omega, \\ \dfrac{\partial w}{\partial n} = 0 = \dfrac{\partial z}{\partial n} & \text{on } \partial\Omega. \end{cases} \tag{5.4}$$

Here we focus only on the critical eigenvalues. So we assume the following for (5.4):

(A6′) There exists an integer $m^* \geq 1$ such that each eigenvalue λ^ϵ and the associated eigenfunctions (w^ϵ, z^ϵ) of (5.4) have the following asymptotic forms:

$$\begin{cases} \lambda^\epsilon = \epsilon \lambda_1 + \epsilon^2 \lambda_2 + o(\epsilon^2), \\ w^\epsilon(x) = \displaystyle\sum_{j=0}^{m^*} \epsilon^j W^{\pm,j}(x) + \omega(r) \cdot \sum_{j=0}^{m^*} \epsilon^j w^{\pm,j}(r/\epsilon, y) + o(\epsilon^{m^*}), \\ & \text{in } \Omega_\pm(\Gamma^*), \\ z^\epsilon(x) = \displaystyle\sum_{j=0}^{m^*} \epsilon^j Z^{\pm,j}(x) + \omega(r) \cdot \epsilon^2 \sum_{j=0}^{m^*} \epsilon^j z^{\pm,j}(r/\epsilon, y) + o(\epsilon^{m^*}), \end{cases}$$

where $W^{\pm,i}(x) \in C^{2,\alpha}(\overline{\Omega}_\pm(\Gamma^*))$ and $Z^{\pm,i}(x) \in C^{2,\alpha}(\overline{\Omega}_\pm(\Gamma^*))$ are outer expansions, $w^{\pm,i}(\xi, y)$ and $z^{\pm,j}(\xi, y)$ are inner expansions bounded for $\pm\xi \in (0, \infty)$ and $y \in \Gamma^*$.

In the activator-inhibitor case, the principal parts of C^1-matching conditions to eigenfunctions become

$$u_\xi^*(s_0(y))[w_\xi^{-,0}(0,y) - w_\xi^{+,0}(0,y)] = m^2\lambda_1\theta(y) - J'(v^*)q_0(y) = 0,$$

$$Z_r^{-,0}(0,y) - Z_r^{+,0}(0,y) = (\Pi_- + \Pi_+)q_0 = 0.$$

Then, the following two cases are considered.

Case I. $q_c \neq 0$ and θ is a constant function $\theta(y) = \theta_c(\neq 0)$. Then λ_1 is given by

$$\lambda_1 = \frac{J'(v^*)q_c}{m^2\theta_c}.$$

Case II. $q_c = 0$ and $\lambda_1 = 0$.

Concerning Case I, we have the following a priori estimate for λ_1.

LEMMA 5.1 (a priori estimate for λ_1). *Assume that the principal eigenvalue λ^ϵ and the associated eigenfunctions have the following asymptotic forms:*

$$\lambda^\epsilon = \epsilon\lambda_1(\epsilon) \quad \textit{and} \quad \lim_{\epsilon\downarrow 0}\lambda_1(\epsilon) = \frac{J'(v^*)q_c}{m^2\theta_c},$$

$$w \approx \epsilon h_v^\pm(v^*)q_c + (w^0(r/\epsilon, y) + \epsilon w^{\pm,1}(r/\epsilon, y)) \cdot \omega(r),$$

$$z \approx q_c + \epsilon Z^{\pm,2}(x) + \epsilon^2(z^{\pm,1}(r/\epsilon, y) + \epsilon z^{\pm,2}(r/\epsilon, y)) \cdot \omega(r)$$

on $\Omega_\pm(\Gamma^*)$. Then the real part of $\lambda_1(\epsilon)$ is strictly negative.

Proof. See Appendix G. □

In view of Lemma 5.1, we expand λ^ϵ as $\lambda^\epsilon = \epsilon^2\lambda_2 + o(\epsilon^2)$. For the eigenvalues of (5.4), we have the following result.

COROLLARY OF THEOREM 1.3. *Assume that* (A1)–(A3) *in section* 1, (A4′)–(A6′) *with $\lambda_1 = 0$ and* (A7) *are satisfied. Then the principal part of the critical eigenvalues is given by $\epsilon^2\lambda^*$, where λ^* is the eigenvalue of the following problem:*

$$L\theta^* = \lambda^*\theta^* \tag{5.5}$$

for $\theta^* \in (I - P)C^{2,\alpha}(\Gamma^*)$.

We assume, for simplicity, that the nonlinearity takes the form $(f, g) = (u - u^3, u - \gamma v)$, $\gamma > 3/2$. Then the constants are precisely computed as

$$v^* = 0, \quad h^\pm(v^*) = \pm 1, \quad J'(v^*) = -2, \quad m^2 = \frac{2\sqrt{2}}{3} = \frac{4}{3\sqrt{2}}.$$

For example, when Ω is a rectangle in (x, y)-plane $\Omega = (0, X) \times (0, Y)$, the reduced ploblems (5.1)–(5.3) are recast as

$$(5.6)_+ \qquad \begin{cases} DV_{xx}^+ + G^+ = 0, & x_0 < x < X, \\ V^+(x_0) = 0, \quad V_x^+(X) = 0, \end{cases}$$

$$(5.6)_- \qquad \begin{cases} DV_{xx}^- + G^- = 0, & 0 < x < x_0, \\ V_x^-(0) = 0, \quad V^-(x_0) = 0, \end{cases}$$

$$V_x^-(x_0) = V_x^+(x_0), \tag{5.7}$$

where $G^\pm := g(h^\pm(v^*), v^*) = \pm 1$, $\Omega_+(\Gamma^*) = \{(x,y) \in \mathbf{R}^2 \,|\, x_0 < x < X,\ 0 < y < Y\}$, $\Omega_-(\Gamma^*) = \{(x,y) \in \mathbf{R}^2 \,|\, 0 < x < x_0, 0 < y < Y\}$, and $\Gamma^* = \{(x_0, y) \in \mathbf{R}^2 \,|\, 0 < y < Y\}$. Here we used the fact that $H^*(y) = 0$. We can easily show that $(5.6)_-$ and $(5.6)_+$ have unique solutions given by

$$V^-(x) = -\frac{1}{2D} G^-[x^2 - x_0^2],$$
$$V^+(x) = -\frac{G^+}{D}\left[\frac{1}{2}(x^2 - x_0^2) - X(x - x_0)\right].$$

Then, by using the C^1-matching condition (5.7), we can uniquely determine x_0 and $V_x(x_0)$ as

$$x_0 = \frac{G^+ X}{[g]} \quad \text{and} \quad V_x(x_0) = -\frac{G^+ G^- X}{D[g]},$$

where $[g] = G^+ - G^-$. The existence results of the planar solution to (1.4) is given by Taniguchi and Nishiura [23] (see Theorem 4.6 in section 4).

The jth eigenvalue Λ_j of $\mathcal{T}$ is given by $\Lambda_j = X\Lambda(\kappa\pi j)$, where

$$\Lambda(z) = \frac{\cosh z + \cosh pz}{2z \sinh z}, \quad p = -\frac{G^-}{[g]}, \quad \kappa = \frac{X}{Y}.$$

Then the associated eigenfunction is $\beta_j(y) = \cos(\tau_j y)$, where $\tau_j = \pi j / Y$. Therefore, the jth eigenvalue of L is real and given by $\Sigma_j = \Sigma(\kappa\pi j, X, D)$, where

$$\Sigma(z, X, D) := -\left(\frac{z}{X}\right)^2 - \frac{J'(v^*)X}{Dm^2}\left[-\frac{G^+G^-}{[g]} + [g]\Lambda(z)\right].$$

The associated eigenfunction is the same as that of $\mathcal{T}$. $\Sigma(z, X, D)$ have the following properties.

LEMMA 5.2. *The nullcline of $\Sigma(z, X, D)$ as a function of $z > 0$ and $X > 0$ is given by $\{(z, X(z, D)) \mid z > z_0\}$, where z_0 is a unique zero of $-\frac{G^+G^-}{[g]} + [g]\Lambda(z) = 0$ and*

$$X(z, D) = \left(\frac{Dm^2}{-J'(v^*)}\right)^{1/3}\left[-\frac{G^+G^-}{[g]} + [g]\Lambda(z)\right]^{-1/3} z^{2/3}.$$

Moreover, $X(z, D)$ has the following properties:

$$\lim_{z \to z_0 + 0} X(z, D) = \infty,$$

$$\lim_{z \to \infty}\left[X(z, D) - \left(\frac{Dm^2}{-J'(v^*)}\right)^{1/3}\left(-\frac{G^+G^-}{[g]}\right)^{-1/3} z^{2/3}\right] = 0,$$

$$\frac{dX}{dz} \quad \begin{cases} < 0 & \textit{for} \quad z_0 < z < z_1, \\ = 0 & \textit{for} \quad z = z_1, \\ > 0 & \textit{for} \quad z_1 < z, \end{cases}$$

for some $z_1 = z_1(D)$.

By using Lemma 5.2, we obtain the following corollary.

COROLLARY OF THEOREM 1.5. (i) *For any $D > 0$, there exists $\underline{X} = \underline{X}(D) > 0$ such that the planar solution is stable for $X < \underline{X}$ and $\kappa > 0$.*

(ii) *For any fixed $\kappa > 0$ and $D > 0$, there exists $\overline{X} = \overline{X}(\kappa, D) > 0$ such that the planar solution is unstable for $X > \overline{X}$.*

Note. After finishing our paper, the anonymous referee noted us an interesting preprint, "On the spectra of 3-D lamellar solutions of the diblock copolymer problem," by X. Ren and J. Wei, which now appears in [27]. Their results are consistent with the part of our stability results for the lamellar patterns, although they employed a different Euler–Lagrange equation and methods for stability analysis.

Appendix A. (proof of Proposition 2.2). First we compute $u_\xi^{\pm,2}(0,y)$.

$$
\begin{aligned}
u_\xi^{\pm,2}(0,y) = &- U^{\pm,2}(0,y)\frac{u^*_{\xi\xi}(s_0)}{u^*_\xi(s_0)} \\
&- \frac{1}{u^*_\xi(s_0)}\int_{\pm\infty}^0 \left[p^\pm_{2,1}+p^\pm_{2,2}+p^\pm_{2,3}+p^\pm_{2,4}+p^\pm_{2,5}\right] u^*_\xi(\xi+s_0)d\xi,
\end{aligned} \tag{A.1}
$$

where

$$
\begin{aligned}
p^\pm_{2,1} &= (N-1)H(0,y)u^1_\xi, \\
p^\pm_{2,2} &= \Delta(0)u^0 + (N-1)H_r(0,y)\xi u^0_\xi, \\
p^\pm_{2,3} &= \frac{1}{2}\tilde{f}_{uu}[U^1(0,y)+u^1]^2, \\
p^\pm_{2,4} &= \tilde{f}_u[\xi U^1_r(0,y)+U^2(0,y)], \\
p^\pm_{2,5} &= -[\xi V^1_r(0,y)+V^2(0,y)].
\end{aligned}
$$

In the following, we compute each term of the integral part in (A.1):

$$
\begin{aligned}
&\int_{\pm\infty}^0 p_{2,2}(\xi,y)u^*_\xi(\xi+s_0)d\xi = \Delta(0)s_0\int_{\pm\infty}^{s_0}[u^*_\xi(\xi)]^2d\xi + (N-1)H_r(0,y) \\
&\quad\int_{\pm\infty}^{s_0}(\xi-s_0)[u^*_\xi(\xi)]^2d\xi\int_{\pm\infty}^0 p_{2,3}(\xi,y)u^*_\xi(\xi+s_0)d\xi \\
&\quad= -\frac{1}{2}f^*_u[U^1(0,y)]^2 - \int_{\pm\infty}^0 f_u(u^*(\xi+s_0))[U^1(0,y)+u^1(\xi,y)]u^1_\xi(\xi,y)d\xi \\
&\quad= -\frac{1}{2}f^*_u[U^1(0,y)]^2 + \int_{\pm\infty}^0 [u^1_{\xi\xi}(\xi,y)+(N-1)H(0,y)u^*_\xi(\xi+s_0)]u^1_\xi(\xi,y)d\xi \\
&\quad= \frac{1}{2}U^1(0,y)b^*_1(y) + \frac{1}{2}[u^{1,\pm}_\xi(0,y)]^2 + (N-1)H(0,y)\int_{\pm\infty}^0 u^1_\xi(\xi,y)u^*_\xi(\xi+s_0)d\xi,
\end{aligned}
$$

where $f^*_u = \frac{d}{du}f(h^\pm(v^*))$ and we used the fact that $f^*_uU^1(y,0) - b^*_1(y) = 0$.

$$
\begin{aligned}
\int_{\pm\infty}^0 p_{2,4}(\xi,y)u^*_\xi(\xi+s_0)d\xi &= f(u^*(s_0))U^2(0,y) - U^1_r(0,y)\int_{\pm\infty}^0 f(u^*(\xi+s_0))d\xi \\
&= f(u^*(s_0))U^2(0,y) + U^1_r(0,y)u^*_\xi(s_0) \\
\int_{\pm\infty}^0 p_{2,5}(\xi,y)u^*_\xi(\xi+s_0)d\xi &= -V^1_r(0,y)\int_{\pm\infty}^{s_0}(\xi-s_0)u^*_\xi(\xi)d\xi - b_2(y)[u^*(s_0)-h^\pm(v^*)].
\end{aligned}
$$

Concerning $p_{2,1}$, we have

$$\begin{aligned}
&\int_{\pm\infty}^{0} u_\xi^1(\xi,y)u_\xi^*(\xi+s_0)d\xi = -\frac{U^1(0,y)}{u_\xi^*(s_0)}\int_{\pm\infty}^{0} u_{\xi\xi}^*(\xi+s_0)u_\xi^*(\xi+s_0)d\xi\\
&\quad -\int_{\pm\infty}^{0} u_{\xi\xi}^*(\xi+s_0)u_\xi^*(\xi+s_0)d\xi\int_0^\xi [u_\xi^*(\tau+s_0)]^{-2}\int_{\pm\infty}^{\tau}[***]u_\xi^*(s+s_0)dsd\tau d\xi\\
&\quad -\int_{\pm\infty}^{0}\int_{\pm\infty}^{\xi}[***]u_\xi^*(s+s_0)d\tau d\xi\\
&\qquad = \underline{\frac{1}{2}U^{1,\pm}(0,y)u_\xi^*(s_0)}_A - \frac{1}{2}\int_{\pm\infty}^{0}\int_{\pm\infty}^{\xi}[(N-1)H(0,y)u_\xi^*(\tau+s_0)\\
&\qquad\quad \underline{+f_u(u^*(\tau+s_0))U^1(0,y)}_A - b_1^*(y)]u_\xi^*(\tau+s_0)d\tau d\xi\\
&\qquad = -\frac{1}{2}\int_{\pm\infty}^{0} 1\int_{\pm\infty}^{\xi}[(N-1)H(0,y)u_\xi^*(\tau+s_0)-b_1^*(y)]u_\xi^*(\tau+s_0)d\tau d\xi\\
&\qquad = \frac{1}{2}\int_{\pm\infty}^{s_0}(\xi-s_0)[(N-1)H(0,y)u_\xi^*(\xi)-b_1^*(y)]u_\xi^*(\xi)d\xi.
\end{aligned}$$

The underlined terms cancel pairwise. By using the above expression,

$$\begin{aligned}
&\int_{-\infty}^{0} u_\xi^{-,1}(\xi,y)u_\xi^*(\xi+s_0)d\xi - \int_{\infty}^{0} u_\xi^{+,1}(\xi,y)u_\xi^*(\xi+s_0)d\xi\\
&\qquad = \frac{1}{2}\int_{-\infty}^{\infty}\xi[(N-1)H(0,y)u_\xi^*(\xi)-b_1^*(y)]u_\xi^*(\xi)d\xi.
\end{aligned}$$

Here we used the fact that $b_1^*(y) = -m^2(N-1)H(0,y)/J'(v^*)$.

Combining the above computations, we have

$$\begin{aligned}
&-u_\xi^*(s_0)[u_\xi^{\pm,2}(0,y)+U_r^{\pm,1}(0,y)]\\
&\quad = 2(N-1)H(0,y)\int_{\pm\infty}^{0} u_\xi^{\pm,1}(\xi,y)u_\xi^*(\xi+s_0)d\xi\\
&\qquad + [\Delta(0)s_0-(N-1)H_r(0,y)s_0]\int_{\pm\infty}^{s_0}[u_\xi^*(\xi)]^2d\xi + (N-1)H_r(0,y)\int_{\pm\infty}^{s_0}\xi[u_\xi^*(\xi)]^2d\xi\\
&\qquad + \frac{1}{2}U^{\pm,1}(0,y)b_1^*(y) + \frac{1}{2}[u_\xi^{\pm,1}(0,y)]^2 + s_0V_r^1(0,y)\int_{\pm\infty}^{s_0}u_\xi^*(\xi)d\xi\\
&\qquad - V_r^1(0,y)\int_{\pm\infty}^{s_0}\xi u_\xi^*(\xi)d\xi - b_2(y)\int_{\pm\infty}^{s_0}u_\xi^*(\xi)d\xi,
\end{aligned}$$

$$\begin{aligned}
&u_\xi^*(s_0)[u_\xi^{+,2}(0,y)+U_r^{+,1}(0,y)-u_\xi^{-,2}(0,y)-U_r^{-,1}(0,y)]\\
&= (N-1)\int_{-\infty}^{\infty}\xi[(N-1)H(0,y)u_\xi^*(\xi)-b_1^*(y)]u_\xi^*(\xi)d\xi\\
&+ [\Delta(0)s_0-(N-1)H_r(0,y)s_0]\int_{-\infty}^{\infty}[u_\xi^*(\xi)]^2d\xi + (N-1)H_r(0,y)\int_{-\infty}^{\infty}\xi[u_\xi^*(\xi)]^2d\xi\\
&- \frac{1}{2}b_1^*(y)\left[U^{+,1}(0,y)-U^{-,1}(0,y)\right] + s_0V_r^1(0,y)[h] - V_r^1(0,y)\int_{-\infty}^{\infty}\xi u_\xi^*(\xi)d\xi - b_2(y)[h].
\end{aligned}$$

Substituting the above results into $\frac{\partial}{\partial r}\mathcal{U}^{+,\epsilon}(0,y) - \frac{\partial}{\partial r}\mathcal{U}^{-,\epsilon}(0,y)$, we have (2.9).

Concerning (2.8), we note only the following:

$$\begin{aligned} v_\xi^{\pm,3}(\xi,y) &= -\int_{\pm\infty}^{s_0}[u^*(\tau)-h^\pm(v^*)]d\tau \\ &= -\int_{\pm\infty}^{0}[u^*(\tau)-h^\pm(v^*)]d\tau - \int_0^{s_0} u^*(\tau)d\tau + h^\pm(v^*)s_0, \\ v_\xi^{-,3}(\xi,y) - v_\xi^{+,3}(\xi,y) &= -\int_{-\infty}^{0}[u^*(\tau)-h^-(v^*)]d\tau + \int_{\infty}^{0}[u^*(\tau)-h^+(v^*)]d\tau - [h]s_0. \end{aligned}$$

Appendix B. (proof of Lemma 3.1). We can assume without loss of generality that the $O(1)$-terms W^0, Z^0, w^0 and z^0 are not equivalent to zero at the same time. We note that when $\lambda^\epsilon = \lambda_0 + \epsilon\lambda_1$, the expansions for w and C^1-matching conditions are the same as in subsections 3.1, 3.2, and 3.3. Also, we easily see that $z^0 \equiv 0$ and $z^1 \equiv 0$ (see (3.9) and (3.10)) and the equation for $z^{\pm,2}$ becomes

$$\begin{cases} z_{\xi\xi}^{\pm,2} + \lambda_0 w^{\pm,0} = 0, \\ z^{\pm,2}(0,y) = 0, \quad \lim_{\xi\to\pm\infty} z^{\pm,2}(\xi,y) = 0. \end{cases}$$

If $\lambda^\epsilon = \lambda_0 + \epsilon\lambda_1$, the equations of W^0 and Z^0 become

$$\begin{cases} f_u^0 W^0 - Z^0 = 0, \\ \Delta Z^0 + \lambda_0 W^0 = 0, \end{cases} \qquad \text{in } \Omega_\pm(\Gamma^*),$$

and these equations are rewritten as

$$\Delta Z^0 + \frac{1}{f_u^0}\lambda_0 Z^0 = 0,$$

since $f_u^0 < 0$. On the other hand, we see $Z^0(0,y) = q_0(y) \equiv 0$ from Lemma 3.4. So we conclude (i) $Z^0 \equiv 0$ (hence $W^0 \equiv 0$) and Re $\lambda_0 \geq 0$, or (ii) $Z^0 \neq 0$ and Re $\lambda_0 < 0$. If $Z^0 \equiv 0$, the equation of Z^1 becomes

$$\Delta Z^1 + \frac{1}{f_u^0}\lambda_0 Z^1 = 0.$$

Then the $O(\epsilon)$-term of C^1-matching conditions (3.19) become

$$u_\xi^*(s_0(y))[W_r^{-,0}(0,y) + w_\xi^{-,1}(0,y) - W_r^{+,0}(0,y) - w_\xi^{+,1}(0,y)] = q_1(y)[h] = 0, \tag{B.1}$$

$$\begin{aligned} &Z_r^{-,1}(0,y) + z_\xi^{-,2}(0,y) - Z_r^{+,1}(0,y) - z_\xi^{+,2}(0,y) \\ &\qquad = (\Pi_- + \Pi_+)q_1 - \lambda_0[h]\theta(y) = 0. \end{aligned} \tag{B.2}$$

From (B.1) and (B.2), we have $Z^1(0,y) = q_1(y) \equiv 0$, and (a) $\lambda_0 = 0$ or (b) $\lambda_0 \neq 0$ and $\theta(y) \equiv 0$. In the case (b), we obtain that $w^0 \equiv 0$ since w^0 is given by

$$w^0(\xi,y) = \Theta(y)\frac{u_\xi^*(\xi+s_0)}{u_\xi^*(s_0)} = \theta(y)u_\xi^*(\xi+s_0)$$

(see (3.14)). This contradicts our assumption. Thus we conclude that Re $\lambda_0 < 0$ or $\lambda_0 = 0$.

Appendix C. (proof of Lemma 3.5). $\tilde{\Phi}(\theta, q_2, \lambda_1, 0) = 0$ is equivalent to

$$u_\xi^*(s_0(y))[w_\xi^{-,2}(0,y) - w_\xi^{+,2}(0,y)] = 0.$$

In order to calculate $w_\xi^{\pm,2}(0,y)$, we display the equation and the boundary conditions of $w_\xi^{\pm,2}$ again

$$w_{\xi\xi}^2 + (N-1)H(0,y)w_\xi^1 + [(N-1)H_r(0,y)\xi\partial_\xi + \Delta^\Gamma]w^0$$
$$+\tilde{f}_u^0[w^2 + W^2(0,y)] + \tilde{f}_u^1 w^1 + \tilde{f}_u^2 w^0 - Z^2(0,y) = 0,$$
$$w^{\pm,2}(0,y) = -W^{\pm,2}(0,y), \quad \lim_{\xi\to\pm\infty} w^{\pm,2}(\xi,y) = 0,$$

where

$$\tilde{f}_u^{\pm,2} = \tilde{f}_{uu}[p_{2,1}] + \frac{1}{2}\tilde{f}_{uuu}[p_{2,3}]^2$$
$$p_{2,1} := \xi U_r^1(0,y) + U^2(0,y) + u^2, \quad p_{2,3} := U^1(0,y) + u^1.$$

Using the expression (3.17) for $w^{\pm,2}(\xi,y)$, we have

$$w_\xi^2(0,y) = -\frac{1}{u_\xi^*(s_0)}\int_{\pm\infty}^0 [(N-1)H(0,y)\underbrace{w_\xi^1}_{\text{(i)}} + (N-1)H_r(0,y)\underbrace{\xi\partial_\xi w^0}_{\text{(ii)}}$$
$$+\underbrace{\Delta^{\Gamma^*}w^0}_{\text{(iii)}} + \underbrace{\tilde{f}_u^1 w^1}_{\text{(iv)}} + \underbrace{\tilde{f}_u^2 w^0}_{\text{(v)}} - Z^2(0,y)]u_\xi^*(s+s_0)d\xi. \tag{C.1}$$

In the following, we calculate each term of integral in (C.1):

Computation of (i). By using

$$w^{\pm,1}(\xi,y) = -\theta(y)u_\xi^*(\xi+s_0)\int_0^\xi [u_\xi^*(t+s_0)]^{-2}$$
$$\times\int_{\pm\infty}^t [(N-1)H(0,y)u_{\xi\xi}^*(s+s_0) + \tilde{f}_u^1 u_\xi^*(s+s_0)]u_\xi^*(s+s_0)dsdt,$$

we have

$$\begin{aligned}\int_{\pm\infty}^0 w_\xi^1(\xi,y)u_\xi^*(\xi+s_0)d\xi &= -\int_{\pm\infty}^0 w^1(\xi,y)u_{\xi\xi}^*(\xi+s_0)d\xi \\ &= -\frac{1}{2}\theta(y)\int_{\pm\infty}^0\int_{\pm\infty}^\xi [(N-1)H(0,y)u_{\xi\xi}^*(s+s_0) \\ &\quad +\tilde{f}_u^1 u_\xi^*(s+s_0)]u_\xi^*(s+s_0)dsd\xi \\ &= \frac{1}{2}\theta(y)[u_\xi^{\pm,1}(0,y)u_\xi^*(s_0) - 2I_3^\pm].\end{aligned}$$

Here we used the fact that

$$-\int_{\pm\infty}^0\int_{\pm\infty}^\xi [(N-1)H(0,y)u_{\xi\xi}^*(s+s_0) + \tilde{f}_u^1 u_\xi^*(s+s_0)]u_\xi^*(s+s_0)dsd\xi$$
$$= u_\xi^{\pm,1}(0,y)u_\xi^*(s_0) - 2I_3^\pm, \tag{C.2}$$

where

$$I_3^{\pm} := \int_{\pm\infty}^{0} u_\xi^1(\xi, y) u_{\xi\xi}^*(\xi + s_0) d\xi.$$

Equation (C.2) is proved in Appendix D.

Computation of (ii).

$$\begin{aligned}\int_{\pm\infty}^{0} \xi w_\xi^0(\xi, y) u_\xi^*(\xi + s_0) d\xi &= \theta(y) \int_{\pm\infty}^{0} \xi u_{\xi\xi}^*(\xi + s_0) u_\xi^*(\xi + s_0) d\xi \\ &= -\frac{1}{2}\theta(y) \int_{\pm\infty}^{0} [u_\xi^*(\xi + s_0)]^2 d\xi.\end{aligned}$$

Computation of (iii).

$$\begin{aligned}\Delta^{\Gamma^*} w^0 &= \Delta^{\Gamma^*}(\theta u_\xi^*) \\ &= u_\xi^* \Delta^{\Gamma^*}\theta + 2u_{\xi\xi}^* \nabla^\Gamma \theta \cdot \nabla s_0 + u_{\xi\xi}^* \theta \Delta^{\Gamma^*} s_0 + u_{\xi\xi\xi}^* \theta |\nabla^\Gamma s_0|^2.\end{aligned}$$

Here note that u_ξ^*, $u_{\xi\xi}^*$, etc., are all evaluated at $\xi + s_0$. Then multiplying $\Delta^{\Gamma^*} w^0$ by u_ξ^*, we have

$$\begin{aligned}\int_{\pm\infty}^{0} (\Delta^{\Gamma^*} w^0) u_\xi^* d\xi = \Delta^{\Gamma^*}\theta \int_{\pm\infty}^{0} [u_\xi^*(\xi + s_0)]^2 d\xi + [u_\xi^*(s_0)]^2 \left[\nabla^\Gamma \theta \cdot \nabla s_0 + \frac{1}{2}\theta \Delta^{\Gamma^*} s_0\right] \\ + \left[u_{\xi\xi}^*(s_0) u_\xi^*(s_0) - \int_{\pm\infty}^{0} [u_{\xi\xi}^*(\xi + s_0)]^2 d\xi\right] \theta |\nabla^\Gamma s_0|^2.\end{aligned}$$

Computation of (iv).

$$\begin{aligned}&\int_{\pm\infty}^{0} \tilde{f}_u^1 u_\xi^*(\xi + s_0) w^1(\xi, y) ds d\xi \\ &= -\int_{\pm\infty}^{0} [(N-1)H(0,y) u_{\xi\xi}^*(\xi + s_0) + u_{\xi\xi\xi}^1(\xi, y) + \tilde{f}_u u_\xi^1(\xi, y)] w^1(\xi, y) d\xi \\ &\quad \times (u_{\xi\xi\xi}^1 + \tilde{f}_u u_\xi^1 + (N-1)H(0,y) u_{\xi\xi}^* + \tilde{f}_u^1 u_\xi^* = 0 \ \text{ is used}) \\ &= (N-1)H(0,y) \int_{\pm\infty}^{0} u_\xi^*(\xi + s_0) w_\xi^1(\xi, y) d\xi + u_\xi^{\pm,1}(0,y) w_\xi^{\pm,1}(0,y) \\ &\quad - \int_{\pm\infty}^{0} [w_{\xi\xi}^{\pm,1}(\xi, y) + \tilde{f}_u w^{\pm,1}(\xi, y)] u_\xi^{\pm,1} d\xi \\ &\quad \times (w_{\xi\xi}^1 + \tilde{f}_u w^1 + (N-1)H(0,y) w_\xi^0 + \tilde{f}_u^1 w^0 = 0 \ \text{ and } \ w^0 = \theta u_\xi^* \ \text{ will be used}) \\ &= (N-1)H(0,y) \int_{\pm\infty}^{0} u_\xi^*(\xi + s_0) w_\xi^1(\xi, y) d\xi + u_\xi^{\pm,1}(0,y) w_\xi^{\pm,1}(0,y) \\ &\quad + \theta(y) \int_{\pm\infty}^{0} [(N-1)H(0,y) u_{\xi\xi}^*(\xi + s_0) + \tilde{f}_u^1 u_\xi^*(\xi + s_0)] u_\xi^{\pm,1}(\xi, y) d\xi \\ &= (N-1)H(0,y) \int_{\pm\infty}^{0} u_\xi^*(\xi + s_0) w_\xi^1(\xi, y) d\xi + u_\xi^{\pm,1}(0,y) w_\xi^{\pm,1}(0,y) \\ &\quad + \theta(y)[(N-1)H(0,y) I_3^{\pm} + I_4^{\pm}],\end{aligned}$$

where

$$I_4^{\pm} := \int_{\pm\infty}^{0} \tilde{f}_u^1 u_\xi^*(\xi+s_0) u_\xi^{\pm,1}(\xi,y)d\xi = \int_{\pm\infty}^{0} \tilde{f}_{uu}[U^1(0,y)+u^1]u_\xi^1(\xi,y)u_\xi^*(\xi+s_0)d\xi.$$

Computation of (v).

$$\int_{\pm\infty}^{0} \tilde{f}_u^{\pm,2} w^0 u_\xi^* d\xi = \theta(y)\int_{\pm\infty}^{0} \tilde{f}_u^{\pm,2}[u_\xi^*]^2 d\xi,$$

$$\int_{\pm\infty}^{0} \tilde{f}_u^{\pm,2}[u_\xi^*]^2 d\xi$$

$$= \Big[\tilde{f}_u u_\xi^* p_{2,1}\Big]_{\pm\infty}^{0} - \int_{\pm\infty}^{0} [U_r^1(0,y)+u_\xi^2]\tilde{f}_u u_\xi^* d\xi - \underbrace{\int_{\pm\infty}^{0} \tilde{f}_u p_{2,1} u_{\xi\xi}^* d\xi}_{A}$$

$$+ \left[\frac{1}{2}\tilde{f}_{uu} u_\xi^* [p_{2,3}]^2\right]_{\pm\infty}^{0} \underbrace{-\int_{\pm\infty}^{0} \tilde{f}_{uu}[U^1(0,y)+u^1]u_\xi^1 u_\xi^* d\xi}_{C} \underbrace{-\int_{\pm\infty}^{0} \frac{1}{2}\tilde{f}_{uu}[p_{2,3}]^2 u_{\xi\xi}^* d\xi}_{A}$$

$$= -\int_{\pm\infty}^{0} \left[u_{\xi\xi}^{\pm,2} + \underbrace{\tilde{f}_u p_{2,1} + \frac{1}{2}\tilde{f}_{uu}[p_{2,3}]^2}_{A}\right] u_{\xi\xi}^* d\xi$$

$$+ u_{\xi\xi}^*(s_0)[U_r^{\pm,1}(0,y) + u_\xi^{\pm,2}(0,y)] - \underbrace{\int_{\pm\infty}^{0} \tilde{f}_{uu}[U^1(0,y)+u^1]u_\xi^1 u_\xi^* d\xi}_{C}$$

$$= \int_{\pm\infty}^{0} [(N-1)H(0,y)u_\xi^1 + \Delta(0)u^0 + (N-1)H_r(0,y)\xi u_\xi^0]u_{\xi\xi}^* d\xi + V_r^1(0,y)\int_{\pm\infty}^{0} u_\xi^* d\xi$$

$$+ u_{\xi\xi}^*(s_0)[U_r^{\pm,1}(0,y) + u_\xi^{\pm,2}(0,y)] - I_4^{\pm}.$$

Here we used the equation of $u^{\pm,1}$.

$$\int_{\pm\infty}^{0} [(N-1)H(0,y)u_\xi^1 + \Delta(0)u^0 + (N-1)H_r(0,y)\xi u_\xi^0]u_{\xi\xi}^* d\xi$$

$$= (N-1)H(0,y)I_3^{\pm} + |\nabla^{\Gamma^*} s_0|^2 \int_{\pm\infty}^{0} [u_{\xi\xi}^*(\xi+s_0)]^2 d\xi + \frac{1}{2}[u_\xi^*(s_0)]^2 \Delta^\Gamma s_0$$

$$-\frac{1}{2}(N-1)H_r(0,y)\int_{\pm\infty}^{0} [u_\xi^*(\xi+s_0)]^2 d\xi.$$

Here we used the fact that $\Delta(0)u^0 = |\nabla^{\Gamma^*} s_0|^2 u_{\xi\xi}^* + \Delta^{\Gamma^*} u_\xi^*$.

Combining the above computations, we obtain

$$
\begin{aligned}
-u_\xi^*(s_0)w_\xi^{\pm,2}(0,y) &= \int_\pm^0 P_2(\xi,y)u_\xi^*(\xi+s_0)d\xi \\
&= \frac{1}{2}(N-1)H(0,y)[u_\xi^{\pm,1}(0,y)u_\xi^*(s_0)-2I_3^\pm]\theta \\
&\quad -\frac{1}{2}(N-1)H_r(0,y)[m^\pm]^2\theta \\
&\quad +[m^\pm]^2\Delta^{\Gamma^*}\theta+[u_\xi^*(s_0)]^2\left[\nabla^{\Gamma^*}\theta\cdot\nabla^{\Gamma^*}s_0+\frac{1}{2}\theta\Delta^{\Gamma^*}s_0\right] \\
&\quad +[u_{\xi\xi}^*(s_0)u_\xi^*(s_0)-[n^\pm]^2]|\nabla^{\Gamma^*}s_0|^2\theta \\
&\quad +\frac{1}{2}(N-1)H(0,y)[u_\xi^{\pm,1}(0,y)u_\xi^*(s_0)-2I_3^\pm]\theta+u_\xi^{\pm,1}(0,y)w_\xi^{\pm,1}(0,y) \\
&\quad +[(N-1)H(0,y)I_3^\pm+I_4^\pm]\theta \\
&+\left[(N-1)H(0,y)I_3^\pm+|\nabla^{\Gamma^*}s_0|^2[n^\pm]^2+\frac{1}{2}[u_\xi^*(s_0)]^2\Delta^{\Gamma^*}s_0-\frac{1}{2}(N-1)H_r(0,y)[m^\pm]^2\right. \\
&\quad \left.+u_{\xi\xi}^*(s_0)[U_r^{\pm,1}(0,y)+u_\xi^{\pm,2}(0,y)]-I_4^\pm+V_r^1(0,y)\int_{\pm\infty}^0 u_\xi^*(\xi+s_0)d\xi\right]\theta,
\end{aligned}
$$

$$
\begin{aligned}
&-Z^2(0,y)\int_{\pm\infty}^0 u_\xi^*(\xi+s_0)d\xi \\
&\quad =[m^\pm]^2\Delta^{\Gamma^*}\theta-(N-1)H_r(0,y)[m^\pm]^2\theta+V_r^1(0,y)\theta\int_{\pm\infty}^0 u_\xi^*(\xi+s_0)d\xi \\
&\quad\quad -Z^2(0,y)\int_{\pm\infty}^0 u_\xi^*(\xi+s_0)d\xi+(N-1)H(0,y)u_\xi^{\pm,1}(0,y)u_\xi^*(s_0)\theta \\
&\quad\quad +[u_\xi^*(s_0)]^2[\nabla^{\Gamma^*}\theta\cdot\nabla^{\Gamma^*}s_0+\theta\Delta^{\Gamma^*}s_0]+u_{\xi\xi}^*(s_0)u_\xi^*(s_0)|\nabla^{\Gamma^*}s_0|^2\theta \\
&\quad\quad +u_\xi^{\pm,1}(0,y)w_\xi^{\pm,1}(0,y)+u_{\xi\xi}^*(s_0)[U_r^{\pm,1}(0,y)+u_\xi^{\pm,2}(0,y)],
\end{aligned}
$$

where

$$
m^\pm := \left[\int_{\pm\infty}^0 [u_\xi^*(\xi+s_0)]^2d\xi\right]^{1/2} \quad \text{and} \quad n^\pm := \left[\int_{\pm\infty}^0 [u_{\xi\xi}^*(\xi+s_0)]^2d\xi\right]^{1/2}.
$$

Finally, by using the relations

$$
\begin{aligned}
&u_\xi^{+,1}(0,y)=u_\xi^{-,1}(0,y), \quad w_\xi^{+,1}(0,y)=w_\xi^{-,1}(0,y), \\
&U_r^{+,1}(0,y)+u_\xi^{+,1}(0,y)=U_r^{-,1}(0,y)+u_\xi^{-,1}(0,y),
\end{aligned}
$$

we obtain

$$-u^*_\xi(s_0)[w^{-,2}_\xi(0,y)-w^{+,2}_\xi(0,y)]d = m^2\Delta^{\Gamma^*}\theta + m^2H^*(y)\theta \\ -V^1_r(0,y)J'(v^*)\theta + J'(v^*)Z^2(0,y),$$

where

$$H^*(y) := -(N-1)H_r(0,y) = \sum_{j=1}^{N-1}[\kappa_j]^2,$$

κ_j $(j=1,\dots,N-1)$, are the principal curvatures of Γ^*. Thus we obtain (3.26).

Next we prove (3.27). By using the fact that $W^{\pm,0}(x)\equiv 0\equiv Z^{\pm,0}(x)$ and $W^{\pm,1}(x)\equiv 0\equiv Z^{\pm,1}(x)$, we see that $Z^{\pm,2}$ satisfies the following equation:

$$\begin{cases} \Delta Z^{\pm,2}=0 \quad \text{in } \ \Omega_\pm(\Gamma^*), \\ \\ Z^{\pm,2}=q_2 \quad \text{on } \ \Gamma^*, \ \ \dfrac{\partial Z^{-,2}}{\partial n}=0 \quad \text{on } \ \partial\Omega. \end{cases}$$

Then we have

$$Z^{-,2}_r(0,y)-Z^{+,2}_r(0,y) = \frac{\partial}{\partial\nu}(\mathcal{P}^-q_2)\Big|_{\Gamma^*} + \frac{\partial}{\partial\nu}(\mathcal{P}^+q_2)\Big|_{\Gamma^*} = (\Pi_-+\Pi_+)q_2. \tag{C.3}$$

In view of (3.15) and (3.22), we have

$$z^{-,3}_\xi(0,y)-z^{+,3}_\xi(0,y) = (1+\lambda_1)[h]\theta(y). \tag{C.4}$$

Combining (C.3) and (C.4), we obtain

$$Z^{-,2}_r(0,y)-Z^{+,2}_r(0,y)+z^{-,3}_\xi(0,y)-z^{+,3}_\xi(0,y) = (\Pi_-+\Pi_+)q_2-(1+\lambda_1)[h]\theta(y).$$

Appendix D. (proof of (3.25) and (C.2)). By using

$$u^1_{\xi\xi\xi}+\tilde f_u u^1_\xi + (N-1)H(0,y)u^*_{\xi\xi}+\tilde f^1_u u^*_\xi = 0 \ \text{ and } \ u^*_{\xi\xi\xi}+\tilde f_u u^*_\xi = 0,$$

we have

$$\int_{\pm\infty}^0 \left[(N-1)H(0,y)u^*_{\xi\xi}(\xi+s_0)+\tilde f^1_u u^*_\xi(\xi+s_0)\right]u^*_\xi(\xi+s_0)d\xi \\ = u^{\pm,1}_\xi(0,y)u^*_{\xi\xi}(s_0)-u^*_\xi(s_0)u^{\pm,1}_{\xi\xi}(0,y)$$

and

$$-\int_{\pm\infty}^0\int_{\pm\infty}^\xi [(N-1)H(0,y)u^*_{\xi\xi}(s+s_0)+\tilde f^1_u u^*_\xi(s+s_0)]u^*_\xi(s+s_0)dsd\xi$$

$$= \int_{\pm\infty}^0 [u^1_{\xi\xi}(\xi,y)u^*_\xi(\xi+s_0)-u^1_\xi(\xi,y)u^*_{\xi\xi}(\xi+s_0)]d\xi$$

$$= u^{\pm,1}_\xi(0,y)u^*_\xi(s_0)-2I^\pm_3.$$

Appendix E. (proof of Lemma 4.3). (i) We study the operator $\Pi_- + \Pi_+$ instead of $\mathcal{T}$. The problems (1.11) and (1.10) are rewritten as

$$(\mathrm{E}.1)_+ \qquad \begin{cases} Z^+_{rr} + \dfrac{N-1}{r} Z^+_r + \dfrac{1}{r^2}\Delta^S Z^+ = 0, & r_0 < r < R, \\ Z^+(r_0, y) = q, \quad Z^+_r(R, y) = 0, & y \in S, \end{cases}$$

$$(\mathrm{E}.1)_- \qquad \begin{cases} Z^-_{rr} + \dfrac{N-1}{r} Z^-_r + \dfrac{1}{r^2}\Delta^S Z^- = 0, & 0 < r < r_0, \\ Z^-_r(0, y) = 0, \quad Z^-(r_0, y) = q, & y \in S, \end{cases}$$

where Δ^S is the Laplacian on $S = S^{N-1}$. Note that $\Pi_\pm$ have the same complete system of eigenfunctions as that of Δ^S.

Let $\{\ell_j, \beta^m_j(y)\}_{j=1}^\infty$ be the complete system of an eigenpair for $-\Delta^S$, where $\ell_j = j(j+N-2)$, $j = 1, 2, \ldots$. If we take $q = \beta_j$, then the solutions of $(E.1)_\pm$ separate as $Z^\pm(r, y) = R^{\pm,j}(r)\beta_j(y)$. The equations for $R^{\pm,j}(r)$ are, respectively,

$$\begin{cases} R^{+,j}_{rr} + \dfrac{N-1}{r} R^{+,j}_r - \dfrac{l_j}{r^2} R^{+,j} = 0, & r_0 < r < R, \\ R^{+,j}(r_0) = 1, \quad R^{+,j}_r(R) = 0, & y \in S, \end{cases}$$

$$\begin{cases} R^{-,j}_{rr} + \dfrac{N-1}{r} R^{-,j}_r - \dfrac{l_j}{r^2} R^{-,j} = 0, & 0 < r < r_0, \\ R^{-,j}_r(0) = 0, \quad R^{-,j}(r_0) = 1, & y \in S. \end{cases}$$

In terms of the solutions $R^{\pm,j}$, the operator $\Pi_- + \Pi_+$ acting on β_j are expressed as

$$(\Pi_- + \Pi_+)\beta_j = (R^{-,j}_r(r_0) - R^{+,j}_r(r_0)) \cdot \beta_j.$$

By using the fundamental solutions r^j and $r^{-(j+N-2)}$, we have

$$R^{-,j}(r) = \frac{1}{r_0^j} r^j, \quad R^{+,j}(r) = A^{+,j} r^j + B^{+,j} r^{-(j+N-2)},$$

where

$$A^{+,j} = \frac{(j+N-2)R^{-(j+N-1)}}{\Delta}, \quad B^{+,j} = \frac{jR^{j-1}}{\Delta},$$
$$\Delta = (j+N-2) r_0^j R^{-(j+N-1)} + j r_0^{-(j+N-2)} R^{j-1}.$$

Then, we have

$$R^{-,j}_r(r_0) - R^{+,j}_r(r_0) = \frac{j(2j+N-2)}{\Delta r_0^{j+N-1} R^{-(j-1)}} \tag{E.2}$$

for $j \geq 1$. Therefore, the eigenvalue of $\mathcal{T}$ is given by (4.3) and (4.4) since it is defined by the inverse of the right-hand side of (E.2).

(ii) This can be shown by simple computation so we omit the proof.

Appendix F. (proof of Lemma 4.7). (i) We study the operator $\Pi_- + \Pi_+$ instead of $\mathcal{T}$. The problems (1.11) and (1.10) are recast as

$$(\mathrm{F.1})_+ \qquad \begin{cases} Z^+_{xx} + Z^+_{yy} = 0, & (x,y) \in (x_0, X) \times (0, Y), \\ Z^+(x_0, y) = q, \quad Z^+_r(X, y) = 0, & y \in (0, Y), \end{cases}$$

$$(\mathrm{F.1})_- \qquad \begin{cases} Z^-_{xx} + Z^-_{yy} = 0, & (x,y) \in (0, x_0) \times (0, Y), \\ Z^-_r(0, y) = 0, \quad Z^-(x_0, y) = q, & y \in (0, Y). \end{cases}$$

Note that $\Pi_\pm$ have the same complete system of eigenfunctions as that of $-\frac{d^2}{dy^2}$ under the Neumann boundary condition.

Let $\{\tau_j^2, \beta_j(y)\}_{j=1}^\infty$ be the complete system of an eigenpair for $-\frac{d^2}{dy^2}$, where

$$\beta_j(y) = \cos(\tau_j y), \quad \tau_j := \frac{\pi j}{Y}.$$

If we take $q = \beta_j$, then the solutions of $(F.1)_\pm$ separate as $Z^\pm(x,y) = \zeta^{\pm,j}(x)\beta_j(y)$. The equations for $\zeta^{\pm,j}(x)$ are, respectively,

$$\begin{cases} \zeta^{+,j}_{xx} - \tau_j^2 \zeta^{+,j} = 0, & x_0 < x < X, \\ \zeta^{+,j}(x_0) = 1, & \zeta^{+,j}_x(X) = 0, \end{cases}$$

$$\begin{cases} \zeta^{-,j}_{xx} - \tau_j^2 \zeta^{-,j} = 0, & 0 < x < x_0, \\ \zeta^{-,j}_x(0) = 0, & \zeta^{-,j}(x_0) = 1. \end{cases}$$

In terms of the solutions $\zeta^{\pm,j}$, the operator $\Pi_- + \Pi_+$ acting on β_j are expressed as,

$$(\Pi_- + \Pi_+)\beta_j = (\zeta^{-,j}_x(x_0) - \zeta^{+,j}_x(x_0)) \cdot \beta_j.$$

By using the fundamental solutions $e^{\tau_j x}$ and $e^{-\tau_j x}$, we have

$$\zeta^{\pm,j}(x) = c_1^\pm e^{\tau_j x} + c_2^\pm e^{-\tau_j x},$$

where

$$\begin{bmatrix} c_1^- \\ c_2^- \end{bmatrix} = \frac{1}{\Delta^-} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} c_1^+ \\ c_2^+ \end{bmatrix} = \frac{1}{\Delta^+} \begin{bmatrix} e^{-\tau_j X} \\ e^{\tau_j X} \end{bmatrix},$$

$$\Delta^- := e^{x_0\tau_j} + e^{-x_0\tau_j}, \quad \Delta^+ := e^{(X-x_0)\tau_j} + e^{-(X-x_0)\tau_j}.$$

Then, we have

$$\zeta^{-,j}_x(x_0) - \zeta^{+,j}_x(x_0) = \frac{2\tau_j \sinh X\tau_j}{\cosh X\tau_j + \cosh \overline{u} X\tau_j} \tag{F.2}$$

for $j \geq 1$. Therefore, the eigenvalue of $\mathcal{T}$ is given by (4.9) and (4.10) since it is defined by the inverse of the right-hand side of (F.2).

(ii) This can be shown by simple computation so we omit the proof.

Appendix G. (proof of Lemma 5.1). Without loss of generality, we can normalize θ_c as $\theta_c = 1$. Multiplying $\overline{z}$ (complex conjugate of z) to the second equation of (5.4) and integrating by parts, we obtain

$$-D\int_\Omega |\nabla z|^2 dx + \int_\Omega g_u w\overline{z} dx + \epsilon\int_\Omega g_v |z|^2 dx = \epsilon^2\lambda_1(\epsilon)\int_\Omega |z|^2 dx, \tag{G.1}$$

where

$$|\nabla z|^2 := \sum_{i=1}^N \frac{\partial z}{\partial x^i}\frac{\partial \overline{z}}{\partial x^i}, \quad |z|^2 := z\overline{z}.$$

Noting that $w^{\pm,i}(\xi, y)$ (i = 0, 1) decays exponentially as $|\xi| \to \infty$, we have

$$\begin{aligned}
\int_\Omega g_u w\overline{z} dx &= \int_{\Omega_+(\Gamma^*)} g_u w^+\overline{z^+} dx + \int_{\Omega_-(\Gamma^*)} g_u w^-\overline{z^-} dx \\
&= g_u^{*,+}\int_{\Omega_+(\Gamma^*)} [\epsilon h_v^+(v^*)q_c + w^0(r,y)\omega(r)]\overline{q_c} dx \\
&\quad + g_u^{*,-}\int_{\Omega_-(\Gamma^*)} [\epsilon h_v^-(v^*)q_c + w^0(r,y)\omega(r)]\overline{q_c} dx + O(\epsilon^2) \\
&= \epsilon[g_u^{*,+}h_v^+(v^*)|\Omega_+(\Gamma^*)| + g_u^{*,-}h_v^-(v^*)|\Omega_-(\Gamma^*)|]|q_c|^2 \\
&\quad + \epsilon[C^+(\epsilon) + C^-(\epsilon)]\overline{q_c} + O(\epsilon^2),
\end{aligned}$$

where $|\Omega_\pm(\Gamma^*)|$ is volume of domain $\Omega_\pm(\Gamma^*)$,

$$C^\pm(\epsilon) := \frac{g_u^{*,\pm}}{\epsilon}\int_{\Omega_\pm(\Gamma^*)} w^0(r/\epsilon, y)\omega(r) dx = \frac{g_u^{*,\pm}}{\epsilon}\int_{\Omega_\pm(\Gamma^*)} u_\xi^*(r/\epsilon)\omega(r) dx > 0$$

for small $\epsilon > 0$ and $g_u^{*,\pm} = g_u(h^\pm(v^*), v^*)$. Using the fact that

$$\sup_{x\in\Omega} |\nabla z| = O(\epsilon) \ \text{ and } \ \int_\Omega |z|^2 dx = O(1) \quad \text{as} \ \ \epsilon \downarrow 0,$$

and dividing (G.1) by ϵ, we have

$$\begin{aligned}
(C^+(\epsilon) + C^-(\epsilon))\overline{q_c} = &-(g_u^{*,+}h_v^+(v^*)|\Omega_+(\Gamma^*)| + g_u^{*,-}h_v^-(v^*)|\Omega_-(\Gamma^*)|)|q_c|^2 \\
&+(g_u^{*,+}|\Omega_+(\Gamma^*)| + g_u^{*,-}|\Omega_-(\Gamma^*)|)|q_c|^2 + O(\epsilon).
\end{aligned}$$

This implies Re $q_c < 0$ since $g_u^{*,\pm} > 0$, $g_v^{*,\pm} = g_v(h^\pm(v^*), v^*) < 0$, and $h_v^\pm(v^*) = 1/f_u(h^\pm(v^*), v^*) < 0$. Then we conclude that

$$\text{Re } \lambda_1^* = \text{Re } \frac{J'(v^*)q_c}{m^2} < 0.$$

Acknowledgment. The authors thank the anonymous referees for many valuable comments and pointing out an important reference.

REFERENCES

[1] M. Bahiana and Y. Oono, *Cell dynamical system approach to block copolymers*, Phys. Rev., 41 (1990), pp. 6763–6771.

[2] R. Choksi, *Scaling laws in microphase separation of diblock copolymers*, J. Nonlinear Sci., 11 (2001), pp. 223–236.

[3] R. Choksi and X. Ren, *On the derivation of a density functional theory for microphase separation of diblock copolymers*, J. Statist. Phys., 113 (2003), pp. 151–176.

[4] A. Doelman and H. van der Ploeg, *Homoclinic stripe patterns*, SIAM J. Appl. Dyn. Syst., 1 (2002), pp. 65–104.

[5] M. del Pino, *Radially symmetric internal layers in a semilinear elliptic system*, Trans. Amer. Math. Soc., 347 (1995), pp. 4807–4837.

[6] H. Hasegawa, H. Tanaka, K. Yamasaki, and T. Hashimoto, Macromolecules, 20 (1987), pp. 1651.

[7] T. Hashimoto, M. Shibayama, and H. Kawai, Macromolecules, 16 (1983), pp. 1093.

[8] T. Hashimoto, H. Tanaka, and H. Hasegawa, *Molecular Conformation and Dynamics of Macromolecules in Condensed Systems*, M. Nagasawa, ed., Elsevier, Amsterdam, 1998.

[9] H. Ikeda, *On the asymptotic solutions for a weakly coupled elliptic boundary value problem with a small parameter*, Hiroshima Math. J., 16 (1986), pp. 227–250.

[10] Y. Nishiura, *Coexistence of infinitely many stable solutions to reaction diffusion systems in the singular limit*, in Dynamics Reported (New Series) Vol. 3, Springer-Verlag, Berlin, 1994, pp. 25–103.

[11] Y. Nishiura and H. Fujii, *Stability of singularly perturbed solutions to systems of reaction-diffusion equations*, SIAM J. Math. Anal., 18 (1987), pp. 1726–1770.

[12] Y. Nishiura and I. Ohnishi, *Some aspect mathematical of the micro-phase separation in diblock copolymers*, Phys. D, 84 (1995), pp. 31–39.

[13] Y. Nishiura and H. Suzuki, *Nonexistence of higher dimensional stable Turing patterns in the singular limit*, SIAM J. Math. Anal., 29 (1998), pp. 1087–1105.

[14] I. Ohnishi, Y. Nishiura, M. Imai, and Y. Matsushita, *Analytical solutions describing the phase separation driven by a free energy functional containing a long-range interaction term*, Chaos, 9 (1999), pp. 329–341.

[15] Y. Oshita, *On stable nonconstant stationary solutions and mesoscopic patterns for FitzHugh–Nagumo equations in higher dimensions*, J. Differential Equations, 188 (2003), pp. 110–134.

[16] T. Ohta and K. Kawasaki, *Equilibrium morphology of block copolymer melts*, Macromolecules, 19 (1986), pp. 2621.

[17] K. Sakamoto, *Internal layers in high-dimensional domains*, Proc. Roy. Soc. Edinb., 128 (1998), pp. 359–401.

[18] K. Sakamoto, *Spatial homogenization and internal layers in a reaction-diffusion system*, Hiroshima Math. J., 30 (2000), pp. 377–402.

[19] K. Sakamoto and H. Suzuki, *Spherically symmetric internal layers for activator-inhibitor systems: I. Existence by a Lyapunov-Schmidt reduction*, J. Differential Equations, to appear.

[20] K. Sakamoto and H. Suzuki, *Spherically symmetric internal layers for activator-inhibitor systems: II. Stability and symmetry breaking bifurcations*, J. Differential Equations, to appear.

[21] H. Suzuki, *Asymptotic characterization for interfacial patterns for reaction diffusion systems*, Hokkaido Math. J., 26 (1997), pp. 631–667.

[22] M. Taniguchi and Y. Nishiura, *Instability of planar interfaces in reaction-diffusion systems*, SIAM J. Math. Anal., 25 (1994), pp. 99–134.

[23] M. Taniguchi and Y. Nishiura, *Stability and characteristic wavelength of planar interfaces in the large diffusion limit of the inhibitor*, Proc. Roy. Soc. Edinburgh Sect. A, 126 (1996), pp. 117–145.

[24] T. Teramoto and Y. Nishiura, *Double gyroid morphology in a gradient system with nonlocal effects*, J. Phys. Soc. Japan, 71 (2002), pp. 1611–1614.

[25] X. Ren and J. Wei, *On the multiplicity of solutions of two nonlocal variational problems*, SIAM J. Math. Anal., 31 (2000), pp. 909–924.

[26] X. Ren and J. Wei, *Concentrically layered energy equilibria of the di-block copolymer problem*, European J. Appl. Math., 13 (2002), pp. 479–496.

[27] X. Ren and J. Wei, *On the spectra of three-dimensional Lamellar solutions of the Diblock copolymer problem*, SIAM J. Math. Anal., 35 (2003), pp. 1–32.

SIAM J. MATH. ANAL.
Vol. 36, No. 3, pp. 967–985

ON THE EXISTENCE OF INFINITELY MANY MODES OF A NONLOCAL NONLINEAR SCHRÖDINGER EQUATION RELATED TO DISPERSION-MANAGED SOLITONS*

MICHAEL KURTH†

Abstract. We present a comprehensive study of a nonlinear Schrödinger equation with additional quadratic potential and general, possibly highly nonlocal, cubic nonlinearity. In particular, this equation arises in a variety of applications and is known as the Gross–Pitaevskii equation in the context of Bose–Einstein condensates with parabolic traps or as a model equation describing average pulse propagation in dispersion-managed fibers. Both global and local bifurcation behavior is determined showing the existence of infinitely many symmetric modes of the equation. In particular, our theory provides a strict theoretical proof of the existence of a symmetric bi-soliton which recently was found by numerical simulations.

Key words. nonlinear Schrödinger equation, harmonic potential, dispersion management, global bifurcation theorem

AMS subject classification. 34C23, 49S05, 78M30

DOI. 10.1137/S0036141003431530

1. Introduction and main results.

In this paper we consider the nonlinear Schrödinger equation (NLS) with additional quadratic potential

$$iu_t + u_{xx} - x^2 u = F(u), \qquad x \in \mathbb{R},\ t \geq 0, \tag{1.1}$$

where $F(u)$ is a cubic, possibly nonlocal, nonlinearity satisfying some assumptions given later in this paper. Nonlocality of the nonlinearity is an important factor in many applications, often approximated by a simpler local nonlinearity. In this paper we will explain that it is possible to determine the full bifurcation behavior without the assumption that F is local.

Equation (1.3) models a variety of phenomena and is known as the Gross–Pitaevskii (GP) equation in context of Bose–Einstein condensates (BEC) with parabolic traps. Assuming a highly anisotropic trap Kivshar, Alexander, and Turitsyn [10] derived the one-dimensional GP-equation (1.1) with $F(u) = \pm|u|^2 u$ as model equation for the macroscopic dynamics of cooled atoms confined in a three-dimensional parabolic potential created by a magnetic trap. Using an approximation technique they explain the existence of infinitely many nonlinear modes of the equation. In the present paper we rigorously prove the existence of such modes identifying them as bifurcating solutions from the eigenvalues of the linear harmonic oscillator. Moreover, the shape of the mode is determined by the corresponding Gauss–Hermite eigenfunctions.

In most applications, including BEC with ultracold atomic gases, the origin of the nonlinearity is nonlocal; cf. [2, 5] and references therein. Usually the nonlocal nonlinearity is assumed to be of Hartree type, i.e.,

$$F(u) = (K * |u|^2)u,$$

*Received by the editors July 17, 2003; accepted for publication (in revised form) February 6, 2004; published electronically November 17, 2004.
http://www.siam.org/journals/sima/36-3/43153.html
†Universität zu Köln, Mathematisches Institut, Weyertal 86-90, D-50931 Cologne, Germany (mkurth@mi.uni-koeln.de).

where the kernel K is in some L^p-space and $*$ denotes evolution. The standard local GP-equation is an approximation of the nonlocal model in the sense that it arises by approximating the kernel with a δ-function. Thus, the NLS with local nonlinearity can be considered as a simplified model and is of interest to understand the original nonlocal problem, although in most applications it is hard to get a clue on properties of the kernel. It should be mentioned that the analysis given in this paper also includes the Hartree-type nonlinearity; for details we refer to [1].

Our main interest, however, lies in the context of nonlinear fiber optics. Modern optical transmission systems successfully use the so-called dispersion management (DM) technique. The idea of DM is to use a dispersion-compensating fiber to overcome the dispersion of the standard monomode fiber which causes dispersive broadening of a pulse. If the residual dispersion is small the signal should evolve nearly periodical, this situation is called strong DM. Numerical and experimental results show that the corresponding pulse is stable over hundreds of periods; analogous to the traditional NLS, it is called the DM-soliton. Using the so-called lens transformation and an averaging technique developed by Zharnitsky et al. [26] we have shown in [13] that the master equation can be transformed into (1.1) with nonlocal nonlinearity of the following form (after normalization):

$$F(u) := -\int_0^1 S^{-1}(z)\left(\frac{1}{T(z)}|S(z)u|^2 S(z)u\right)dz, \tag{1.2}$$

with $S(z) := U(R^{\mathrm{eff}}(z))$, where $U(z)$ denotes the group generated by the harmonic oscillator, $T(z)$ is a characteristic pulse width, and $R^{\mathrm{eff}}(z)$ is the effective residual dispersion (for details see section 2 below).

Regarding (1.1) we notice that the linear part is nothing other than the harmonic oscillator which has the well-known basis of Gauss–Hermite eigenfunctions. The presence of the quadratic potential helps us to overcome the problems due to the unboundedness of the underlying spatial domain such as the noncompactness or the continuous spectrum of the original problem.

In this paper we are interested in the existence of nonlinear bound states; the corresponding ansatz $u(t,x) = \exp(-\lambda t)v(x)$ results in the nonlinear eigenvalue problem

$$-u_{xx} + x^2 u + F(u) = \lambda u, \tag{1.3}$$

where we have required $F(\exp(i\theta)u) = \exp(i\theta)F(u)$ to derive the equation. The natural space to consider (1.3) is the weighted Hilbert space [9, 12]

$$X := \left\{ u \in H^1(\mathbb{R}) \,\middle|\, \int_{\mathbb{R}} x^2|u|^2 dx < \infty \right\} \tag{1.4}$$

with inner product $\langle u, v\rangle_X := (u_x, v_x) + (xu, xv)$, where $(\cdot,\cdot)$ denotes the standard inner product in $L^2(\mathbb{R})$ and corresponding energy norm

$$\|u\|_X^2 = \int_{\mathbb{R}} |u_x|^2 + x^2|u|^2 dx = \|u_x\|_2^2 + \|xu\|_2^2. \tag{1.5}$$

The main conditions on the nonlinearity are the following.

($\mathcal{F}_1$) $F : X \to L^2(\mathbb{R}) : F(\exp(i\theta)u) = \exp(i\theta)F(u)$ and $u : \mathbb{R} \to \mathbb{R} \Rightarrow F \circ u : \mathbb{R} \to \mathbb{R}$.

($\mathcal{F}_2$) There exist $0 < \alpha < 7/2, \beta \geq 4 - \alpha$ such that

$$\|F(u) - F(v)\|_2^2 \leq C(\|u\|_X^\alpha \|u\|_2^\beta + \|v\|_X^\alpha \|v\|_2^\beta)\|u - v\|_X^{1/2}\|u - v\|_2^{3/2} \ \forall u, v \in X.$$

($\mathcal{F}_3$) F is sufficiently smooth, i.e., $F \in C^1(X, L^2)$.

By virtue of assumption $(\mathcal{F}_1)$ we have

$$\langle F(u), u\rangle \in \mathbb{R}\ \forall u \in X_{\mathcal{C}} = \left\{ u \in H^1(\mathbb{R}, \mathcal{C}) \;\middle|\; \int_{\mathbb{R}} x^2 |u|^2 dx < \infty \right\}$$

and hence we consider throughout the paper only real-valued functions and consequently X instead of $X_{\mathbb{C}}$. Assumption $(\mathcal{F}_2)$ is a general growth condition for a cubic nonlinearity which appears quite natural, resulting in $F(u) = \mathcal{O}(\|u\|_2^3)$ for $u \to 0$. $(\mathcal{F}_3)$ is a minimum smoothness condition which will later be replaced by some stronger condition in order to determine the local bifurcation behavior. Necessary for the validity of the variational approach is the potential property of F, that is, the following assumption:

$(\mathcal{F}_4)$ There exists $G \in \mathcal{C}^1(X, \mathbb{R})$ with $G(0) = 0$ such that
$$G'(u)v = (F(u), v) \quad \forall u, v \in X.$$

Later we will restrict ourselves to the practical relevant case, where ground states exist. In order to determine the direction of bifurcation one has to fix the sign of the nonlinearity, that is to consider only "focusing" nonlinearities with an additional technical assumption, i.e.,

$(\mathcal{F}_5)$ $G(u) < 0\ \forall u \in X \setminus \{0\}$ and $(F(su), u) \geq s^\delta (F(u), u)$ for $0 < s < 1$ and $\delta > 1$.

Moreover, we are interested in the symmetry of the solutions. Thus we consider at some stage only symmetric potentials

$(\mathcal{F}_6)$ $$G(u(x)) = G(u(-x)) \text{ and } G(u(x)) = G(-u(-x)).$$

The above assumptions allow us to determine the direction of bifurcation and orbital stability of the solution. It should be noted that in our applications $G(u)$ is only $(F(u), u)/4$ which is typically of one sign, but much more general nonlinearities can be treated as well.

Note that the nonlinearity $F(u) = \sigma|u|^2 u$ with $\sigma < 0$ satisfies assumptions $(\mathcal{F}_1)$–$(\mathcal{F}_6)$. Equation (1.3) with standard cubic nonlinearity was investigated by several authors in the past: Existence and stability of the solutions of (1.3) was discussed by Fukuizumi [4], Oh [17], and Zhang [28]. Kivshar, Alexander, and Turitsyn [10] observe the existence of infinitely many nonlinear modes of (1.3), but they did not give a theoretical explanation. The NLS with quadratic potential is also discussed in section 9.3 in the book by Cazenave [1]. Both global and local bifurcation results are obtained by Kunze et al. [12]. Moreover, the corresponding solutions decay very fast, i.e., Gaussian-like, and there exists a positive solution [6, 9].

Allowing the nonlinearity to be nonlocal, we will explain in this paper that some results and methods can be adapted, whereas other properties of the solutions are lost, e.g., the Gaussian decay. The main result of this paper can be summarized as follows: In each eigenvalue of the harmonic oscillator bifurcates an unbounded branch of nonlinear bound states in the sense of the global bifurcation theorem of Rabinowitz which give rise to the existence of infinitely many nonlinear modes. Under slightly more restrictive conditions on the nonlinearity, the bifurcating solutions can be characterized as minimizers (resp., saddle points) of the corresponding energy functional. Moreover, there exist infinitely many even (resp., odd solutions). Furthermore, stability and decay properties of the solutions are discussed. These assertions

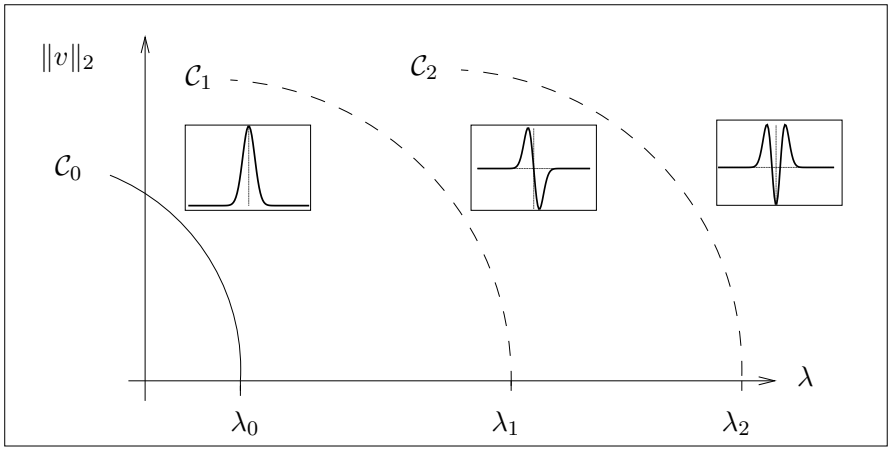

FIG. 1.1. *Bifurcation diagram of* $-v_{xx} + x^2 v + F(v) = \lambda v$.

can be visualized in a bifurcation diagram as shown in Figure 1.1, the direction of bifurcation depends on the sign of the nonlinearity.

In the context of fiber optics the ground state of the corresponding energy functional is close to the true DM-soliton in the sense of the averaging procedure, whereas the other branches correspond to modes of arbitrary order. Our method guarantees the uniqueness (up to a phase factor) of the ground state close to the bifurcation point for fixed energy and shows that it is even. This is a new theoretical result, well supported by numerical simulations. The DM-soliton as a ground state of a macroscopic quantum oscillator as (1.3) has recently been studied by Schäfer et al. [24], but they consider only reduced models and consequently our results are a verification of their approximation method.

Of great practical interest is the theoretical verification that DM-systems support the bi-soliton in addition to the well-known single-soliton. The bi-soliton was recently numerically observed by Maruta et al. [15]; see also the paper of Pare and Belanger [21]. It is a promising candidate for the improvement of today's systems and will help to increase transmission rates by using new encoding schemes.

2. Derivation of the main equation in the case of strong DM. In this section a brief motivation of the nonlinear eigenvalue problem (1.1) with nonlinearity (2.7) for the application of dispersion-managed optical fibers is given. For a more detailed derivation we refer to [13]. From a mathematical point of view the model equation describing pulse propagation in optical fibers with dispersion management is given by the cubic nonlinear Schrödinger equation (DM-NLS) with periodically varying coefficients

$$iA_z(z,t) + D(z)A_{tt}(z,t) + c|A(z,t)|^2 A(z,t) = 0. \tag{2.1}$$

Here, A is the complex envelope of the electric field, t is retarded time, z is propagation distance, D is the dispersion coefficient, and $c > 0$ represents loss and influence of the amplifiers and is assumed to be constant (loss-less model). The dispersion profile D is periodic with normalized period one. In the case of strong dispersion management the residual dispersion $\langle D \rangle$ is small compared to local dispersion, i.e., $\langle D \rangle \ll D$, where $\langle \cdot \rangle$ denotes averaging over one period. Equation (2.1) is mostly

studied for dispersion profiles having the form of a symmetric two-step map:

$$D(z) = D_{\mathrm{loc}} + \langle D\rangle = \begin{cases} d + \langle D\rangle : 0 \le z \le L, 1-L \le z \le 1, \\ -d + \langle D\rangle : L < z < 1-L \end{cases} \tag{2.2}$$

with $\langle D_{\mathrm{loc}}\rangle = 0$. Throughout the paper we restrict ourselves to the case of positive residual dispersion, i.e., we require $\langle D\rangle > 0$. In a series of papers Turitsyn and Gabitov (cf. [25]) suggest applying the following transformation to (2.1), which is known as lens transformation or pseudo-conformal transformation:

$$A(z,t) = N\frac{Q(z, t/T(z))}{\sqrt{T(z)}} \exp\left(it^2 \frac{M(z)}{T(z)}\right). \tag{2.3}$$

Here (T, M) is a periodic solution of the so-called nonlinear TM-equations which arise in the context of lens transformation (see [13, 25] for details):

$$T'(z) = 4D(z)M(z), \qquad T(0) = T_0 > 0, \tag{2.4}$$

$$M'(z) = \frac{D(z)}{T(z)^3} - \frac{N^2}{T(z)^2}, \quad M(0) = 0. \tag{2.5}$$

Thereby, T_0 has to be determined in such a way that for a given N^2 the corresponding solution is periodic or vice versa. T and M have the physical meaning of pulse width and chirp, N^2 is the pulse energy. In [13] it was shown that after applying lens transformation to the (strong) dispersion managed NLS (2.1) with a two-step map as in (2.2) and averaging of the resulting equation one arrives at a Schrödinger-type equation with additional quadratic potential, i.e.,

$$iu_z + au_{xx} - bx^2u + \int_0^1 S^{-1}(z)\left(\frac{N^2}{T(z)}|S(z)u|^2 S(z)u\right) dz = 0, \quad z \ge 0, \quad x \in \mathbb{R}, \tag{2.6}$$

where

$$a = \left\langle \frac{D}{T^2} \right\rangle - N^2 \left\langle \frac{1-\cos(4R^{\mathrm{eff}})}{2T} \right\rangle,$$
$$b = \left\langle \frac{D}{T^2} \right\rangle - N^2 \left\langle \frac{1+\cos(4R^{\mathrm{eff}})}{2T} \right\rangle.$$

In (2.6), $S(z)$ is defined as $S(z) = U(R^{\mathrm{eff}}(z))$, where R^{eff} has the physical meaning of accumulative effective dispersion, i.e., $R^{\mathrm{eff}}(z) = \int_0^z D_{\mathrm{loc}}/T_{\mathrm{lin}}^2$. Furthermore, T_{lin} is the periodic solution of the linear TM-equations

$$T'_{\mathrm{lin}}(z) = 4D_{\mathrm{loc}}(z)M_{\mathrm{lin}}(z), \quad T_{\mathrm{lin}}(0) = T_0 > 0,$$
$$M'_{\mathrm{lin}}(z) = \frac{D_{\mathrm{loc}}(z)}{T_{\mathrm{lin}}(z)^3}, \qquad M_{\mathrm{lin}}(0) = 0,$$

which is explicitly known. $U(z)$ denotes the group generated by the harmonic oscillator, i.e., $U(z) = \exp(iAz)$ with $Au = \triangle u - x^2u$. It is essential for the whole approach that $S(z)$ is 1-periodic since $\langle D_{\mathrm{loc}}/T_{\mathrm{lin}}^2\rangle = 0$. Equation (2.6) describes averaged pulse propagation in a strong dispersion-managed system after lens transformation. In the

region of moderate energy values $N^2 < \overline{N^2}(\langle D \rangle)$ we have shown by numerical simulations in [13] that

$$a, b > 0$$

in contrast to former discussions of the problem (see [25] for details), and hence the potential is attracting. In order to transform (2.6) to the standard bifurcation problem (1.3) we consider steady-state solutions of the following form:

$$u(x,z) = \phi(\gamma x)\exp(-i\sqrt{ab}\lambda z), \quad \text{with } \gamma = \left(\frac{b}{a}\right)^{1/4}.$$

In the new variable $\xi = \gamma x$ we have $-\phi_{\xi\xi} + \xi^2\phi + F(\phi) = \lambda\phi$ with

$$F(\phi) := -\frac{N^2}{\sqrt{ab}} \int_0^1 S^{-1}(z) \left(\frac{1}{T(z)} |S(z)\phi|^2 S(z)\phi \right) dz. \tag{2.7}$$

Writing again u instead of ϕ and x instead of ξ we end up at (1.3) with a highly nonlocal nonlinearity which can be expressed in terms of Mehler's kernel and satisfies all the assumptions $(\mathcal{F}_1)$ throughout $(\mathcal{F}_6)$; see the appendix.

3. Analysis of the nonlinear eigenvalue problem. In this section we present the analysis of the nonlinear eigenvalue problem (1.3) and state our main results. Thereby the developed theory is sufficiently general to cover both applications, BEC and dispersion-managed optical fibers. However, our main goals are to generalize the bifurcation result of Kunze et al. [12] to nonlocal nonlinearities and to characterize the bifurcating solutions more precisely by variational arguments.

3.1. Bifurcation analysis. In this subsection we investigate the bifurcation behavior of (1.3). We strongly rely on the paper by Kunze et al. [12].

3.1.1. Preliminaries. The key property of the space X is the following [28].

LEMMA 3.1. *The embedding $X \hookrightarrow L^q(\mathbb{R})$ is compact for $2 \le q < \infty$.*

Next we consider the linear problem corresponding to (1.3), i.e.,

$$-u_{xx} + x^2 u - \lambda u = 0. \tag{3.1}$$

The following properties of the linear harmonic oscillator are well known [12].

LEMMA 3.2. *Let $\lambda_n = 2n+1, n \in \mathbb{N}_0$, and*

$$u_n(x) := \frac{1}{\sqrt{2^n n! \sqrt{\pi}}} \exp(-x^2/2) H_n(x),$$

where H_n is the nth Hermite polynomial

$$H_n(x) := (-1)^n \exp(x^2) \frac{d^n}{dx^n} \exp(-x^2).$$

(i) *λ_n are exactly the eigenvalues of (3.1). They are simple and the corresponding eigenfunctions are given by $u_n \in X$.*
(ii) *The eigenfunctions u_n of (3.1) form a complete orthonormal system of $L^2(\mathbb{R})$.*

3.1.2. Global bifurcation behavior. In this section we apply the global bifurcation theorem of Rabinowitz [22] to (1.3). The proof is similar to the one by Kunze et al. [12]. Due to the potential property of F we can define a weak solution. Let

$$S_0 := \{(\lambda, u) \in \mathbb{R} \times X : u \neq 0 \text{ is a weak solution of } -u_{xx} + x^2 u + F(u) = \lambda u\}$$

denote the set of all nontrivial solutions of (1.3) and $S = \overline{S_0}^{\mathbb{R}\times X}$ denote its closure. It is clear that $u \in X \subset H^1(\mathbb{R})$ implies $u \in C_0^0(\mathbb{R})$, the set of continuous functions on $\mathbb{R}$ vanishing for $x \pm \infty$. Using a bootstrapping argument it follows then that $u \in C^\infty(\mathbb{R})$, cf. [26], and consequently a weak solution is a classical solution. Our global bifurcation result then reads as follows

THEOREM 3.3. *Let F satisfy assumptions $(\mathcal{F}_1)$–$(\mathcal{F}_3)$. Then for all $n \in \mathbb{N}_0$, $(\lambda_n, 0)$ is a bifurcation point. Let $\mathcal{C}_n$ denote the component of $\mathcal{S}$ with $(\lambda_n, 0) \in \mathcal{C}_n$. Then the following alternative holds: Either*

(i) *$\mathcal{C}_n$ is unbounded in $\mathbb{R} \times X$ or*

(ii) *$\mathcal{C}_n$ is compact and there exists $m \neq n$ such that $(\lambda_m, 0) \in \mathcal{C}_n$.*

Proof. Using the Green's function $g(x, \xi)$ the problem is transformed to an integral problem; see [12] for details. To apply the global bifurcation theorem of Rabinowitz the resulting nonlinearity should be compact and of higher order. Roughly speaking, this is guaranteed by the growth condition together with the compact embedding. In particular, assumptions $(\mathcal{F}_2)$, $(\mathcal{F}_3)$ yield the assertions of Lemma 5 in [12]; the proof then is along the lines of [12]. ∎

Usually the second alternative is ruled out by nodal arguments; see [12] for details. These arguments are no longer valid for nonlocal nonlinearities. However, in section 3.2 we will show by variational arguments that the bifurcating branches are unbounded.

3.1.3. Local bifurcation behavior and orbital stability. In order to determine the local behavior in the vicinity of $(\lambda_n, 0)$ we introduce a nondegeneracy condition on the nonlinearity

$$(\mathcal{F}_3^n) \qquad F \in C^3(X, L^2) \text{ with } \langle \delta^3 F(0)[u_n]^3, u_n\rangle \neq 0.$$

Note that by virtue of $(\mathcal{F}_2)$, $(\mathcal{F}_3)$ the first derivatives of F vanish, i.e., $\delta F(0)[u_n] = 0$ and $\delta^2 F(0)[u_n]^2 = 0$. The local bifurcation behavior then can be determined as follows.

LEMMA 3.4. *Let F satisfy $(\mathcal{F}_1)$, $(\mathcal{F}_2)$, $(\mathcal{F}_3^n)$. Then there exists $\epsilon > 0$ such that $(\lambda, u) \in \mathcal{C}_n \cap U_\epsilon(\lambda_n, 0)$ implies*

$$\lambda = \lambda_n + \lambda(s), \quad u = s u_n + s v_n(s),$$

where $0 < |s| < \epsilon$ and $\lambda(0) = 0$, $\lambda'(0) = 0$, and

$$\operatorname{sgn}(\lambda''(0)) = \operatorname{sgn}\big(\langle \delta^3 F(0)[u_n]^3, u_n\rangle\big). \tag{3.2}$$

Moreover, $v_n(0) = 0$ with $(v_n(s), u_n)_X = 0$.

Proof. The assertions follow by standard Lyapunov–Schmidt theory since the eigenvalues are simple. ∎

Thus, the direction of bifurcation is determined by the sign of $\lambda''(0)$. In the case of negative sign the bifurcating solutions from the smallest eigenvalue λ_0 are orbitally stable by the method of Rose and Weinstein [23]; otherwise they are unstable. Instead of discussing this in detail we refer to the next section where the solutions are characterized as ground states of the energy functional in the situation where the nonlinearity is focusing.

3.2. Variational calculus. In this section we characterize the bifurcating solutions by variational arguments and identify them as minimizers or saddle points of the corresponding energy functional. Throughout the section we assume that F has the potential property.

Considering the energy functional corresponding to (1.3)

$$J(u) = \frac{1}{2}\|u\|_X^2 + G(u), \tag{3.3}$$

it is easy to observe that critical points of J correspond to nonlinear bound states of (1.3). Now we are in a position to state our main theorem, which characterizes the bifurcating solutions. In particular, it shows the existence of infinitely many nonlinear modes of arbitrary energy.

THEOREM 3.5. *Suppose $(\mathcal{F}_1)$–$(\mathcal{F}_6)$ are satisfied and $\omega > 0$ is given. Then there exists an unbounded sequence $\{u_n^\omega\}_{n\in\mathbb{N}_0} \subset X$ of critical points of J with $\|u\|_2^2 = \omega$ and corresponding Lagrange multipliers $\lambda_n^\omega \leq \lambda_n$ with the following properties:*

(i) *u_0^ω is the ground state of the energy functional, and it is even and orbitally stable. Moreover, $(\lambda_0^\omega, u_0^\omega) \in \mathcal{C}_0$.*

(ii) *u_1^ω is odd and minimizes J among all odd functions. Furthermore, we have $(\lambda_1^\omega, u_1^\omega) \in \mathcal{C}_1$.*

(iii) *The u_n^ω with $n > 1$ are saddle points with $(\lambda_n^\omega, u_n^\omega) \in \mathcal{C}_n$. Moreover, u_{2k}^ω is even and u_{2k+1}^ω is odd.*

In the case of a simple nonlinearity $F(u) = -|u|^2u$, it is shown that the ground states are positive [6], using Kato's inequality and maximum principle. Moreover, the solutions of the simpler equation decay like a Gaussian and are unique. However, all these arguments need information about nodal properties of the solutions, which are not at hand for the nonlocal nonlinearity we have in mind.

In particular the second alternative of the global bifurcation theorem can be ruled out by Theorem 3.5.

COROLLARY 3.6.

1. *$\mathcal{C}_n$ is unbounded in both u and λ.*
2. *$\mathcal{C}_n \cap \mathcal{C}_m = \emptyset$ for $n \neq m$.*

Proof of Theorem 3.5. In order to increase the clarity and due to the practical importance of the first two modes, the proof is divided into three parts.

Since the ground states of the equation are of fundamental importance, we discuss them first. Define Γu as $\Gamma u(x) = u(-x)$. Due to assumption $(\mathcal{F}_6)$ the functional J is invariant under Γ. With $X_\Gamma := \{u \in X | \Gamma u = u\}$ the following lemma holds.

LEMMA 3.7 (characterization of the ground state). *For all $\omega > 0$ the minimization problem*

$$J_\Gamma^\omega = \min\{J(u) | u \in X_\Gamma, \|u\|_2^2 = \omega\} \tag{3.4}$$

has a nontrivial solution $u^\omega \in X_\Gamma$ which corresponds to a weak solution of (1.3). *Moreover, u^ω is orbitally stable as the ground state of the equation, that is,*

$$J(u^\omega) = \min\{J(u) | u \in X, \|u\|_2^2 = \omega\}. \tag{3.5}$$

Furthermore, $\{(\lambda^\omega, u^\omega) | \omega \in (0, \infty)\} \subset \mathcal{C}_0$.

Proof. The proof relies on the principle of symmetric criticality [18], which allows us to reduce the problem to even functions, i.e., a minimizer of problem (3.4) is a critical point for the whole problem; see [11] for a similar result. The following lemma is essential to show that J is bounded from below.

LEMMA 3.8. *For $u, v \in X$ the following estimate holds with $\alpha' := \alpha/2 + 1/4 < 2$ and $\beta' = \beta/2 + 3/4$:*

$$|G(u) - G(v)| \le C(\|u\|_X^{\alpha'} + \|v\|_X^{\alpha'})(\|u\|_2^{\beta'} + \|v\|_2^{\beta'})\|u - v\|_2. \tag{3.6}$$

Proof. We calculate as follows:

$$\begin{aligned}
|G(u) - G(v)| &= \left| \int_0^1 \frac{d}{ds} G(su + (1-s)v) ds \right| = \left| \int_0^1 G'(su + (1-s)v)(u - v) ds \right| \\
&= \int_0^1 |(F(su + (1-s)v), u)| ds \le C \int_0^1 \|F(su + (1-s)v)\|_2 \|u - v\|_2 ds \\
&\le C \int_0^1 \|su + (1-s)v\|_X^{\alpha/2+1/4} \|su + (1-s)v\|_2^{\beta/2+3/4} \|u - v\|_2 ds \\
&\le C(\|u\|_X^{\alpha'} + \|v\|_X^{\alpha'})(\|u\|_2^{\beta'} + \|v\|_2^{\beta'})\|u - v\|_2
\end{aligned}$$

which is the assertion of the lemma. ∎

The above lemma together with assumption $(\mathcal{F}_2)$ shows

$$J(u) = \frac{1}{2}\|u\|_X^2 + G(u) \ge \frac{1}{2}\|u\|_X^2 - C\|u\|_X^{\alpha'} \omega^{(\beta'+1)/2} \tag{3.7}$$

which is bounded from below since $\alpha' < 2$. Making use of the fact that the embedding $X_\Gamma \subset\subset L^p(\mathbb{R})$ is compact for $p \ge 2$ we are able to show that a minimizer on X_Γ exists. Let $\{u_n\}$ be a minimizing sequence; that is, $\|u_n\|_2^2 = \omega$ and $J(u_n) \to J_\Gamma^\omega$, which implies that $J(u_n)$ is bounded, say $J(u_n) \le M$, and using (3.7) we conclude that u_n is bounded in X. Passing to a subsequence we may assume that $u_n \rightharpoonup u$ weakly in X and $u_n \to u$ strongly in $L^2(\mathbb{R})$, which yields $\|u\|_2^2 = \omega$. From (3.6) we obtain

$$|G(u_n) - G(u)| \le C\|u - v\|_2 \tag{3.8}$$

since u_n, u are bounded in X. Finally, it follows that $J(u) \le \lim_{n\to\infty} J(u_n) = J_\Gamma^\omega$, and accordingly $u \in X_\Gamma$ is the desired minimizer. The principle of symmetric criticality then reveals that u is a critical point of J and, consequently, a weak solution of (1.3) with Lagrange-multiplier λ^ω.

$$-u_{xx}^\omega + x^2 u^\omega + F(u^\omega) = \lambda^\omega u. \tag{3.9}$$

Next we show $\lim_{\omega\to 0} \lambda^\omega = \lambda_0$, where λ_0 is the eigenvalue of the harmonic oscillator corresponding to u_0.

Therefore, we take the inner product of the Euler–Lagrange equation (3.9) with u^ω to obtain

$$\|u^\omega\|_X^2 + (F(u^\omega), \omega) = \lambda^\omega \omega; \quad \text{hence } \lambda^\omega = \frac{\|u^\omega\|_X^2}{\omega} + \frac{(F(u^\omega), u^\omega)}{\omega}. \tag{3.10}$$

Note that λ_0 can be characterized by the Rayleigh quotient as follows [3]:

$$\lambda_0 = \|u_0\|_X^2 = \inf\left\{ \frac{\|u\|_X^2}{\|u\|_2^2} \middle| u \in X, u \ne 0 \right\} = \inf\left\{ \frac{\|u\|_X^2}{\|u\|_2^2} \middle| u \in X_\Gamma, u \ne 0 \right\},$$

where the last equality holds since u_0 is even. Accordingly, since $\beta' > 3/4$

$$\begin{aligned}\lambda^\omega &\geq \lambda_0 + \frac{(F(u^\omega), u^\omega)}{\omega} \geq \lambda_0 - \frac{\|F(u^\omega)\|_2}{\sqrt{\omega}} \\ &\geq \lambda_0 - \|u^\omega\|_X^{\alpha'} \|u^\omega\|_2^{\beta'-1/2} \geq \lambda_0 - \omega^{\beta'/2-1/4} \to \lambda_0 \quad \text{for } \omega \to 0.\end{aligned}$$

On the other hand G can be written as

$$0 > G(u) = \int_0^1 (F(su), u) ds \geq \int_0^1 s^\delta ds (F(u), u) = \frac{1}{\delta+1}(F(u), u)$$

which shows, in particular, $(F(u), u) < 0$. By definition of J^ω we can write

$$\begin{aligned}\lambda^\omega &= 2\frac{J(u^\omega) - G(u^\omega)}{\omega} + \frac{(F(u^\omega), \omega)}{\omega} \leq 2\frac{J(\sqrt{\omega}u_0)}{\omega} + \frac{(F(u^\omega), \omega) - 2G(u^\omega)}{\omega} \\ &= \lambda_0 + \frac{2G(\sqrt{\omega}u_0) - 2G(u^\omega) + (F(u^\omega), u^\omega)}{\omega} \qquad (3.11) \\ &\leq \lambda_0 + \frac{2G(\sqrt{\omega}u_0)}{\omega} + \left(1 - \frac{2}{\delta+1}\right)(F(u^\omega), u^\omega) \leq \lambda_0.\end{aligned}$$

Thus we have shown $\lambda^\omega \to \lambda_0$ for $\omega \to 0$ which implies that the u^ω are bifurcating from the first eigenvalue. Moreover, due to $\lambda^\omega \leq \lambda_0$ the direction of bifurcation is determined. In a similar way to the argument of Zhang [28] it follows that the bifurcating solutions are orbitally stable.

The uniqueness of the bifurcating solutions from the global bifurcation theorem (up to phase translation) and the fact that the same arguments apply for the (possibly nonsymmetric) ground state $\tilde{u}^\omega$ lead to $u^\omega = \tilde{u}^\omega$ at least for small ω and hence the ground states are even. In addition the solutions u^ω exist for arbitrary $\omega > 0$, and consequently the branch is unbounded in $L^2(\mathbb{R})$ and hence also in X which rules out the second alternative of the global bifurcation theorem.

Next we will show that the solutions bifurcating in the second eigenvalue λ_1 are odd. It should be noted that these modes are somewhat orbitally stable among all odd functions which yields their practical relevance. In order to find odd solutions we introduce the action $\Gamma_2 u(x) = -u(-x)$ and the corresponding space X_{Γ_2} of fixed points of Γ_2 and apply the principle of symmetric criticality again to obtain the following lemma.

LEMMA 3.9. *For all $\omega > 0$ the minimization problem*

$$J^\omega_{\Gamma_2} = \min\{J(u) | u \in X_{\Gamma_2}, \|u\|_2^2 = \omega\} \qquad (3.12)$$

has a nontrivial solution $u^\omega \in X_{\Gamma_2}$, which corresponds to a weak solution of (1.3). *Furthermore, $\{(\lambda^\omega, u^\omega,) | \omega \in (0, \infty)\} \subset \mathcal{C}_1$.*

Proof. The proof of the existence of a minimizer can be adapted by replacing X_Γ with X_{Γ_2} from Lemma 3.7. It remains to verify the behavior for $\omega \to 0$. Note that λ_1 can be characterized as

$$\lambda_1 = \|u_1\|_X^2 = \inf\left\{ \frac{\|u\|_X^2}{\|u\|_2^2} \middle| u \in X_{\Gamma_2}, u \neq 0 \right\}.$$

and using (3.10) it follows that

$$\lambda^\omega \geq \lambda_1 - \|u^\omega\|_X^{\alpha'} \|u^\omega\|_2^{\beta'-1/2} \geq \lambda_1 - \omega^{\beta'/2-1/4} \to \lambda_1 \quad \text{for } \omega \to 0.$$

In the same way as in (3.11) we then can verify the direction of bifurcation, i.e., $\lambda^\omega \leq \lambda_1$. Hence $\lambda^\omega \to \lambda_1$. ∎

Next we show the existence of infinitely many modes of the equation which correspond to saddle points of the energy functional J. Obviously this includes the previously discussed situations, but the proof is more technical and we have therefore discussed the previous cases separately.

LEMMA 3.10. *For arbitrary $\omega > 0$ there exists an unbounded sequence of even (odd) solutions $u_n \in X$ of (1.3) with $\|u\|_2^2 = \omega$.*

The proof relies on the following result [8] which is a generalization of the theorem of Ljusternik–Schnirelmann for infinite-dimensional Hilbert spaces.

THEOREM 3.11. *Let X be an infinite-dimensional Hilbert space, $J, K \in C^1(X, \mathbb{R})$ with $K'(v) \neq 0$ for all $v \in X - \{0\}$ and $S := \{v \in X : K(v) = 0\}$. If $J|_S$ is even and bounded from below and satisfies the Palais–Smale condition on S, then there exist infinitely many critical values, that is, $c_k \in \mathbb{R}$ with $\lim_{k\to\infty} c_k = \infty$ and for all $k \in \mathbb{N}$ there exists a pair $(\lambda_k, v_k) \in \mathbb{R} \times S$ with $J(v_k) = c_k$ and $J'(v_k) - \lambda_k K'(u_k) = 0$. Moreover, $c_k \to \infty$.*

Proof of Lemma 3.10. We apply the theorem with $K(u) = \|u\|_2^2 - \omega$ and $J(u)$ defined as in (3.3) for $X = X_\Gamma$ (resp., $X = X_{\Gamma_2}$) separately. It suffices to show that J satisfies the Palais–Smale condition on S. Let (u_n, λ_n) be a Palais–Smale sequence, that is,

$$J(u_n) \to c \in \mathbb{R}, \qquad J'(u_n) - \lambda_n K'(u_n) \to 0 \quad \text{in } X',$$

where X' denotes the dual space of X. We have to show the existence of a strongly convergent subsequence. As in the proof of Lemma 3.7 we can extract a subsequence still denoted as (u_n, λ_n) with $u_n \rightharpoonup u$ weakly in X and $u_n \to u$ strongly in X. It remains to show $u_n \to u$ strongly in X. Therefore we calculate

$$\begin{aligned}\|u_n - u\|_X^2 &= (J'(u_n) - J'(u))(u_n - u) - (F(u_n) - F(u), u_n - u)\\ &\le C\|J'(u_n)\|_{X'} + |J'(u)(u_n - u)| + \|F(u_n) - F(u)\|_2 \|u_n - u\|_2\\ &\le C\|J'(u_n)\|_{X'} + |J'(u)(u_n - u)| + C\|u_n - u\|_X^{1/4} \|u_n - u\|_2^{7/4}.\end{aligned}$$

Thus, it follows that $u_n \to u$ in X, and from (3.10) it follows that

$$\begin{aligned}|\lambda_n - \lambda| &\le \big|\|u_n\|_X^2 - \|u\|_X^2\big| + \|(F(u_n), u_n) - (F(u), u)\|\\ &\le \big|\|u_n\|_X^2 - \|u\|_X^2\big| + |(F(u_n) - F(u), u_n)| + |(F(u), u_n - u)|\end{aligned}$$

which together with $(\mathcal{F}_2)$ and $u_n \to u \in X$ yields the desired convergence. ∎

We still have to show $(\lambda_n^\omega, u_n^\omega) \in \mathcal{C}_n$. Instead of discussing this in detail we refer to the characterization of the critical values in [8],

$$c_n^\omega = \inf_{A \in \mathcal{B}_n} \max_{u \in A} \frac{J(u)}{\omega}, \quad \text{where } \mathcal{B}_n := \{A \in s(X) : A \subset S, \text{compact with } \gamma(A) \ge k\},$$

and $s(X)$ denotes the set of all nonempty, closed subsets S of X which are symmetric to the origin and satisfy $0 \notin S$. Moreover, $\gamma(A)$ denotes the genus of A, i.e.,

$$\gamma(A) := \inf\{n \ge 1 : \exists \phi : A \to \mathbb{R}^n - \{0\} \text{ continuous and odd}\}.$$

Using (3.10) and calculating as in (3.11) give $\lambda_n^\omega \le 2c_n^\omega$. The assertion then again follows from $G(u) < 0$ combined with the min-max characterization of the eigenvalues, i.e.,

$$\inf_{A \in \mathcal{B}_n} \max_{v \in A} \frac{\|u\|_X^2}{\|u\|_2^2} = \min_{X_n \subset X} \max_{u \in X_n} \frac{\|u\|_X^2}{\|u\|_2^2} = \max_{X_{n-1} \subset X} \min_{u \in X_{n-1}^\perp} \frac{\|u\|_X^2}{\|u\|_2^2} = \lambda_n, \tag{3.13}$$

where X_n denotes a subspace of X of dimension n. Thereby, the first equation can be found in [9], while the second holds due to the equivalence of the principle of Courant–Fischer and the characterization of the nth eigenvalue by Poincare; cf. [3]. Again, (3.13) implies $\lambda_0 \leq \lim_{\omega\to 0} \lambda_n^\omega \leq \lambda_n$. Hence λ_n^ω must converge to an eigenvalue which must be λ_n by induction. ∎

Thus, all statements of the theorem are proven. ∎

For some applications it may be of interest to fix the wave-number. Due to the direction of bifurcation we have the following corollary.

COROLLARY 3.12. *For all* $\lambda \in \mathbb{R}$ (1.3) *has infinitely many solutions* $\{u_n^\lambda\}_{n\in\mathbb{N}_0}$.

Proof. Since all the branches are unbounded in both u and λ, the existence of infinitely many solutions u_n with fixed wave-number λ is obvious. Note that due to $\lambda_n^\omega \leq \lambda_n$ we have $\lambda_n^\omega \to -\infty$ for $\omega \to \infty$. ∎

Remark 3.13. Requiring $\|u\|_p \leq |G(u)|$ for some $p \geq 2$ one can verify that the sequence $\{u_n^\lambda\}$ is unbounded in X [14].

4. Exponential decay of the bound states. In this section it is shown that in case of the nonlinearity (2.7) arising from the context of fiber optics all solutions decay exponentially fast as $x \to \infty$.

THEOREM 4.1. *Let* (λ, u) *be a solution of the nonlinear eigenvalue problem* (1.3) *with nonlinearity* (2.7). *Then there exists* $C > 0$ *such that*

$$|u(x)| + |u_x(x)| \leq C \exp(-|x|/2).$$

Remark 4.2. It is not clear to us whether the decay rate is sharp or not in a rigorous mathematical way. But regarding numerical results or reduced models it turns out that the solution indeed decays only exponentially fast. In fact it is well known that the DM-soliton has a Gaussian kernel and the envelope of its oscillating tails decays exponentially; cf. ([24, 25]). Thus, the above decay rate is the best that one can expect. In conclusion, the Gaussian decay which occurs for simpler nonlinearities is lost due to the nonlocal properties of F.

Proof. The proof is similar to Theorem 8.1.1 in [1], where exponential decay is verified without potential. For $\epsilon > 0$ define the function

$$f^\epsilon(x) := \exp\left(\frac{x}{1+\epsilon x}\right)$$

which has the following properties [1]:

- f^ϵ is bounded for all $\epsilon > 0$,
- $f_x^\epsilon(x) \leq f^\epsilon(x)$,
- $\lim_{\epsilon\to 0} f^\epsilon(x) = \exp(x)$.

Multiplication of (1.3) with $f^\epsilon \overline{u}$ and integration give in the real part

$$\int_{\mathbb{R}} (x^2 - \lambda) f^\epsilon |u|^2 dx = I - \Re\left(\int_{\mathbb{R}} u_x (f^\epsilon \overline{u})_x dx\right), \tag{4.1}$$

$$I := \Re\left(\int_0^1 \int_{\mathbb{R}} S^{-1}(z)\left(\frac{1}{T(z)}|S(z)u|^2 S(z)u\right) f^\epsilon \overline{u} dz dx\right). \tag{4.2}$$

At first we bound the left-hand side from below. Defining $R_1 := \sqrt{|\lambda|+1}$ we find

$$\int_{\mathbb{R}} (x^2-\lambda) f^\epsilon |u|^2 dx \geq \int_{|x|\leq R_1} (x^2-\lambda) f^\epsilon |u|^2 dx + (R_1^2 - \lambda) \int_{|x|>R_1} f^\epsilon |u|^2 dx \tag{4.3}$$

which due to $R_1^2 - \lambda \geq 1$ implies the inequality

$$\int_{|x|>R_1} f^\epsilon |u|^2 dx \leq \int_{\mathbb{R}} (x^2 - \lambda) f^\epsilon |u|^2 dx - \int_{|x|\leq R_1} f^\epsilon |u|^2 dx. \tag{4.4}$$

Next we estimate the right-hand side: The second term is bounded in ϵ, whereas for the first term, regarding (4.1) and using $f_x^\epsilon \leq f^\epsilon$,

$$\begin{aligned}
\Re\left(\int_{\mathbb{R}} u_x (f^\epsilon u)_x dx\right) &= \Re\left(\int_{\mathbb{R}} f^\epsilon |u_x|^2 + f_x^\epsilon u_x \overline{u} dx\right) \geq \int_{\mathbb{R}} f^\epsilon |u_x|^2 - f^\epsilon |u||u_x| dx \\
&\geq \int_{\mathbb{R}} f^\epsilon |u_x|^2 dx - \left(\int_{\mathbb{R}} f^\epsilon |u_x|^2 dx\right)^{1/2} \left(\int_{\mathbb{R}} f^\epsilon |u|^2 dx\right)^{1/2} \\
&\geq \int_{\mathbb{R}} f^\epsilon |u_x|^2 dx - \frac{1}{2}\left(\int_{\mathbb{R}} f^\epsilon |u|^2 dx + \int_{\mathbb{R}} f^\epsilon |u_x|^2 dx\right).
\end{aligned}$$

Hence, by collecting all the terms and splitting the last integral we can conclude that

$$\frac{1}{2}\left(\int_{\mathbb{R}} f^\epsilon |u_x|^2 dx + \int_{|x|>R_1} f^\epsilon |u|^2 dx\right) \leq I - \int_{|x|\leq R_1} (x^2 - \lambda) f^\epsilon |u|^2 dx + \frac{1}{2}\int_{|x|\leq R_1} f^\epsilon |u|^2 dx$$

and it remains to estimate I. Splitting the integral into $|x| \leq R$ and $|x| > R$ with $R > R_1$ to be determined later, the following is true due to $f^\epsilon u \in X$ for all $\epsilon > 0$:

$$\begin{aligned}
&\left|\int_0^1 \int_{|x|>R} S^{-1}(z)\left(\frac{1}{T(z)}|S(z)u|^2 S(z)u\right) f^\epsilon \overline{u} dx dz\right| \\
&\quad \leq \max_{z\in[0,1]}\left(\left\|\frac{1}{T(z)} S(z)u\right\|^2_{L^\infty(x>R)}\right)\left|\int_0^1 \int_{x>R} S(z)u\overline{S(z)\,(f^\epsilon u)} dx dz\right| \\
&\quad = \max_{z\in[0,1]}\left(\frac{1}{T(z)}\|S(z)u\|^2_{L^\infty(x>R)}\right)\int_{|x|>R} f^\epsilon |u|^2 dx.
\end{aligned}$$

Since $S(z)u \in L^2(\mathbb{R})$ in particular $S(z)u(x) \to 0$ for $x \to \pm\infty$. Accordingly there exists $R_2 > R_1$ with

$$\max_{z\in[0,1]}\left(\frac{1}{T(z)}\|S(z)u\|^2_{L^\infty(x>R_2)}\right) < \frac{1}{4}.$$

Using

$$\frac{1}{4}\int_{|x|>R_2} f^\epsilon |u|^2 dx \leq \frac{1}{2}\int_{|x|>R_1} f^\epsilon |u|^2 dx - \frac{1}{4}\int_{|x|>R_2} f^\epsilon |u|^2 dx$$

the following estimate holds:

$$\begin{aligned}
\frac{1}{2}\int_{\mathbb{R}} f^\epsilon |u_x|^2 dx + \frac{1}{4}\int_{|x|>R_2} f^\epsilon |u|^2 dx \leq &-\int_{|x|\leq R_1} (x^2 - \lambda) f^\epsilon |u|^2 dx + \frac{1}{2}\int_{|x|\leq R_1} f^\epsilon |u|^2 dx \\
&+ \left|\int_0^1 \int_{|x|\leq R_2} S^{-1}(z)\left(\frac{1}{T(z)}|S(z)u|^2 S(z)u\right) dz f^\epsilon \overline{u} dx\right|.
\end{aligned}$$

On the right-hand side the limit $\epsilon \to 0$ is finite due to the boundedness of the domain of integration. Hence

$$\int_{x>R_2} \exp(|x|)|u(x)|^2 dx < \infty \quad \text{and} \quad \int_{\mathbb{R}} \exp(|x|)|u_x|^2 dx < \infty,$$

which implies

$$\int_{\mathbb{R}} \exp(|x|)(|u(x)|^2 + |u_x(x)|^2) dx < \infty. \tag{4.5}$$

Using the Lipschitz continuity of u, we are then able to derive exactly as in [1] the existence of a $C > 0$ with

$$\exp(|x|)(|u(x)|^2 + |u_x(x)|^2) < C \quad \forall x \in \mathbb{R},$$

which gives the desired decay estimate. □

5. Conclusion. In this section we explain the meaning of the derived results in the context of dispersion-managed optical fibers and discuss their practical relevance.

Since N^2 in the nonlinear TM-equations has the physical meaning of pulse energy, we are interested in unit-norm solutions of the DM-NLS after lens tranformation. Accordingly we apply our main theorem with $\omega = 1$ to obtain the following.

THEOREM 5.1. *There exists a sequence* $\{\lambda_n, u_n\}$ *of solutions of the averaged equation* (2.6) *having the form*

$$u_n(x,z) = \phi_n(\gamma x)\exp(-i\sqrt{ab}\lambda z), \quad \gamma = \left(\frac{b}{a}\right)^{1/4},$$

where ϕ_{2k} *is even and* ϕ_{2k+1} *is odd. Moreover, there exists* $C > 0$ *such that*

$$|\phi_n(x)| + |\phi_n'(x)| < C\exp\left(-\gamma\frac{|x|}{3}\right).$$

Thereby, the well-known Gaussian decay [9] which occurs for $F(u) = -|u|^2u$ is lost due to the nonlocal properties of F which prevents nodal arguments or a maximum principle. Thus we have shown the existence of infinitely many even (resp., odd) solutions of (1.3). Note that symmetry of the DM-soliton was not rigorously proven up to now. Kunze [11] considered only the case of two spatial dimensions, but in the context of nonlinear optics space and time variables are interchanged and hence the one-dimensional case is of practical interest.

It should be noted that a rigorous averaging theorem as in [26] is not at hand due to the problem of expanding the parameters a and b in powers of $\epsilon = \langle D \rangle$. However, heuristically it is clear that a similar averaging theorem is valid, i.e., the solution of the averaged equation is ϵ-close to the solution of the original (lens transformed) problem on the time scale $1/\epsilon$ in some H^s-space. Consequently, we expect the periodic solution of the averaged problem to be close to the original pulse and the approximation to be accurate if z is not too large.

Before discussing the relevance of our results to dispersion-managed solitons, we compare them to the relevant publications and add some comments on the differences.

(i) In recent publications (cf. [19, 20, 27]), it is explained that DM-solitons do not exist as an exactly periodic solution of the original DM-NLS (2.1) due to parametric

resonance. Using a perturbation series expansion, a coupling between bound states and linear Bloch waves is observed which results in a decay of the DM pulse for $z \to \infty$. In contrast, the averaged equation possesses a true periodic solution which is close to the solution of the original problem in terms of the averaging procedure.

(ii) Due to the lens transformation the continuous spectrum of the DM-NLS becomes discrete due to the parabolic potential and the parametric resonance is destroyed. Therefore, we have been able to apply standard bifurcation theory to our problem in contrast to the equation considered by Zharnitsky et al. [26].

This can be understood as follows: Since the lens transformation makes use of some well-known properties of the DM-soliton some resonant terms in the leading order are removed. Of course the problems with and without lens transformation are equivalent before averaging, but regarding the averaged variational principle it turns out that we reduce the problem to particular solutions, i.e., bound states. Nevertheless, both averaged models give raise to solutions close to the real DM pulse but they describe different problems; i.e., they lead to different approximations for the solution. Hence, it is not surprising that the spectrum has changed. Another way to understand the discrepancy is that some characteristic features of the DM-soliton as the existence of a pulse chirp are implemented in the lens transformation, and consequently, the equation studied in our paper is a better approximation from a practical point of view. However, in a rigorous mathematical way both are just first order approximations of the same equation.

Finally, we discuss the relevance of our result to the original problem. We have shown the existence of infinitely many solutions of the original problem which are close

$$A(z,t) = \frac{N^2}{T(z)} U(R^{\text{eff}}(z)) \left\{ \exp(i\lambda z) u_n(t/T(z)) \right\} \exp\left(i t^2 \frac{M(z)}{T(z)} \right),$$

where u_n decays exponentially fast. Note that the ground state obtained by Zharnitsky et al. [26] was only shown to be in $L^2(\mathbb{R})$. The relevance of our theoretical results can be summarized as follows:

- The ground state ϕ_0 corresponds to the DM-soliton; we have shown that it is an even function at least for small input energies. This is a new theoretical result already known from numerical simulations.
- Moreover, it is shown that the DM-soliton decays exponentially fast and has a Gaussian core. Numerical simulations show the existence of an "optimal" energy N^2, where γ and, accordingly, the decay rate is maximized [14].
- Uniqueness of the DM-soliton is still an open question, reduced models indicate that the DM-solitons form a one-parameter family. We have shown in the present paper the uniqueness of the DM-soliton close to the bifurcation point (that is for small energies), but there could exist a secondary (symmetry-breaking) bifurcation.
- The odd solution ϕ_1 corresponds to the bi-soliton which was first observed by Maruta, Nonaka, and Yoshika [16] by numerical simulations; see also the work of Pare and Belanger [21]. It minimizes the energy functional with respect to all odd functions and is hence stable against odd perturbations. It is a promising candidate for the reduction of intrachannel interactions which play an important role in todays multichannel systems. Numerical simulations [15] show that the bi-soliton propagates stable over long distances and the bit rate is increased significantly by a new encoding scheme.

- Furthermore, we have verified the existence of modes of arbitrary order, a fact which was unknown up to now. Similar to the basis of Gauss–Hermite functions for the linear oscillator there exists a family of nonlinear modes with shape close to the corresponding eigenfunction. Maruta et al. [15] also observed a tri-soliton which corresponds to solutions on the third branch in the bifurcation diagram. They conjectured the existence of a periodic pulse of arbitrary order which is guaranteed by our result.
- A very effective way to derive approxmiations of the DM-soliton is to use a Hermite–Gaussian ansatz in the lens transformed equation; cf. [24]. From the results derived in this paper it is now clear why this method gives reasonable results. We have shown that the DM-soliton in the averaged equation is close to the first eigenmode. In [24], u is expanded in terms of the Gauss–Hermite eigenfunctions, the expansion is truncated after a few modes. Bearing the bifurcation result in mind it is now obvious that the error is small although infinitely many modes are omitted. With this method it should also be possible to obtain approximations for the solutions bifurcating from the other eigenvalues by considering the corresponding eigenfunction and its neighbors as perturbations.

In conclusion we have explained various facts on the DM-soliton which were only known from numerical simulations in an analytical way.

Appendix: Verification of assumptions. In this section we show that F as in (1.2) satisfies assumptions $(\mathcal{F}_1)$ to $(\mathcal{F}_6)$. The same holds obviously for the nonlinearity $F(u) = \sigma|u|^2u$ with $\sigma < 0$.

$(\mathcal{F}_1)$: By definition we have $S(z) = U(R^{\text{eff}}(z))$, where $U(t)$ denotes the group of the harmonic oscillator. Using

$$\overline{U(-t)u} = U(t)\overline{u}$$

together with $-R^{\text{eff}}(z) = R^{\text{eff}}(z + 1/2)$ implies

$$S(z)\overline{u} = U(R^{\text{eff}}(z))\overline{u} = \overline{U(-R^{\text{eff}}(z))u} = \overline{U(R^{\text{eff}}(z + 1/2))u}.$$

Hence, by $S^{-1}(z) = U(-R^{\text{eff}}(z))$,

$$\begin{aligned} &S^{-1}(z)\left(\frac{1}{T(z)}|S(z)\overline{u}|^2S(z)\overline{u}\right) \\ &\quad = \overline{S^{-1}\left(z+\frac{1}{2}\right)\left(\frac{1}{T(z)}\left|S\left(z+\frac{1}{2}\right)u\right|^2 S\left(z+\frac{1}{2}\right)u\right)}. \end{aligned}$$

Due to symmetry $T(z) = T(z + 1/2)$ we can conclude

$$\begin{aligned} F(\overline{u}) &= -\int_0^1 \overline{S^{-1}(z+1/2)\left(\frac{1}{T(z+1/2)}|S(z+1/2)u|^2S(z+1/2)u\right)}dz \\ &= \overline{-\int_{1/2}^{3/2} S^{-1}(z)\left(\frac{1}{T(z)}|S(z)u|^2S(z)u\right)dz} = \overline{F(u)}, \end{aligned}$$

where the last equality holds due to the 1-periodicity of S and T which gives the assertion for real-valued u.

($\mathcal{F}_2$): Due to $\|u\|_{L^1(0,1)} \le \|u\|_{L^2(0,1)}$ we can estimate in the following way:

$$\begin{aligned}\|F(u)-F(v)\|_2^2 &\le C\int_{\mathbb{R}}\left(\int_0^1 |S^{-1}(z)\big(|S(z)u|^2S(z)u-|S(z)v|^2S(z)v\big)\,|dz\right)^2 dx\\ &\le \int_{\mathbb{R}}\left(\int_0^1 \big|S^{-1}(z)\left(|S(z)u|^2S(z)u-|S(z)v|^2S(z)v\right)\big|^2 dz\right)dx\\ &= \int_0^1 \||S(z)u|^2S(z)u-|S(z)v|^2S(z)v\|_2^2\;dz.\end{aligned}$$

Making use of the inequality $\big||a|^2a-|b|^2b\big| \le \frac{3}{2}(|a|^2+|b|^2)|a-b|$ which holds for $a,b\in\mathbb{C}$ and the Cauchy–Schwarz inequality we conclude

$$\begin{aligned}&\||S(z)u|^2S(z)u-|S(z)v|^2S(z)v\|_2^2\\ &\le\ C\ \|\big(|S(z)u|^2+|S(z)v|^2\big)\,|S(z)u-S(z)v|\,\|_2^2\\ &\le\ C\ \|(|S(z)u|^2+|S(z)v|^2)^2\|_2\||S(z)(u-v)|^2\|_2\\ &\le\ C\ \left(\|\,|S(z)u|^4\|_2+\||S(z)v|^4\|_2\right)\||S(z)(u-v)|^2\|_2\\ &=\ C\ \left(\|S(z)u\|_8^4+\|S(z)v\|_8^4\right)\|S(z)(u-v)\|_4^2.\end{aligned}$$

From Oh [17] it is known that $S(z)$ is a bounded operator from $L^p(\mathbb{R})$ to $L^q(\mathbb{R})$, where $q=p/(p-1)$. Since X is compactly embedded in $L^p(\mathbb{R})$ for all $p\ge 2$ we have $S(z)u\in L^q(\mathbb{R})$ for all $q\ge 2$. Estimating the terms separately we interpolate

$$\|S(z)u\|_8 \le \|S(z)u\|_4^{\lambda}\|S(z)u\|_q^{1-\lambda},$$

where

$$\lambda=\frac{1/8-1/q}{1/4-1/q}.$$

Using the estimate of Sobolev type $\|v\|_4^4\le C\|v_x\|_2\|v\|_2^3$ we obtain

$$\begin{aligned}\|S(z)u\|_8 &\le C(\|S(z)u_x\|_2\|S(z)u\|_2^3)^{\lambda/4}\|S(z)u\|_X^{1-\lambda}\\ &\le C\|u\|_X^{\lambda/4}\|u\|_X^{1-\lambda}\|u\|_2^{3\lambda/4}=C\|u\|_X^{1-3\lambda/4}\|u\|_2^{3\lambda/4}.\end{aligned}$$

In the same way it follows that

$$\|S(z)(u-v)\|_4^2\le C\|u-v\|_X^{1/2}\|u-v\|_2^{3/2}.$$

Hence, by collecting all the terms,

$$\|F(u)-F(v)\|_2^2\le C\left(\|u\|_X^{4-3\lambda}\|u\|_2^{3\lambda}+\|v\|_X^{4-3\lambda}\|v\|_2^{3\lambda}\right)\|u-v\|_X^{1/2}\|u-v\|_2^{3/2}.$$

We still have the freedom to choose q. In order to satisfy $\alpha=4-3\lambda<7/2$ we need $\lambda>1/6$ which holds for $q>10$. Note that a choice $q=16$ would give $\alpha=3,\beta=1$.

($\mathcal{F}_3$): It is easy to observe that $F\in C^3(X,L^2)$ with $\delta F(0)[u_n]=0$, $\delta^2F(0)[u_n]^2=0$ and

$$\delta^3F(0)[u_n]^3=-6\int_0^1 S^{-1}(z)\left(\frac{1}{T(z)}|S(z)u_n|^2S(z)u_n\right)dz. \tag{5.1}$$

For a normalized eigenfunction ($\|u_n\|_2 = 1$) we have

$$(\delta^3 F(0)[u_n]^3, u_n)_{L^2} = -6 \int_0^1 \left(S^{-1}(z) \left(\frac{1}{T(z)} |S(z)u_n|^2 S(z)u_n \right), u_n \right)_{L^2} dz$$
$$= -6 \int_0^1 \frac{1}{T(z)} \|S(z)u_n\|_4^4 \, dz < 0.$$

($\mathcal{F}_4$): Defining $G(u) = (F(u), u)/4$ gives

$$G(u) = -\frac{1}{4} \int_0^1 \frac{1}{T(z)} \|S(z)u\|_4^4 dz$$

with $G'(u)v = (F(u), v)$.

($\mathcal{F}_5$): This assumption is fulfilled with $\delta = 3$.

($\mathcal{F}_6$): With $\Gamma u(x) = u(-x)$ we have to show $G(u) = G(\Gamma u)$. This is true since $S(z)$ and Γ commute which is shown as follows: Let $v(z) := \Gamma S(z)u$, then $v(0) = \Gamma u$ holds and with $Au = u_{xx} - x^2 u$ one can observe that

$$iv_z = i\Gamma \left(iR^{\text{eff}}(z) A S(z)u \right) = -R^{\text{eff}}(z) A \Gamma S(z)u,$$

which gives the assertion for Γ, and the same arguments apply for Γ_2.

Acknowledgments. The author thanks Tassilo Küpper for his support and discussions and C. K. R. T. Jones for valuable suggestions.

REFERENCES

[1] T. Cazenave, *An introduction to Nonlinear Schrödinger Equations*, Textos de Métods Matématicos, UFRJ, Rio de Janeiro, 26 (1993).

[2] B. Deconinck and J. N. Kutz, *Singular Instability of Exact Stationary Solutions of the Nonlocal Gross–Pitaevskii Equation*, e-print cond-mat/0208441.

[3] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, Vol. 3: Spectral Theory and Applications, Springer-Verlag, Berlin, 1990.

[4] R. Fukuizumi, *Stability and instability of standing waves for the nonlinear Schrödinger equation with harmonic potential*, Discrete Contin. Dynam. Systems, 7 (2001), pp. 525–544.

[5] J. J. Garcia-Ripoll, V. V. Konotop, B. Malomed, and V. M. Pérez-Garcia, *A quasi-local Gross–Pitaevskii equation for Bose–Einstein condensates*, Math. Comput. Simulation, 62 (2003), pp. 21–30.

[6] M. Hirose and M. Ohta, *Structure of positive radial solution to scalar field equations with harmonic potential*, J. Differential Equations, 178 (2002), pp. 519–540.

[7] R. K. Jackson, C. K. R. T. Jones, and V. Zharnitsky, *Dispersion-managed solitons via an averaged variational principle*, preprint.

[8] O. Kavian, *Introduction à la théorie des points critiques*, Math. Appl. 13, Springer-Verlag, Paris, 1993.

[9] O. Kavian and F. B. Weissler, *Self-similar solutions of the pseudo-conformally invariant nonlinear Schrödinger equation*, Michigan Math. J., 41 (1994), pp. 151–173.

[10] Y. S. Kivshar, T. J. Alexander, and S. K. Turitsyn, *Nonlinear modes of a macroscopic quantum oscillator*, Phys. Lett. A, 278 (2001), pp. 225–230.

[11] M. Kunze, *Infinitely many radial solutions of a variational problem related to dispersion-managed optical fibers*, Proc. Amer. Math. Soc., 131 (2003), pp. 2181–2188.

[12] M. Kunze, T. Küpper, V. K. Mezentsev, E. G. Shapiro, and S. K. Turitsyn, *Nonlinear solitary waves with Gaussian tails*, Phys. D, 128 (1999), pp. 273–295.

[13] M. Kurth, *Optical solitons as ground states of NLS in the regime of strong dispersion management*, Phys. D, 185 (2003), pp. 227–249.

[14] M. Kurth, *Verzweigung von DM-Solitonen in optischen Übertragungssystemen*, Ph.D. thesis, University of Cologne, Cologne, Germany, 2003.

[15] A. Maruta, T. Inoue, Y. Nonaka, and Y. Yoshika, *Bi-soliton solution in a dispersion managed system and its application to high speed and long haul communication*, in Optical Solitons—Theoretical and Experimental Challenges, K. Porsezian and V. C. Kuriakose, eds., Lecture Notes in Phys. 613, Springer-Verlag, Berlin, 2003, pp. 247–264.

[16] A. Maruta, Y. Nonaka, and T. Inoue, *Symmetric bi-soliton solution in a dispersion-managed system*, postdeadline poster PD4, Topical Meeting on Nonlinear Guided Waves and Their Applications, Clearwater, FL, 2001.

[17] Y.-G. Oh, *Cauchy problem and Ehrenfest's law of nonlinear Schrödinger equations with potentials*, J. Differ. Equations, 81 (1989), pp. 255–274.

[18] R. S. Palais, *The principle of symmetric criticality*, Commun. Math. Phys., 69 (1979), pp. 19–30.

[19] D. E. Pelinovsky and V. Zharnitsky, *Averaging of dispersion-managed solitons: Existence and stability*, SIAM J. Appl. Math., 63 (2003), pp. 745–776.

[20] D. E. Pelinovsky and J. Yang, *Parametric resonance and radiative decay of dispersion-managed solitons*, SIAM J. Appl. Math., 64 (2004), pp. 1360–1382.

[21] C. Paré and P. A. Bélanger, *Antisymmetric soliton in a dispersion-managed system*, Optics Communications, 168 (1999), pp. 103–109.

[22] P. H. Rabinowitz, *Some global results for nonlinear eigenvalue problems*, J. Functional Analysis, 7 (1971), pp. 487–513.

[23] H. A. Rose and M. I. Weinstein, *On the bound states of the nonlinear Schrödinger equation with a linear potential*, Phys. D, 30 (1988), pp. 207–218.

[24] T. Schäfer, V. K. Mezentsev, K. H. Spatschek, and S. K. Turitsyn, *The dispersion-managed soliton as a ground state of a macroscopic nonlinear quantum oscillator*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 457 (2001), pp. 273–282.

[25] S. K. Turitsyn, T. Schäfer, K. H. Spatschek, and V. K. Mezentsev, *Path-averaged chirped optical soliton in dispersion-managed fiber communication lines*, Optics Communications, 163 (1999), pp. 122–158.

[26] V. Zharnitsky, E. Grenier, S. K. Turitsyn, C. K. R. T. Jones, and J. S. Hesthaven, *Stabilizing effects of dispersion management*, Phys. D, 152/153 (2001), pp. 794–817.

[27] T.-S. Yang and W. L. Kath, *Radiation loss of dispersion-managed solitons in optical fibers*, Phys. D, 149 (2001), pp. 80–94.

[28] J. Zhang, *Stability of standing waves for nonlinear Schrödinger equations with unbounded potentials*, Z. Angew. Math. Phys., 51 (2000), pp. 498–503.

SIAM J. MATH. ANAL.
Vol. 36, No. 3, pp. 986–1013

ADIABATIC APPROXIMATION OF THE SCHRÖDINGER–POISSON SYSTEM WITH A PARTIAL CONFINEMENT*

NAOUFEL BEN ABDALLAH†, FLORIAN MÉHATS†, AND OLIVIER PINAUD†

Abstract. Asymptotic quantum transport models of a two-dimensional electron gas are presented. The starting point is a singular perturbation of the three-dimensional Schrödinger–Poisson system. The small parameter ε is the scaled width of the electron gas and appears as the lengthscale on which a one-dimensional confining potential varies. The rigorous $\epsilon \to 0$ limit is performed by projecting the three-dimensional wavefunction on the eigenfunctions corresponding to the confining potential. This leads to a two-dimensional Schrödinger–Poisson system with a modified Poisson equation keeping track of the third dimension. This limit model is proven to be a first-order approximation of the initial model. An intermediate model, called the "2.5D adiabatic model" is then introduced. It shares the same structure as the limit model but is shown to be a second-order approximation of the three-dimensional model.

Key words. adiabatic approximation, energy estimates, Strichartz estimates, error estimates, nonlinear analysis, two-dimensional electron gas

AMS subject classifications. 35B25, 35Q40, 82C22

DOI. 10.1137/S0036141003437915

1. Introduction.

Systems with reduced dimensionality are the basis of operation of most of nanoscale electronic devices. Among them is the two-dimensional electron gas (2DEG) [1, 2, 9], in which the electrons are strongly confined in one direction so that collisionless transport is allowed in the two remaining ones. Although the transport is quasi bidimensional, the Coulomb interaction results in a fully three-dimensional structure. Indeed, the particle density is a sheet density concentrated on the two-dimensional electron gas plane, which generates through mean field interaction a fully three-dimensional potential. In [17], an approximate Schrödinger–Poisson model taking into account the quasi-bidimensional nature of electron transport, while maintaining a three-dimensional description of the electrostatic potential, was proposed and numerically implemented in the stationary framework for electron waveguide structures. The model has been shown to be numerically in very good agreement with the fully three-dimensional Schrödinger–Poisson system, while having a much lower numerical complexity. The aim of this paper is to prove by a rigorous asymptotic analysis that the model introduced in [17] is a good approximation of the fully three-dimensional model and quantify the discrepancy between the two models. In order to simplify the setting and to avoid additional technicalities induced by stationarity and by boundary effects, we shall consider the time-dependent problem in the whole space. The case of stationary boundary value problems will be the subject of a forthcoming work by the third author of this paper [16].

Denoting by z the confined direction, we shall consider the following singularly

*Received by the editors December 2, 2003; accepted for publication (in revised form) May 7, 2004; published electronically January 5, 2005. This work was supported by the European IHP network Ref. HPRN-CT-2002-00282 entitled "Hyperbolic and Kinetic Equations: Asymptotics, Numerics, Analysis" and by the CNRS project "Transport dans les nanostructures" (Action Spécifique MATH-STIC).

http://www.siam.org/journals/sima/36-3/43791.html

†MIP, Laboratoire CNRS (UMR 5640), Université Paul Sabatier, 118, route de Narbonne, 31062 Toulouse Cedex 04, France (naoufel@mip.ups-tlse.fr, mehats@mip.ups-tlse.fr, pinaud@mip.ups-tlse.fr).

perturbed Schrödinger–Poisson system:

$$i\partial_t \psi^\varepsilon = -\frac{1}{2}\Delta_{x,z}\,\psi^\varepsilon + \frac{1}{\varepsilon^2}V_c\left(\frac{z}{\varepsilon}\right)\psi^\varepsilon + V^\varepsilon\psi^\varepsilon, \tag{1.1}$$

$$\psi^\varepsilon(0,x,z) = \psi_0^\varepsilon(x,z), \tag{1.2}$$

$$V^\varepsilon = \frac{1}{4\pi r} * \left(|\psi^\varepsilon|^2\right), \tag{1.3}$$

where $x \in \mathbb{R}^2$, $z \in \mathbb{R}$, $r = \sqrt{|x|^2 + z^2}$, the potential V^ε is the self-consistent potential due to space charge effects, and the external confinement potential $V_c^\varepsilon(z) = \frac{1}{\varepsilon^2}V_c\left(\frac{z}{\varepsilon}\right)$ is given. In this work, the asymptotic behavior of the solution of this nonlinear system is studied when ε goes to 0. Two approximate models are exhibited: the limit model (*2D surface density model*) and an intermediate ε-dependent model (*2.5D adiabatic model*), which is shown to be a more accurate approximation of the initial model.

Quantum systems confined on a surface have been studied previously in [8, 10, 15, 21]. Starting from a similar scaling on the transverse Hamiltonian, these authors consider the linear Schrödinger equation with a confinement on a general surface and derive an effective Hamiltonian which locally depends on the curvature properties of the surface. In our case, the effective Hamiltonian at the leading order is trivial since the surface is the plane $z = 0$. The main difficulty here stems from the nonlinear character of the problem due to the self-consistent potential.

As remarked in [21], quantum constrained systems can be linked to the Born–Oppenheimer approximation in molecular dynamics [12, 19, 21]. In order to analyze this link, let us rescale the variables z, t by setting $\tilde{z} = \frac{z}{\varepsilon}$, $\tilde{t} = \frac{t}{\varepsilon}$ and let $\tilde{x} = x$. To keep densities of order $\mathcal{O}(1)$, we also need to rescale ψ by a factor $\frac{1}{\sqrt{\varepsilon}}$; hence the self-consistent potential is rescaled by $\frac{1}{\varepsilon}$. Denoting again (with an abuse of notation) by ψ^ε and V^ε the functions of the new variables, the system takes the form

$$i\varepsilon\partial_{\tilde{t}}\psi^\varepsilon = -\frac{\varepsilon^2}{2}\Delta_{\tilde{x}}\psi^\varepsilon - \frac{1}{2}\partial_{\tilde{z}}^2\psi^\varepsilon + (V_c + \varepsilon V^\varepsilon)\psi^\varepsilon. \tag{1.4}$$

The above problem (in the linear case) has been studied in particular in [3, 19]. However, the problem (1.1)–(1.3) is not just a rescaling of the Born–Oppenheimer asymptotics for two reasons. The first reason is, again, the nonlinear character of this system, which might induce rapid time oscillations of V^ε. The second reason is the time scale. Indeed, if the asymptotics is done for times $\tilde{t}$ of order 1 for the Born–Oppenheimer problem (1.4), then t is of order ε in the initial problem (1.1)–(1.3). Therefore, since we are here interested in time intervals of order 1 for the variable t, working in the variable $\tilde{t}$ would necessitate longer time intervals (of the order of $1/\varepsilon$) which is more difficult. The two problems, however, share similar properties of adiabatic decoupling. The systems can be diagonalized by using the eigenspaces of the transverse Hamiltonian $-\frac{1}{2}\partial_z^2 + V$ (in which t and x are frozen). Within each eigenspace the dynamics is governed by an effective potential and is quantum in our case, wheras semiclassical behavior is expected in the Born–Oppenheimer approximation.

The paper is organized as follows. In section 2, we first make precise the properties of the confinement operator and define the two approximate models (namely, the two-dimensional and 2.5D models). Then we state the main results of this paper, namely Theorems 2.5, 2.6, and 2.7. Section 3 is devoted to the proof of ε-independent

estimates for (1.1)–(1.3). In section 4, we put both approximate models into a more general framework allowing us to prove existence and uniqueness of their solutions. The 2.5D adiabatic model is shown to be a second-order approximation in section 5, while in section 6 the 2D surface density model is proven to be only a first-order approximation. Finally, the appendices contains some basic results on the Schrödinger equation and the Poisson equation which are used throughout the paper.

Remark on the scaling. Before going further, and in order to make clear the physical assumptions made here, let us show how the system (1.1)–(1.3) can be obtained by a rescaling of the Schrödinger–Poisson system written in the physical dimensional variables. Let $\Psi(T,X,Z)$, $\mathcal{V}(T,X,Z)$ be the solution of

$$i\hbar\partial_T\Psi = -\frac{\hbar^2}{2m}\Delta_{X,Z}\,\Psi + (\mathcal{V}_c+\mathcal{V})\Psi, \tag{1.5}$$

$$\mathcal{V} = \frac{e^2}{4\pi\varepsilon_M}\,\frac{1}{\sqrt{|X|^2+Z^2}} * \left(|\Psi|^2\right), \tag{1.6}$$

where m is the effective mass, e is the elementary charge of the electrons, and ε_M is the electric permittivity of the material. We introduce two characteristic energies, E_c and E, which are, respectively, the typical energy of the confinement and the typical kinetic energy of the electrons. The assumption of a strong confinement is

$$\varepsilon^2 = \frac{E}{E_c} \ll 1. \tag{1.7}$$

The confinement operator is the partial Hamiltonian defined on $\mathbb{R}$ by $-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial Z^2}+\mathcal{V}_c$. Hence we deduce that the typical length L_c of the confinement, defined as the spatial extension of the eigenvalues of this operator, satisfies $\frac{\hbar^2}{2mL_c^2} = E_c$, and the confinement potential takes the form $\mathcal{V}_c(Z) = E_c V_c(\frac{Z}{L_c})$, where V_c denotes a dimensionless potential. Since we are interested in quantum models for the transport of the electrons, the typical space length L and the typical time $\mathcal{T}$ are deduced from the kinetic energy (this crucial assumption says that the initial data are not oscillating): $\frac{\hbar}{\mathcal{T}} = \frac{\hbar^2}{2mL^2} = E$; thus (1.7) gives $\frac{L_c}{L} = \varepsilon$. Finally, we assume that the self-consistent potential is of the same order of magnitude as the kinetic energy, which means that if N_0 is the typical density (the scale of $|\Psi|^2$), we have

$$\frac{e^2N_0L^2}{\varepsilon_M} = E.$$

With these assumptions, setting

$$t = \frac{T}{\mathcal{T}}, \qquad (x,z) = \left(\frac{X}{L},\frac{Z}{L}\right), \qquad \psi^\varepsilon = \frac{\Psi}{\sqrt{N_0}}, \qquad V^\varepsilon = \frac{\mathcal{V}}{E},$$

the system (1.5)–(1.6) is written (1.1)–(1.3) in the dimensionless variables.

2. Notation and main results. Throughout this paper, for any $q\in[1,\infty]$, we shall denote by q' its conjugate, and for any $q\in[2,\infty]$ we denote by q^* its 2-conjugate, respectively, defined by

$$q' = \frac{q}{q-1}; \qquad q^* = \frac{2q}{q-2}.$$

We define the following functional spaces.

DEFINITION 2.1. *Let $1 \le p, q, r \le +\infty$. The spaces $L^p_x L^q_z$ and $L^r_t L^p_x L^q_z$ are defined by*

$$L^p_x L^q_z(\mathbb{R}^3) = \left\{ u \in L^1_{loc}(\mathbb{R}^3), \quad \|u\|_{L^p_x L^q_z(\mathbb{R}^3)} = \left(\int_{\mathbb{R}^2} \|u(x,\cdot)\|^p_{L^q(\mathbb{R})}\, dx \right)^{1/p} < +\infty \right\}$$

(with an obvious generalization of this definition for $p = +\infty$),

$$L^r_t L^p_x L^q_z((0,T)\times\mathbb{R}^3) = L^r((0,T), L^p_x L^q_z(\mathbb{R}^3)).$$

When there is no ambiguity, we shall simply denote these spaces by $L^p_x L^q_z$ and $L^r_t L^p_x L^q_z$ and the corresponding norms by $\|\cdot\|_{p,q}$ and $\|\cdot\|_{r,p,q}$ (when there are two indices, the variables are (x,z); when there are three indices, the variables are (t,x,z)).

For a function $f = f(z)$ belonging to $L^1(\mathbb{R})$, we denote $\langle f \rangle = \int_{\mathbb{R}} f(z)dz$. In particular, if $n(t,x,z)$ is the particle density, the surface particle density is defined by $n_s(t,x) = \langle n(t,x,\cdot)\rangle = \int_{\mathbb{R}} n(t,x,z)\, dz$.

The symbol $*$ denotes a convolution with respect to all the variables $(x,z) \in \mathbb{R}^3$; partial convolutions are denoted by $*_x$ and $*_z$.

2.1. Properties of the confinement operator. Let us now introduce the basic assumptions made on the confining potential.

ASSUMPTION 2.2. (i) *The rescaled confining potential $V_c = V_c(z)$ is a nonnegative real-valued function in $L^2_{loc}(\mathbb{R})$.*

(ii) *The operator $A = -\frac{1}{2}\frac{d^2}{dz^2} + V_c$ defined on $L^2(\mathbb{R})$ with the domain*

$$\mathcal{D}(A) = \left\{ u \in H^2(\mathbb{R}) \text{ such that } V_c\, u \in L^2(\mathbb{R}) \right\}$$

admits a nondegenerate eigenvalue E associated to an eigenfunction $\chi(z)$ such that $z\chi \in L^2(\mathbb{R})$.

The first part of this assumption implies that the operator A is self-adjoint and nonnegative (see, e.g., [18]). The partial Hamiltonian involved in (1.1) is obtained by rescaling the operator A:

$$A^\varepsilon = -\frac{1}{2}\frac{d^2}{dz^2} + V_c^\varepsilon = -\frac{1}{2}\frac{d^2}{dz^2} + \frac{1}{\varepsilon^2} V_c\left(\frac{z}{\varepsilon}\right)$$

and we obtain an eigenfunction/eigenvalue pair of A^ε by setting

$$\chi^\varepsilon(z) = \frac{1}{\sqrt{\varepsilon}}\,\chi\left(\frac{z}{\varepsilon}\right); \quad E^\varepsilon = \frac{E}{\varepsilon^2}.$$

Note that the assumption on the eigenfunction given in Assumption 2.2 implies that

$$\forall \beta \in [0,1], \qquad \|z^\beta\, \chi^\varepsilon\|_{L^2(\mathbb{R})} = \mathcal{O}(\varepsilon^\beta). \tag{2.1}$$

We shall denote by $X^\varepsilon = \mathit{span}(\chi^\varepsilon)$ the corresponding eigenspace and by Π^ε the orthogonal projector on this eigenspace. Following the physical literature [1, 2], we shall refer to the *subband of energy level E^ε* as the space $L^2(\mathbb{R}^2, X^\varepsilon)$. With an abuse of notation, we shall also denote by Π^ε the orthogonal projector $\mathbb{I} \otimes \Pi^\varepsilon$ of $L^2(\mathbb{R}^3)$ on $L^2(\mathbb{R}^2, X^\varepsilon)$.

The following technical lemma will be used several times.

LEMMA 2.3. *Let $V^\varepsilon \in W^{1,\alpha}(\mathbb{R})$ with $\alpha \in [1,+\infty]$. Then there exists a constant $C>0$ such that*

$$\|[\Pi^\varepsilon, V^\varepsilon]\|_{\mathcal{L}(L^2(\mathbb{R}))} \leq C\,\varepsilon^{1-1/\alpha}\,\|\partial_z V^\varepsilon\|_{L^\alpha(\mathbb{R})},$$

where $[\cdot,\cdot]$ denotes the commutator between the two operators.

Proof. Noting that

$$[\Pi^\varepsilon, V^\varepsilon] = \Pi^\varepsilon\, V^\varepsilon\,(\mathbb{I}-\Pi^\varepsilon) - (\mathbb{I}-\Pi^\varepsilon)\,V^\varepsilon\,\Pi^\varepsilon$$

and that in this difference the second operator is the adjoint of the first one, one can see that the lemma stems from

$$\|\Pi^\varepsilon\, V^\varepsilon\,(\mathbb{I}-\Pi^\varepsilon)\|_{\mathcal{L}(L^2(\mathbb{R}))} \leq C\,\varepsilon^{1-1/\alpha}\,\|\partial_z V^\varepsilon\|_{L^\alpha(\mathbb{R})}.$$

In order to prove the above estimate, let $U^\varepsilon(z) = V^\varepsilon(z) - V^\varepsilon(0)$. By orthogonality of Π^ε and $\mathbb{I}-\Pi^\varepsilon$, we have, clearly,

$$\Pi^\varepsilon\, V^\varepsilon\,(\mathbb{I}-\Pi^\varepsilon) = \Pi^\varepsilon\, U^\varepsilon\,(\mathbb{I}-\Pi^\varepsilon).$$

Therefore

$$\|\Pi^\varepsilon\, V^\varepsilon\,(\mathbb{I}-\Pi^\varepsilon)\|_{\mathcal{L}(L^2(\mathbb{R}))} \leq \|\Pi^\varepsilon\, U^\varepsilon\|_{\mathcal{L}(L^2(\mathbb{R}))} \leq \|\chi^\varepsilon\, U^\varepsilon\|_{L^2(\mathbb{R})},$$

where a Cauchy–Schwarz inequality was used. Additionally, we have

$$|U^\varepsilon(z)| = \left|\int_0^z \partial_z V^\varepsilon(y)\,dy\right| \leq |z|^{1-1/\alpha}\,\|\partial_z V^\varepsilon\|_{L^\alpha(\mathbb{R})}.$$

Thus we conclude the proof, thanks to

$$\begin{aligned}\|\chi^\varepsilon\, U^\varepsilon\|^2_{L^2(\mathbb{R})} &\leq \|\partial_z V^\varepsilon\|^2_{L^\alpha(\mathbb{R})}\left\|z^{1-1/\alpha}\chi^\varepsilon\right\|^2_{L^2(\mathbb{R})}\\ &\leq C\,\varepsilon^{2-2/\alpha}\,\|\partial_z V^\varepsilon\|^2_{L^\alpha(\mathbb{R})},\end{aligned}$$

where we used (2.1). ∎

2.2. Definitions of the approximate models and main results. We shall assume that the initial wavefunction belongs to the subband of energy level E^ε.

ASSUMPTION 2.4 (well-prepared data). *The initial data ψ_0^ε of the three-dimensional Schrödinger–Poisson problem* (1.1)–(1.3) *satisfies*

$$\psi_0^\varepsilon = \phi_0\,\chi^\varepsilon \in H^1(\mathbb{R}^2, X^\varepsilon).$$

Let us now write the two approximate models for the three-dimensional Schrödinger–Poisson system (1.1)–(1.3).

The 2D surface density model. The 2D surface density model is obtained by coupling a two-dimensional Schrödinger equation and the Poisson equation with a modified Green function. It is given by

$$i\partial_t\phi = -\frac{1}{2}\Delta_x\,\phi + W\phi, \tag{2.2}$$

$$W = \frac{1}{4\pi|x|} *_x \left(|\phi|^2\right), \tag{2.3}$$

with the initial data $\phi(0,x)=\phi_0(x)=\langle\psi_0^\varepsilon(x,\cdot)\chi^\varepsilon\rangle$. The unknowns are $\phi(t,x)$, $W(t,x)$ and the surface density $n_s(t,x)=|\phi|^2(t,x)$, where $x\in\mathbb{R}^2$. Note that $W(t,x)=V(t,x,0)$, where V is the Coulomb potential generated by the sheet density supported in the plane $z=0$ with a surface density n_s:

$$n(t,x,z)=n_s(t,x)\delta(z);\quad V=\frac{1}{4\pi r}*n. \tag{2.4}$$

The 2.5D adiabatic model. The 2.5D adiabatic model is an intermediate model between the fully three-dimensional model and the 2D surface density model. It takes into account the small thickness of the electron gas and consists in coupling a two-dimensional Schrödinger equation and the three-dimensional Poisson equation. The unknowns are $\phi^\varepsilon(t,x)$, $V^\varepsilon(t,x,z)$ and the density $n^\varepsilon(t,x,z)$, where $x\in\mathbb{R}^2$ and $z\in\mathbb{R}$. This system is written

$$i\partial_t\phi^\varepsilon=-\frac{1}{2}\Delta_x\,\phi^\varepsilon+\left\langle V^\varepsilon|\chi^\varepsilon|^2\right\rangle\phi^\varepsilon, \tag{2.5}$$

$$V^\varepsilon=\frac{1}{4\pi r}*\left(|\phi^\varepsilon|^2\,|\chi^\varepsilon|^2\right), \tag{2.6}$$

with the initial data $\phi^\varepsilon(0,x)=\phi_0(x)=\langle\psi_0^\varepsilon(x,\cdot)\chi^\varepsilon\rangle$ and where the function $\chi^\varepsilon(z)$ is defined as in section 2.1. The population of electrons is described by a pure quantum state which belongs at any time to the subband of energy level E. One can see that in the 2.5D adiabatic model the dynamics on the subband is induced by the effective potential $\left\langle V^\varepsilon|\chi^\varepsilon|^2\right\rangle$, which is the potential "modulated" by the wavefunction χ^ε. Moreover, applying formally the standard perturbation theory (see [14]), the transverse Hamiltonian $-\frac{1}{2}\frac{d^2}{dz^2}+V_c^\varepsilon+V^\varepsilon$ admits an eigenvalue $\epsilon(t,x)$ given by

$$\epsilon=\frac{E}{\varepsilon^2}+\left\langle V^\varepsilon|\chi|^2\right\rangle+\mathcal{O}(\varepsilon^2).$$

Thus, the above 2.5D adiabatic model can be seen—at least formally—as an ε^2-perturbation of the model given by the adiabatic quantum theory [19] (the constant E/ε^2 can be forgotten in (2.5) since it only induces a phase factor).

The main results of the paper, summarized in the three following theorems, state that the 2.5D adiabatic model is (almost) a second-order approximation of the three-dimensional model, while the 2D surface density model is exactly a first-order approximation.

THEOREM 2.5. *Suppose that Assumptions* 2.2 *and* 2.4 *are satisfied. Then the three-dimensional Schrödinger–Poisson system* (1.1)–(1.3) *and the* 2.5*D adiabatic model* (2.5), (2.6) *admit unique global weak solutions denoted by* (ψ^{3D}, V^{3D}) *and* ($\phi^{2.5D}$, $V^{2.5D}$), *respectively. Moreover, for any* T *we have*

$$\|\psi^{3D}-\phi^{2.5D}\chi^\varepsilon\,e^{-itE/\varepsilon^2}\|_{q^*,q,2}=\mathcal{O}(\varepsilon)\qquad\forall q\in[2,\infty), \tag{2.7}$$

$$\|V^{3D}-V^{2.5D}\|_{L^1((0,T),L^\infty(\mathbb{R}^3))}=\mathcal{O}(\varepsilon^{2-\alpha})\qquad\forall\alpha>0. \tag{2.8}$$

Furthermore, the surface densities defined by $n_s^{3D}=\langle|\psi^{3D}|^2\rangle$ *and* $n_s^{2.5D}=|\phi^{2.5D}|^2$ *satisfy*

$$\|n_s^{3D}-n_s^{2.5D}\|_{L^1((0,T),L^q(\mathbb{R}^2))}=\mathcal{O}(\varepsilon^{2-\alpha})\qquad\forall\alpha>0,\qquad\forall q\in[1,\infty). \tag{2.9}$$

THEOREM 2.6. *Suppose that Assumptions* 2.2 *and* 2.4 *are satisfied. Then as* $\varepsilon \to 0$ *and for any* $T > 0$ *the solution* $(\phi^{2.5D},\, n^{2.5D},\, V^{2.5D})$ *of the* $2.5D$ *adiabatic model converges to the unique solution* $(\phi^{2D},\, n_s^{2D},\, V^{2D})$ *of the* $2D$ *surface density model* (2.2), (2.4) *in the following sense:*

$$\|\phi^{2.5D} - \phi^{2D}\|_{L^{q^*}((0,T),W^{1,q}(\mathbb{R}^2))} = \mathcal{O}(\varepsilon) \qquad \forall q \in [2,\infty), \tag{2.10}$$

$$\|V^{2.5D} - V^{2D}\|_{L^q((0,T),L^\infty(\mathbb{R}^3))} = \mathcal{O}(\varepsilon) \qquad \forall q \in [1,\infty), \tag{2.11}$$

$$\|n_s^{2.5D} - n_s^{2D}\|_{L^q((0,T),L^\infty(\mathbb{R}^2))} = \mathcal{O}(\varepsilon) \qquad \forall q \in [1,\infty). \tag{2.12}$$

THEOREM 2.7. *Suppose that Assumptions* 2.2 *and* 2.4 *are satisfied. If, moreover, we have*

$$0 < \|x\phi_0\|_{L^2(\mathbb{R}^2)} < +\infty \quad \textit{and} \quad \phi_0 \in H^2(\mathbb{R}^2), \tag{2.13}$$

then for any $T > 0$ *there exists a constant* $C > 0$ *such that the solutions of the* $2.5D$ *adiabatic model and the* $2D$ *surface density model satisfy*

$$\|(V^{2.5D} - V^{2D})(t,\cdot,0)\|_{L^\infty(\mathbb{R}^2)} + \|(n_s^{2.5D} - n_s^{2D})(t,\cdot)\|_{L^q(\mathbb{R}^2)} \geq C\,\varepsilon \tag{2.14}$$

for any $t \in [0,T]$, $q \in [1,\infty)$, *where* C *depends on* T *and* q *but not on* ε.

An immediate consequence of these theorems is the following.

COROLLARY 2.8. *Under Assumptions* 2.2 *and* 2.4, *the three-dimensional Schrödinger–Poisson system converges as* $\varepsilon \to 0$ *to the* $2D$ *surface density model. Moreover, if in addition* (2.13) *is satisfied, we have for any* $T > 0$ *and* $q \in [1,\infty)$,

$$C_1\,\varepsilon \leq \|V^{3D} - V^{2D}\|_{L^1((0,T),L^\infty(\mathbb{R}^3))} + \|n_s^{3D} - n_s^{2D}\|_{L^1((0,T),L^q(\mathbb{R}^2))} \leq C_2\,\varepsilon,$$

where the notations of Theorems 2.5 *and* 2.6 *were used.*

3. Estimates for the three-dimensional model. In this section we prove some ε-independent estimates for the three-dimensional Schrödinger-Poisson problem (1.1)–(1.3). We first claim that a straightforward adaptation of the proofs of [4, 13] allows us to show that for any initial data

$$\psi_0^\varepsilon \in \mathcal{H} := \{\phi \in H^1(\mathbb{R}^3) \,:\, \sqrt{V_c^\varepsilon}\,\psi \in L^2(\mathbb{R}^3)\}, \tag{3.1}$$

(which may depend on ε) and for an arbitrary $T > 0$, this system admits a unique weak solution ψ^ε, V^ε, such that

$$\psi^\varepsilon \in C([0,T],\mathcal{H}),$$

$$V^\varepsilon \in L^\infty((0,T)\times\mathbb{R}^3); \qquad \nabla_{x,z} V^\varepsilon \in L^\infty((0,T),L^q(\mathbb{R}^3)) \quad \forall q \in (3/2,\infty).$$

Let us define the kinetic energy along the x direction and along the z direction, respectively, by

$$\mathcal{E}^\varepsilon_{kin,x}(t) = \iint_{\mathbb{R}^3} \frac{1}{2}\,|\nabla_x \psi^\varepsilon(t,x,z)|^2\,dx\,dz; \quad \mathcal{E}^\varepsilon_{kin,z}(t) = \iint_{\mathbb{R}^3} \frac{1}{2}\,|\partial_z \psi^\varepsilon(t,x,z)|^2\,dx\,dz.$$

The self-consistent potential energy and the external potential energy are then, respectively, defined by

$$\mathcal{E}^\varepsilon_{pot}(t) = \iint_{\mathbb{R}^3} \frac{1}{2} |\nabla_{x,z} V^\varepsilon|^2 \, dx \, dz; \quad \mathcal{E}^\varepsilon_{ext}(t) = \iint_{\mathbb{R}^3} V^\varepsilon_c(z) |\psi^\varepsilon(t,x,z)|^2 \, dx \, dz$$

and the total energy of the system is

$$\mathcal{E}^\varepsilon_{tot}(t) = \mathcal{E}^\varepsilon_{kin,x}(t) + \mathcal{E}^\varepsilon_{kin,z}(t) + \mathcal{E}^\varepsilon_{pot}(t) + \mathcal{E}^\varepsilon_{ext}(t).$$

The standard energy estimate for the Schrödinger–Poisson system [4] gives the conservation of the total energy:

$$\forall t \geq 0 \qquad \mathcal{E}^\varepsilon_{tot}(t) = \mathcal{E}^\varepsilon_{tot}(0). \tag{3.2}$$

Unfortunately, due to the strong confinement potential V^ε_c, the external energy $\mathcal{E}^\varepsilon_{ext}$ is of order $\mathcal{O}(1/\varepsilon^2)$. Therefore, (3.2) does not provide directly a bound for the kinetic energy (except for the special case where the initial data is concentrated on the ground state). Nevertheless the Strichartz estimates of Appendix A enable us to obtain some estimates independent of ε, without using the energy conservation. The first step is the following lemma.

LEMMA 3.1. *Let $\psi^\varepsilon_0 \in L^2(\mathbb{R}^3)$ and let ψ^ε, V^ε be a solution of* (1.1)–(1.3). *If Assumption* 2.2(i) *is satisfied, then for any $T > 0$ we have*

$$\forall q \in [2, \infty) \qquad \|\psi^\varepsilon\|_{q^*,q,2} \leq C(\psi_0), \tag{3.3}$$

$$\forall q \in [1, 3) \qquad \|V^\varepsilon\|_{L^q((0,T),L^\infty(\mathbb{R}^3))} \leq C(\psi_0), \tag{3.4}$$

where $C(\psi_0)$ denotes a generic constant, which depends only on $\|\psi^\varepsilon_0\|_{L^2(\mathbb{R}^3)}$ (and q), and $q^* = 2q/(q-2)$.

Proof. This proof relies on the Strichartz estimates and on the properties of the Poisson equation studied in Appendices A and B. Let us first recall that the L^2 estimate for the Schrödinger equation gives

$$\forall t \in [0, T] \qquad \|\psi(t)\|_{2,2} \leq \|\psi^\varepsilon_0\|_{L^2(\mathbb{R}^3)}.$$

Additionally, from (B.3) and a Hölder inequality, we deduce that

$$\forall q \in (2, \infty) \qquad \left\| \frac{1}{r} * (fg) \right\|_{q,\infty} \leq C \, \|f\|_{q,2} \, \|g\|_{2,2};$$

thus for all $t \in (0, T)$ we have

$$\forall q \in (2, \infty) \qquad \|V^\varepsilon(t)\|_{q,\infty} \leq C(\psi_0) \, \|\psi^\varepsilon(t)\|_{q,2} \, .$$

Hence

$$\|V^\varepsilon(t) \, \psi^\varepsilon(t)\|_{2,2} \leq \|V^\varepsilon(t)\|_{q,\infty} \, \|\psi^\varepsilon(t)\|_{q^*,2} \leq C(\psi_0) \, \|\psi^\varepsilon(t)\|_{q,2} \, \|\psi^\varepsilon(t)\|_{q^*,2} \, .$$

Let q be fixed such that $q \in [4, \infty)$. It is readily seen that

$$\|\psi^\varepsilon\|_{q^*,2} \leq \|\psi^\varepsilon\|^{2/(q-2)}_{q,2} \, \|\psi^\varepsilon\|^{(q-4)/(q-2)}_{2,2} \, ,$$

which leads to

$$\|V^\varepsilon(t)\,\psi^\varepsilon(t)\|_{2,2} \le C(\psi_0)\,\|\psi^\varepsilon(t)\|_{q,2}^{q^*/2}. \tag{3.5}$$

For any $t \ge 0$, let

$$Y(t) := \|\psi^\varepsilon\|_{L^{q^*}((0,t),L^q_x L^2_z)}.$$

By using (3.5) and a Hölder inequality, we get

$$\|V^\varepsilon\,\psi^\varepsilon\|_{L^1((0,t),L^2(\mathbb{R}^3))} \le C(\psi_0)\,\sqrt{t}\,(Y(t))^{q^*/2}.$$

Consequently the Strichartz inequality stated in Lemma A.2 gives

$$Y(t) \le C(\psi_0)\left(1 + \sqrt{t}\,(Y(t))^{q^*/2}\right).$$

Since $Y(0) = 0$, this is enough to conclude by continuity that there exists $\widetilde{T}$ and C_0 depending only on $\|\psi_0^\varepsilon\|_{L^2(\mathbb{R}^3)}$ and q such that $Y(\widetilde{T}) \le C_0$. We deduce (3.3) for $q \ge 4$ by iterating this procedure on the interval $(\widetilde{T}, 2\widetilde{T})$, then on $(2\widetilde{T}, 3\widetilde{T})$, etc. By interpolation, we also deduce that (3.3) holds true for $q \in (2,4)$. To obtain (3.4), it is enough to apply (B.5) with p close to 2 and to use (3.3) with q close to 4. ∎

From this lemma, one can deduce the main result of this section as follows.

PROPOSITION 3.2. *Assume that the initial data $\psi_0^\varepsilon \in \mathcal{H}$ (defined by* (3.1)*) satisfies*

$$\|\psi_0^\varepsilon\|_{L^2(\mathbb{R}^3)} + \|\nabla_x \psi_0^\varepsilon\|_{L^2(\mathbb{R}^3)} \le C \tag{3.6}$$

and let ψ^ε, V^ε be the solution of (1.1)–(1.3)*. Then, if Assumption* 2.2(i) *is satisfied, we have the following estimates:*

$$\forall q \in [2,\infty) \qquad \|\psi^\varepsilon\|_{q^*,q,2} + \|\nabla_x \psi^\varepsilon\|_{q^*,q,2} \le C, \tag{3.7}$$

$$\|V^\varepsilon\|_{L^\infty((0,T)\times\mathbb{R}^3)} \le C, \tag{3.8}$$

$$\forall q \in (2,\infty) \qquad \|\nabla_{x,z} V^\varepsilon\|_{\infty,q,\infty} \le C. \tag{3.9}$$

Here C denotes a generic constant independent of ε.

Proof. We first remark is that, thanks to (3.6), the estimates (3.3) and (3.4) given in the previous lemma are independent of ε. Differentiating (1.1) with respect to x leads to

$$i\partial_t \nabla_x \psi^\varepsilon = -\frac{1}{2}\Delta_x\,\nabla_x \psi^\varepsilon + A^\varepsilon \nabla_x \psi^\varepsilon + V^\varepsilon \nabla_x \psi^\varepsilon + \nabla_x V^\varepsilon \psi^\varepsilon. \tag{3.10}$$

From (B.4), we deduce that for all $t \in (0,T)$ we have

$$\begin{aligned} \forall q \in (2,\infty) \qquad \|\nabla_x V^\varepsilon(t)\|_{q,\infty} &\le C\,\|\nabla_x(|\psi^\varepsilon(t)|^2)\|_{2q/(2+q),1} \\ &\le C\,\|\nabla_x \psi^\varepsilon(t)\|_{L^2(\mathbb{R}^3)}\,\|\psi^\varepsilon(t)\|_{q,2}\,. \end{aligned}$$

Hence we get, for any $q \in (2,\infty)$,

$$\|\nabla_x V^\varepsilon\|_{q',q,\infty} \le C\,\|\nabla_x \psi^\varepsilon\|_{2,2,2}\,\|\psi^\varepsilon\|_{q^*,q,2}\,, \tag{3.11}$$

since $\frac{1}{q^*}+\frac{1}{2}=\frac{1}{q'}$. Therefore we have

$$\|\nabla_x V^\varepsilon\psi^\varepsilon\|_{1,2,2} \le \|\nabla_x V^\varepsilon\|_{q',q,\infty}\,\|\psi^\varepsilon\|_{q,q^*,2} \le C\,\|\nabla_x\psi^\varepsilon\|_{2,2,2}\,\|\psi^\varepsilon\|_{q^*,q,2}\,\|\psi^\varepsilon\|_{q,q^*,2}\,.$$

Since for any $q\in(2,\infty)$ we have $(q^*)^*=q$, by using (3.3) we obtain

$$\|\nabla_x V^\varepsilon\psi^\varepsilon\|_{1,2,2} \le C\|\nabla_x\psi^\varepsilon\|_{2,2,2}\,.$$

This inequality, combined with the L^2 estimate for (3.10), gives

$$\begin{aligned}\|\nabla_x\psi^\varepsilon\|_{\infty,2,2} &\le \|\nabla_x\psi_0^\varepsilon\|_{L^2(\mathbb{R}^3)} + \|\nabla_x V^\varepsilon\psi^\varepsilon\|_{1,2,2}\\ &\le \|\nabla_x\psi_0^\varepsilon\|_{L^2(\mathbb{R}^3)} + C\,\|\nabla_x\psi^\varepsilon\|_{2,2,2}\,,\end{aligned}$$

which leads, thanks to a Gronwall argument, to

$$\|\nabla_x\psi^\varepsilon\|_{\infty,2,2} + \|\nabla_x V^\varepsilon\psi^\varepsilon\|_{1,2,2} \le C\,. \tag{3.12}$$

In a second step, we apply the Strichartz estimate (A.5) to (3.10) and obtain

$$\forall q\in[2,\infty)\qquad \|\nabla_x\psi^\varepsilon\|_{q^*,q,2} \le C\|\nabla_x\psi_0^\varepsilon\|_{L^2(\mathbb{R}^3)} + C\|V^\varepsilon\nabla_x\psi^\varepsilon\|_{1,2,2} + C\|\nabla_x V^\varepsilon\psi^\varepsilon\|_{1,2,2}\,.$$

Since (3.4) implies

$$\|V^\varepsilon\nabla_x\psi^\varepsilon\|_{1,2,2} \le \|V^\varepsilon\|_{1,\infty,\infty}\,\|\nabla_x\psi^\varepsilon\|_{\infty,2,2} \le C\|\nabla_x\psi^\varepsilon\|_{\infty,2,2}\,,$$

we deduce the estimate (3.7) from (3.6) and (3.12).

For the last step of this proof, we apply a Sobolev estimate pointwise in time to the function

$$u(t,x) = \|\psi^\varepsilon(t,x,\cdot)\|_{L^2(\mathbb{R})}\,.$$

To this aim, by using the Cauchy–Schwarz inequality, we first get

$$|\nabla_x u(t,x)| \le \left(\int |\nabla_x\psi^\varepsilon(t,x,z)|^2\,dz\right)^{1/2},$$

which yields

$$\|u(t,\cdot)\|_{H^1(\mathbb{R}^2)} \le C\|\psi^\varepsilon\|_{\infty,2,2} + C\|\nabla_x\psi^\varepsilon\|_{\infty,2,2} \le C$$

(apply (3.7) with $q=2$ for the last inequality). By Sobolev embeddings, we have

$$\forall p\in[2,\infty)\qquad \|\psi^\varepsilon\|_{\infty,p,2} = \|u\|_{L^\infty_t L^p_x} \le C, \tag{3.13}$$

which can be rewritten

$$\forall q\in[1,\infty)\qquad \left\||\psi^\varepsilon|^2\right\|_{\infty,q,1} \le C\,.$$

From (B.5) we deduce the $L^\infty((0,T)\times\mathbb{R}^3)$ estimate (3.8). Finally, by combining (3.13) and (3.7), we deduce that

$$\forall q\in(1,2)\qquad \left\|\nabla_x\left(|\psi^\varepsilon|^2\right)\right\|_{\infty,q,1} \le \|\psi^\varepsilon\|_{\infty,2q/(2-q),2}\,\|\nabla_x\psi^\varepsilon\|_{\infty,2,2} \le C,$$

and (3.9) is obtained by applying (B.4). ∎

We end this section with a useful lemma concerning the linear Schrödinger equation with a strong confining potential. It states that, up to—at least—the first order in ε, the subspace X^ε is stable under the action of the Schrödinger group.

LEMMA 3.3. *Let $\psi_0^\varepsilon \in L^2(\mathbb{R}^2, X^\varepsilon)$. Assume that $V^\varepsilon \in L^1((0,T), L^\infty(\mathbb{R}^3))$ and that $\partial_z V^\varepsilon \in L^{r',r,\infty}((0,T)\times(\mathbb{R}^3))$ for some $r \in (2,\infty]$. Then any solution ψ^ε of (1.1) satisfies, for all $s \in [2,\infty)$,*

$$\|(\mathbb{I}-\Pi^\varepsilon)\psi^\varepsilon\|_{s^*,s,2} \le C\,\varepsilon\,\|\partial_z V^\varepsilon\|_{r',r,\infty}\,\|\psi_0^\varepsilon\|_{L^2(\mathbb{R}^3)}\,,$$

where C depends only on $\|V^\varepsilon\|_{1,\infty,\infty}$.

Proof. Thanks to the conservation of the L^2 norm for the Schrödinger equation, a solution ψ^ε of (1.1) satisfies

$$\|\psi^\varepsilon\|_{\infty,2,2} \le \|\psi_0^\varepsilon\|_{L^2(\mathbb{R}^3)}.$$

By using (A.5), we get for any $q \in [2,\infty)$

$$\|\psi^\varepsilon\|_{q^*,q,2} \le C\|\psi_0^\varepsilon\|_{L^2(\mathbb{R}^3)} + C\|V^\varepsilon\,\psi^\varepsilon\|_{1,2,2} \le C\|\psi_0^\varepsilon\|_{L^2(\mathbb{R}^3)} \tag{3.14}$$

(in this lemma, C is a generic constant depending only on $\|V^\varepsilon\|_{1,\infty,\infty}$).

Denote $\omega^\varepsilon = (\mathbb{I}-\Pi^\varepsilon)\psi^\varepsilon$. The assumption on ψ_0^ε implies $\omega^\varepsilon(0,x,z) = 0$ for $(x,z)\in\mathbb{R}^3$. Additionally, the operator $\mathbb{I}-\Pi^\varepsilon$ commutes with ∂_t, with Δ_x, and with A^ε (since Π^ε is a spectral projector of A^ε). Hence (1.1) gives, after direct calculations,

$$\begin{cases} i\partial_t\omega^\varepsilon = -\dfrac{1}{2}\Delta_x\omega^\varepsilon + A^\varepsilon\omega^\varepsilon + V^\varepsilon\omega^\varepsilon - [\Pi^\varepsilon, V^\varepsilon]\psi^\varepsilon, \\ \omega^\varepsilon(0,x,z) = 0. \end{cases} \tag{3.15}$$

Because of source terms, the L^2 conservation becomes

$$\|\omega^\varepsilon\|_{\infty,2,2} \le C\,\|[\Pi^\varepsilon, V^\varepsilon]\psi^\varepsilon\|_{1,2,2};$$

thus from (A.5) with $\sigma \in [2,\infty)$ we deduce

$$\|\omega^\varepsilon\|_{\sigma^*,\sigma,2} \le C\,\|[\Pi^\varepsilon, V^\varepsilon]\psi^\varepsilon\|_{1,2,2}\,. \tag{3.16}$$

Additionally, Lemma 2.3 yields

$$\|[\Pi^\varepsilon, V^\varepsilon]\psi^\varepsilon(t,x,\cdot)\|_{L^2(\mathbb{R})} \le C\,\varepsilon\,\|\partial_z V^\varepsilon(t,x,\cdot)\|_{L^\infty(\mathbb{R})}\,\|\psi^\varepsilon(t,x,\cdot)\|_{L^2(\mathbb{R})}.$$

Hence

$$\|[\Pi^\varepsilon, V^\varepsilon]\psi^\varepsilon\|_{1,2,2} \le C\,\varepsilon\,\|\partial_z V^\varepsilon\|_{r',r,\infty}\,\|\psi^\varepsilon\|_{r,r^*,2}.$$

An application of (3.14) with $q = r^*$ gives

$$\|[\Pi^\varepsilon, V^\varepsilon]\psi^\varepsilon\|_{1,2,2} \le C\,\varepsilon\,\|\partial_z V^\varepsilon\|_{r',r,\infty}\,\|\psi_0^\varepsilon\|_{L^2(\mathbb{R}^3)}\,.$$

Therefore we deduce the result from this estimate and (3.16). ∎

4. Existence results for the approximate models. In this section we show that the two approximate models (2.5), (2.6) and (2.2), (2.3) presented in section 2 are well-posed. Let us first remark that the 2.5D adiabatic model can be rewritten as a two-dimensional Schrödinger–Poisson system with a modified Green function. Indeed, denoting $W^\varepsilon(x) = \langle V^\varepsilon | \chi^\varepsilon|^2 \rangle$, (2.5), (2.6) is equivalent to

$$i\partial_t \phi^\varepsilon = -\frac{1}{2}\Delta_x \phi^\varepsilon + W^\varepsilon \phi^\varepsilon, \tag{4.1}$$

$$W^\varepsilon(x) = G^{2.5D} *_x \left(|\phi^\varepsilon|^2\right), \tag{4.2}$$

where

$$G^{2.5D}(x) = \int_{\mathbb{R}}\int_{\mathbb{R}} \frac{1}{4\pi\left(|x|^2 + (z-z')^2\right)^{1/2}} |\chi^\varepsilon(z')|^2 |\chi^\varepsilon(z)|^2 \, dz' \, dz. \tag{4.3}$$

With this formulation, both approximate systems have the same structure; they differ by the kernel of the "Poisson" equation, respectively, $G^{2.5D}(x)$ for (4.1), (4.2) and $G^{2D}(x) = \frac{1}{4\pi|x|}$ for (2.2), (2.3). We shall see below that these kernels share the same properties and that their difference is small (see the proof of Theorem 2.6 in section 4.2).

4.1. A Schrödinger–Poisson system with a general kernel. Let $G^\varepsilon(x)$ be a general convolution kernel such that $G^\varepsilon \in L^1_{loc}(\mathbb{R}^2)$. Consider the system

$$i\partial_t \phi^\varepsilon = -\frac{1}{2}\Delta_x \phi^\varepsilon + W^\varepsilon \phi^\varepsilon, \tag{4.4}$$

$$W^\varepsilon = G^\varepsilon * |\phi^\varepsilon|^2, \tag{4.5}$$

with the initial data $\phi^\varepsilon(0,\cdot) = \phi_0$. In this problem, the dependency of the functions in ε comes from the dependency of G^ε in this parameter. The energy of this system has two terms: the kinetic energy along x and the potential energy, respectively, defined by

$$\mathcal{E}^\varepsilon_{kin}(t) = \frac{1}{2}\int_{\mathbb{R}^2} |\nabla_x \phi^\varepsilon(t,x)|^2 \, dx,$$

$$\mathcal{E}^\varepsilon_{pot}(t) = \frac{1}{2}\int_{\mathbb{R}^2} W^\varepsilon \, n^\varepsilon_s \, dx = \frac{1}{2}\iint_{\mathbb{R}^4} G^\varepsilon(x-x') n^\varepsilon_s(x) n^\varepsilon_s(x') \, dx dx'.$$

By analogy with the function $\frac{1}{|x|}$ (see Lemma B.1), we assume that the kernel G^ε satisfies the following property.

ASSUMPTION 4.1. *The kernel G^ε is a nonnegative, even function which belongs to $L^1_{loc}(\mathbb{R}^2)$. Moreover, we assume the following estimates:*

(i) *For $f \in L^q(\mathbb{R}^2)$ with $1 < q < 2$, we have*

$$\|G^\varepsilon * f\|_{L^{q^\#}(\mathbb{R}^2)} \le C \|f\|_{L^q(\mathbb{R}^2)}, \tag{4.6}$$

where $q^\# = \frac{2q}{2-q}$.

(ii) *For $f \in L^q(\mathbb{R}^2) \cap L^1(\mathbb{R}^2)$ with $2 < q \le +\infty$, the following estimate holds:*

$$\|G^\varepsilon * f\|_{L^\infty(\mathbb{R}^2)} \le C \|f\|^\theta_{L^q(\mathbb{R}^2)} \|f\|^{1-\theta}_{L^1(\mathbb{R}^2)}, \tag{4.7}$$

where $\theta = \frac{q}{2q-2}$. *The constants* C *are assumed independent of* ε *and* f.

Remark. Any kernel of the type $G^\varepsilon(x) = g^\varepsilon(|x|)$, with $g^\varepsilon(|x|)$ satisfying $g^\varepsilon(t) < C/t$, verifies Assumption 4.1.

The following proposition shows that system (4.4), (4.5) is well-posed and gives some ε-independent estimates.

PROPOSITION 4.2. *Under Assumption* 4.1 *and for* $\phi_0 \in H^1(\mathbb{R}^2)$, *system* (4.4), (4.5) *admits a unique global weak solution. Moreover, the total energy of the system is conserved:*

$$\mathcal{E}^\varepsilon_{kin}(t) + \mathcal{E}^\varepsilon_{pot}(t) = \mathcal{E}^\varepsilon_{kin}(0) + \mathcal{E}^\varepsilon_{pot}(0), \tag{4.8}$$

and for any $T > 0$ *the following estimates hold independently of* ε:

$$\|\phi^\varepsilon\|_{L^{q^*}((0,T),W^{1,q}(\mathbb{R}^2))} \le C \qquad \forall q \in [2,\infty), \tag{4.9}$$

$$\|W^\varepsilon\|_{L^\infty((0,T),W^{1,q}(\mathbb{R}^2))} \le C \qquad \forall q \in (2,\infty). \tag{4.10}$$

Proof. The local-in-time existence of a unique weak solution is obtained via a standard fixed point procedure and is only sketched here. For more details we refer to [4, 13]. Denoting $W^\varepsilon(\psi) = G^\varepsilon * |\psi|^2$, it is enough to show that the application $\mathcal{F} : \psi \mapsto W^\varepsilon(\psi)\psi$ is locally Lipschitz in $H^1(\mathbb{R}^2)$ uniformly in time. To this aim, we shall use the following inequalities obtained by simple arguments such as Sobolev embeddings and Cauchy–Schwarz inequalities:

(4.11)
$$\|fg\|_{H^1(\mathbb{R}^2)} \le C\|f\|_{W^{1,4}(\mathbb{R}^2)}\|g\|_{H^1(\mathbb{R}^2)}; \quad \|fg\|_{W^{1,4/3}(\mathbb{R}^2)} \le C\|f\|_{H^1(\mathbb{R}^2)}\|g\|_{H^1(\mathbb{R}^2)}.$$

Let Φ and Ψ be two functions in $H^1(\mathbb{R}^2)$. We have

$$\|\mathcal{F}(\Psi) - \mathcal{F}(\Phi)\|_{H^1(\mathbb{R}^2)} \le \|W^\varepsilon(\Psi)\,(\Psi - \Phi)\,\|_{H^1(\mathbb{R}^2)} + \|\,(W^\varepsilon(\Psi) - W^\varepsilon(\Phi))\,\Phi\|_{H^1(\mathbb{R}^2)}.$$

Using the first inequality of (4.11), the right-hand side is controlled by

$$\|W^\varepsilon(\Psi)\|_{W^{1,4}(\mathbb{R}^2)}\|\Psi - \Phi\|_{H^1(\mathbb{R}^2)} + \|W^\varepsilon(\Psi) - W^\varepsilon(\Phi)\|_{W^{1,4}(\mathbb{R}^2)}\|\Phi\|_{H^1(\mathbb{R}^2)}.$$

Additionally,

$$\begin{aligned}\|W^\varepsilon(\Psi) - W^\varepsilon(\Phi)\|_{W^{1,4}(\mathbb{R}^2)} &\le \left\|G^\varepsilon * \left(|\Psi|^2 - |\Phi|^2\right)\right\|_{W^{1,4}(\mathbb{R}^2)} \\ &\le C\left\||\Psi|^2 - |\Phi|^2|\right\|_{W^{1,4/3}(\mathbb{R}^2)} \\ &\le C\|\Psi - \Phi\|_{H^1(\mathbb{R}^2)}\|\Psi + \Phi\|_{H^1(\mathbb{R}^2)},\end{aligned}$$

where (4.6) is used as well as the second inequality of (4.11). By noticing that $W^\varepsilon(0) = 0$, we conclude that

$$\begin{aligned}\|\mathcal{F}(\Psi) - \mathcal{F}(\Phi)\|_{H^1(\mathbb{R}^2)} \le{}& C\|\Psi\|^2_{H^1(\mathbb{R}^2)}\|\Psi - \Phi\|_{H^1(\mathbb{R}^2)} \\ &+ C\|\Psi - \Phi\|_{H^1(\mathbb{R}^2)}\|\Psi + \Phi\|_{H^1(\mathbb{R}^2)}\|\Phi\|_{H^1(\mathbb{R}^2)},\end{aligned}$$

which proves that $\mathcal{F}$ is locally Lipschitz on $H^1(\mathbb{R}^2)$.

The energy estimate (4.8) shows that the solution is global in time. It can be obtained in a standard manner by multiplying (4.4) by $\partial_t\overline{\phi^\varepsilon}$, integrating on $\mathbb{R}^2$, and

taking the real part. The key point is that the nonlinear term can be written as follows:

$$\begin{aligned}\mathcal{R}e\int_{\mathbb{R}^2} W^\varepsilon\phi^\varepsilon\,\partial_t\overline{\phi}^\varepsilon\,dx &= \int_{\mathbb{R}^2} G^\varepsilon * |\phi^\varepsilon|^2(x)\partial_t|\phi^\varepsilon(x)|^2\,dx\\ &= \frac{1}{4}\frac{d}{dt}\iint_{\mathbb{R}^4} G^\varepsilon(x-x')\,|\phi^\varepsilon(x')|^2\,|\phi^\varepsilon(x)|^2\,dx = \frac{1}{2}\frac{d}{dt}\mathcal{E}^\varepsilon_{pot}(t),\end{aligned}$$

where we have symmetrized the formula by using the properties of G^ε. The proof of (4.9) and (4.10) can be done without any difficulty by an adaptation of Lemma 3.1 and Proposition 3.2. The starting point is the $L^\infty((0,T),H^1(\mathbb{R}^2))$ bound of ϕ^ε given by the energy estimate and the conservation of charge density. Then we use successively Assumption 4.1 and standard Strichartz estimates in dimension 2 (see, for instance, [7]). □

The following proposition shows the Lipschitz dependency of the solution of (4.4), (4.5) with respect to the kernel.

PROPOSITION 4.3. *Let G^ε and $\widetilde{G^\varepsilon}$ satisfy Assumption* 4.1 *such that $G^\varepsilon - \widetilde{G^\varepsilon} \in L^1(\mathbb{R}^2)$. Let $\phi_0 \in H^1(\mathbb{R}^2)$ and denote by $(\phi^\varepsilon, W^\varepsilon)$ and $(\widetilde{\phi^\varepsilon}, \widetilde{W^\varepsilon})$, respectively, the solutions of* (4.4), (4.5) *corresponding to these kernels. Then we have*

$$\|\phi^\varepsilon - \widetilde{\phi^\varepsilon}\|_{L^{q^*}((0,T),W^{1,q}(\mathbb{R}^2))} \le C\|G^\varepsilon - \widetilde{G^\varepsilon}\|_{L^1(\mathbb{R}^2)} \qquad \forall q\in[2,\infty), \tag{4.12}$$

$$\|W^\varepsilon - \widetilde{W^\varepsilon}\|_{L^q((0,T),L^\infty(\mathbb{R}^2))} \le C\|G^\varepsilon - \widetilde{G^\varepsilon}\|_{L^1(\mathbb{R}^2)} \qquad \forall q\in[1,\infty), \tag{4.13}$$

where C is independent of ε.

Proof. Let us denote $\eta = \|G^\varepsilon - \widetilde{G^\varepsilon}\|_{L^1(\mathbb{R}^2)}$. For any function $f \in L^p(\mathbb{R}^2)$, $p \in [1,+\infty]$, we have

$$\left\|(G^\varepsilon - \widetilde{G^\varepsilon}) * f\right\|_{L^p(\mathbb{R}^2)} \le \eta\,\|f\|_{L^p(\mathbb{R}^2)}. \tag{4.14}$$

Setting

$$R^\varepsilon(x) = (G^\varepsilon - \widetilde{G^\varepsilon}) * |\widetilde{\phi}^\varepsilon|^2,$$

we have

$$W^\varepsilon - \widetilde{W}^\varepsilon = G^\varepsilon * \left(|\phi^\varepsilon|^2 - |\widetilde{\phi}^\varepsilon|^2\right) + R^\varepsilon. \tag{4.15}$$

By applying (4.9) and the Sobolev embeddings $W^{1,2}(\mathbb{R}^2) \hookrightarrow L^q(\mathbb{R}^2)$ for all $q \in [2,+\infty)$, and $W^{1,p}(\mathbb{R}^2) \hookrightarrow L^\infty(\mathbb{R}^2)$ for all $p > 2$, we have

$$\|\widetilde{\phi}^\varepsilon\|_{L^\infty((0,T),L^q(\mathbb{R}^2))} + \|\widetilde{\phi}^\varepsilon\|_{L^q((0,T),L^\infty(\mathbb{R}^2))} \le C \qquad \forall q\in[2,\infty). \tag{4.16}$$

Therefore (4.14) yields, for any $q \in [2,\infty)$,

$$\|R^\varepsilon\|_{L^\infty((0,T),L^q(\mathbb{R}^2))} + \|R^\varepsilon\|_{L^q((0,T),L^\infty(\mathbb{R}^2))} \le C\eta. \tag{4.17}$$

In order to estimate the difference $W^\varepsilon - \widetilde{W}^\varepsilon$, we set $u^\varepsilon := \phi^\varepsilon - \widetilde{\phi}^\varepsilon$. This function solves

$$\begin{cases} i\partial_t u^\varepsilon = -\dfrac{1}{2}\Delta_x u^\varepsilon + W^\varepsilon u^\varepsilon + (W^\varepsilon - \widetilde{W}^\varepsilon)\widetilde{\phi}^\varepsilon,\\ u^\varepsilon(0,\cdot) \equiv 0. \end{cases} \tag{4.18}$$

Thanks to (4.16) we deduce that for any $p \in (2,\infty]$ and any $t \in [0,T]$

$$\|u^\varepsilon\|_{L^\infty((0,t),L^2(\mathbb{R}^2))} \leq C\|W^\varepsilon - \widetilde{W}^\varepsilon\|_{L^1((0,t),L^p(\mathbb{R}^2))}, \tag{4.19}$$

and, by using (4.10) and Strichartz estimates in dimension 2 [7], we deduce that for any $s \in [2,\infty)$ and $q \in (2,\infty]$ we have

$$\|u^\varepsilon\|_{L^{s^*}((0,t),L^s(\mathbb{R}^2))} \leq C\|W^\varepsilon - \widetilde{W}^\varepsilon\|_{L^1((0,t),L^q(\mathbb{R}^2))}. \tag{4.20}$$

Let $q \in (2,+\infty)$. By using (4.16) (and the same estimate for ϕ^ε) and (4.19), we obtain

$$\left\||\phi^\varepsilon|^2 - |\widetilde{\phi}^\varepsilon|^2\right\|_{L^\infty((0,t),L^s(\mathbb{R}^2))} \leq C\|W^\varepsilon - \widetilde{W}^\varepsilon\|_{L^1((0,t),L^q(\mathbb{R}^2))}, \tag{4.21}$$

where $s = \frac{2q}{2+q}$. By (4.6), we deduce

$$\left\|G^\varepsilon * \left(|\phi^\varepsilon|^2 - |\widetilde{\phi}^\varepsilon|^2\right)\right\|_{L^q(\mathbb{R}^2)}(t) \leq C\int_0^t \|W^\varepsilon - \widetilde{W}^\varepsilon\|_{L^q(\mathbb{R}^2)}(\tau)\,d\tau.$$

Consequently (4.15) yields

$$\|W^\varepsilon - \widetilde{W}^\varepsilon\|_{L^q(\mathbb{R}^2)}(t) \leq C\int_0^t \|W^\varepsilon - \widetilde{W}^\varepsilon\|_{L^q(\mathbb{R}^2)}(\tau)\,d\tau + \|R^\varepsilon\|_{L^q(\mathbb{R}^2)}(t).$$

Thanks to (4.17), we deduce from a Gronwall argument applied to the above inequality that

$$\|W^\varepsilon - \widetilde{W}^\varepsilon\|_{L^\infty((0,T),L^q(\mathbb{R}^2))} \leq C\eta \qquad \forall q \in (2,\infty). \tag{4.22}$$

From this estimate together with (4.20), (4.16), and (4.19), we deduce that for any $r \in (2,\infty)$, $s \in (2,r^*)$, we have

$$\left\||\phi^\varepsilon|^2 - |\widetilde{\phi}^\varepsilon|^2\right\|_{L^r((0,T),L^s(\mathbb{R}^2))} + \left\||\phi^\varepsilon|^2 - |\widetilde{\phi}^\varepsilon|^2\right\|_{L^\infty((0,T),L^1(\mathbb{R}^2))} < C\eta,$$

which leads to (4.13) in view of (4.15), (4.7), and (4.17).

Let us now improve the estimate on $\phi^\varepsilon - \widetilde{\phi}^\varepsilon$ and show that (4.12) holds. To this aim, we first differentiate (4.18) with respect to x and obtain

$$\begin{cases} i\partial_t v^\varepsilon = -\dfrac{1}{2}\Delta_x v^\varepsilon + W^\varepsilon v^\varepsilon + (\nabla_x W^\varepsilon)u^\varepsilon, + (\nabla_x W^\varepsilon - \nabla_x \widetilde{W}^\varepsilon)\widetilde{\phi}^\varepsilon + (W^\varepsilon - \widetilde{W}^\varepsilon)\nabla_x\widetilde{\phi}^\varepsilon, \\ v^\varepsilon(0,\cdot) \equiv 0, \end{cases} \tag{4.23}$$

where we have denoted $v^\varepsilon = \nabla_x u^\varepsilon$. By combining (4.9), (4.10), (4.20), and (4.22), we get

$$\|(\nabla_x W^\varepsilon)u^\varepsilon + (W^\varepsilon - \widetilde{W}^\varepsilon)\nabla_x\widetilde{\phi}^\varepsilon\|_{L^1((0,T),L^2(\mathbb{R}^2))} \leq C\eta;$$

thus, for any $q \in (2,\infty]$ and $t \in [0,T]$, we have

$$\|v^\varepsilon\|_{L^2(\mathbb{R}^2)}(t) \leq C\eta + C\|\nabla_x W^\varepsilon - \nabla_x\widetilde{W}^\varepsilon\|_{L^1((0,t),L^q(\mathbb{R}^2))}. \tag{4.24}$$

Additionally, by (4.9) and (4.16) and Young's inequality (4.14), we get

$$\left\|(G^\varepsilon - \widetilde{G^\varepsilon}) * \nabla_x |\widetilde{\phi}^\varepsilon|^2\right\|_{L^r((0,t),L^q(\mathbb{R}^2))} \leq C\eta \qquad \forall q \in (2,\infty), \quad \forall r \in [1,q^*).$$

Moreover, using (4.16), (4.9), and (4.20), we have for any $s \in (1,2)$

$$\left\|\nabla_x \left(|\phi^\varepsilon|^2 - |\widetilde{\phi}^\varepsilon|^2\right)\right\|_{L^1((0,t),L^s(\mathbb{R}^2))} \leq C\eta + C\|v^\varepsilon\|_{L^1((0,T),L^2(\mathbb{R}^2))}.$$

Thus, writing

$$\nabla_x W^\varepsilon - \nabla_x \widetilde{W}^\varepsilon = G^\varepsilon * \nabla_x \left(|\phi^\varepsilon|^2 - |\widetilde{\phi}^\varepsilon|^2\right) + (G^\varepsilon - \widetilde{G^\varepsilon}) * \nabla_x |\widetilde{\phi}^\varepsilon|^2 \tag{4.25}$$

and using (4.6), we deduce that for any $q \in (2,\infty)$

$$\left\|\nabla_x W^\varepsilon - \nabla_x \widetilde{W}^\varepsilon\right\|_{L^1((0,t),L^q(\mathbb{R}^2))} \leq C\eta + C\|v^\varepsilon\|_{L^1((0,t),L^2(\mathbb{R}^2))}.$$

Inserting this inequality in (4.24) leads, through a Gronwall argument, to

$$\|v^\varepsilon\|_{L^\infty((0,T),L^2(\mathbb{R}^2))} \leq C\eta.$$

Going back to (4.23), it is readily seen from the above two estimates and from Proposition 4.2 that

$$\left\|i\partial_t v^\varepsilon + \frac{1}{2}\Delta v^\varepsilon\right\|_{L^1((0,T),L^2(\mathbb{R}^2))} \leq C\eta,$$

which leads to (4.12) through a Strichartz estimate. ∎

In section 6, in order to get estimate (2.14), we will need to deal with strong solutions.

LEMMA 4.4. *Under Assumption* 4.1, *let* $\phi_0 \in H^2(\mathbb{R}^2)$. *Then for any* $T > 0$ *the solution* ϕ^ε *of* (4.4), (4.5) *belongs to* $L^\infty((0,T),H^2(\mathbb{R}^2))$ *and its norm is bounded independently of* ε.

Proof. Denote $u^\varepsilon = \Delta_x \phi^\varepsilon$. By differentiating twice (4.4) with respect to x, we get

$$i\partial_t u^\varepsilon = -\frac{1}{2}\Delta_x u^\varepsilon + W^\varepsilon u^\varepsilon + 2\nabla_x W^\varepsilon \cdot \nabla_x \phi^\varepsilon + \Delta_x W^\varepsilon \phi^\varepsilon.$$

The source term in this Schrödinger equation on u^ε is written as

$$2\nabla_x W^\varepsilon \cdot \nabla_x \phi^\varepsilon + \phi^\varepsilon G^\varepsilon * \left(2|\nabla_x \phi^\varepsilon|^2\right) + 2\phi^\varepsilon \, \mathcal{R}e\, G^\varepsilon * \left(\overline{\phi^\varepsilon} u^\varepsilon\right).$$

The first term $\nabla_x W^\varepsilon \cdot \nabla_x \phi^\varepsilon$ can be estimated thanks to (4.9) and (4.10):

$$\|\nabla_x W^\varepsilon \cdot \nabla_x \phi^\varepsilon\|_{L^1((0,t),L^2(\mathbb{R}^2))} \leq \|\nabla_x W^\varepsilon\|_{L^{4/3}((0,t),L^4(\mathbb{R}^2))} \|\nabla_x \phi^\varepsilon\|_{L^4((0,t),L^4(\mathbb{R}^2))} \leq C.$$

The second term can be estimated thanks to (4.6):

$$\begin{aligned}
&\left\|\phi^\varepsilon G^\varepsilon * \left(2|\nabla_x \phi^\varepsilon|^2\right)\right\|_{L^1((0,t),L^2(\mathbb{R}^2))} \\
&\qquad \leq \|\phi^\varepsilon\|_{L^{3/2}((0,t),L^3(\mathbb{R}^2))} \left\|G^\varepsilon * \left(2|\nabla_x \phi^\varepsilon|^2\right)\right\|_{L^3((0,t),L^6(\mathbb{R}^2))} \\
&\qquad \leq C\|\phi^\varepsilon\|_{L^{3/2}((0,t),L^3(\mathbb{R}^2))} \|\nabla_x \phi^\varepsilon\|^2_{L^6((0,t),L^3(\mathbb{R}^2))} \leq C.
\end{aligned}$$

To treat the third term, we also apply (4.6), (4.9), and (4.10):

$$\begin{aligned}
&\|\phi^\varepsilon G^\varepsilon * (2\phi^\varepsilon u^\varepsilon)\|_{L^1((0,t),L^2(\mathbb{R}^2))} \\
&\qquad \le C\|\phi^\varepsilon\|_{L^\infty((0,t),L^3(\mathbb{R}^2))}\ \|G^\varepsilon * (2\phi^\varepsilon u^\varepsilon)\|_{L^1((0,t),L^6(\mathbb{R}^2))} \\
&\qquad \le C\|\phi^\varepsilon\|_{L^\infty((0,t),L^3(\mathbb{R}^2))}\ \|\phi^\varepsilon\|_{L^\infty((0,t),L^6(\mathbb{R}^2)}\ \|u^\varepsilon\|_{L^1((0,t),L^2(\mathbb{R}^2))} \\
&\qquad \le C\,\|u^\varepsilon\|_{L^1((0,t),L^2(\mathbb{R}^2))}\,.
\end{aligned}$$

Hence, for any $t \le T$,

$$\|u^\varepsilon(t)\|_{L^2(\mathbb{R}^2)} \le C + C\int_0^t \|u^\varepsilon(\tau)\|_{L^2(\mathbb{R}^2)}\ d\tau,$$

which leads to the result thanks to a Gronwall argument. ∎

4.2. Application: Proof of Theorem 2.6. Thanks to Lemma B.1 given in Appendix B, the kernel

$$G^{2D}(x) = \frac{1}{4\pi|x|}$$

of the 2D surface density model (2.2)–(2.3) clearly satisfies Assumption 4.1. Moreover, by using Lemma B.2 and the fact that $\int_{\mathbb{R}} |\chi^\varepsilon|^2 dz = 1$, it is readily seen that the kernel of the 2.5D adiabatic model given by

$$G^{2.5D}(x) = \iint_{\mathbb{R}^2} \frac{1}{4\pi\left(|x|^2 + (z-z')^2\right)^{1/2}} |\chi^\varepsilon(z')|^2\, |\chi^\varepsilon(z)|^2\, dz'\, dz$$

also satisfies Assumption 4.1. Therefore an application of Proposition 4.2 gives the existence of unique weak solutions and estimates independent of ε for the two approximate models. The first parts of Theorems 2.5 and 2.6 are thus proved.

To conclude the proof of Theorem 2.6, it suffices to apply Proposition 4.3. Indeed, setting

$$\begin{aligned}
H^\varepsilon(x) &= \frac{1}{4\pi|x|} - G^{2.5D}(x) \\
&= \frac{1}{4\pi|x|} - \iint_{\mathbb{R}^2} \frac{1}{4\pi\left(|x|^2 + (z-z')^2\right)^{1/2}} |\chi^\varepsilon(z')|^2\, |\chi^\varepsilon(z)|^2\, dz\, dz' \\
&= \frac{1}{4\pi} \iint_{\mathbb{R}^2} \int_0^{\varepsilon|z-z'|} \frac{\xi}{(|x|^2+\xi^2)^{3/2}} |\chi(z)|^2\, |\chi(z')|^2\, d\xi\, dz\, dz',
\end{aligned}$$

and noticing that

$$\int_{\mathbb{R}^2} \frac{\xi}{(|x|^2+\xi^2)^{3/2}}\, dx = 2\pi \quad \text{for } \xi > 0,$$

we deduce from (2.1) that

$$\|H^\varepsilon\|_{L^1(\mathbb{R}^2)} = \frac{\varepsilon}{2} \iint_{\mathbb{R}^2} |z-z'|\, |\chi(z)|^2\, |\chi(z')|^2\, dz\, dz' = C\varepsilon.$$

This leads to (2.10), from which we deduce (2.12). In order to prove (2.11), we write

$$V^{2.5D} - V^{2D} = \frac{1}{4\pi r} *_x \left(n_s^{2.5D} - n_s^{2D}\right) + \widetilde{H^\varepsilon} *_x n_s^{2.5D},$$

where

$$\widetilde{H^\varepsilon}(x,z) = -\frac{1}{4\pi r} + \frac{1}{4\pi r} *_z |\chi^\varepsilon|^2. \tag{4.26}$$

It is then enough to remark that

$$\begin{aligned}\left\| \frac{1}{4\pi r} *_x \left(n_s^{2.5D} - n_s^{2D}\right)\right\|_{L^q((0,T),L^\infty(\mathbb{R}^3))} &\le \left\| \frac{1}{4\pi|x|} *_x \left(n_s^{2.5D} - n_s^{2D}\right)\right\|_{L^q((0,T),L^\infty(\mathbb{R}^2))} \\ &\le C\varepsilon\end{aligned}$$

and that

$$\widetilde{H^\varepsilon}(x,z) = \int_{\mathbb{R}} \int_{(z-z')} \frac{\xi}{(|x|^2+\xi^2)^{3/2}} |\chi^\varepsilon(z')|^2 \, d\xi \, dz',$$

which implies

$$\begin{aligned}&|\widetilde{H^\varepsilon} *_x n_s^{2.5D}|(t,x,z) \\ &\quad \le \|n_s^{2.5D}(t,\cdot)\|_{L^\infty(\mathbb{R}^2)} \int_{\mathbb{R}} \int_{min(z,z-z')}^{max(z,z-z')} \int_{\mathbb{R}^2} \frac{|\xi|}{(|x|^2+\xi^2)^{3/2}} |\chi^\varepsilon(z')|^2 \, dx d\xi dz', \\ &\quad = 2\pi \|n_s^{2.5D}(t,\cdot)\|_{L^\infty(\mathbb{R}^2)} \int_{\mathbb{R}} |z'| \, |\chi^\varepsilon(z')|^2 \, dz' = C\varepsilon \|n_s^{2.5D}(t,\cdot)\|_{L^\infty(\mathbb{R}^2)},\end{aligned}$$

and the right-hand side is an $O(\varepsilon)$ is view of (2.12).

5. The 2.5D adiabatic model is a second-order approximation. In this section we end the proof of Theorem 2.5 initiated in section 4.2. Consider the solution ψ^{3D}, V^{3D} of (1.1)–(1.3) with the initial data $\psi_0^\varepsilon = \phi_0 \, \chi^\varepsilon$ and the solution $\phi^{2.5D}$, $V^{2.5D}$ of (2.5), (2.6) corresponding to the initial data ϕ_0. Assumption 2.4 leads in particular to the uniform-in-ε estimate

$$\|\psi_0^\varepsilon\|_{L^2(\mathbb{R}^3)} + \|\nabla_x \psi_0^\varepsilon\|_{L^2(\mathbb{R}^3)} \le C.$$

Proposition 3.2 then implies the following uniform bounds:

$$\|V^{3D}\|_{L^\infty((0,T)\times\mathbb{R}^3)} + \|\nabla_{x,z} V^{3D}\|_{\infty,q,\infty} \le C, \qquad 2 < q < \infty, \tag{5.1}$$

$$\|\psi^{3D}\|_{q^*,q,2} + \|\nabla_x \psi^{3D}\|_{q^*,q,2} \le C, \qquad 2 \le q < \infty. \tag{5.2}$$

Furthermore, Lemma 3.3 implies

$$\left\|(\mathbb{I} - \Pi^\varepsilon)\, \psi^{3D}\right\|_{q^*,q,2} = \mathcal{O}(\varepsilon), \qquad 2 \le q < \infty. \tag{5.3}$$

We start by proving (2.8). To this aim, we write

$$\begin{aligned} V^{3D} - V^{2.5D} &= \frac{1}{4\pi r} * \left(n^{3D} - n^{2.5D}\right) \\ &= \frac{1}{4\pi r} * \left(|\Pi^\varepsilon \psi^{3D}|^2 - |\chi^\varepsilon \phi^{2.5D}|^2\right) + R_a^\varepsilon + R_b^\varepsilon, \end{aligned} \tag{5.4}$$

where the remainder terms are

$$R_a^\varepsilon = \frac{1}{4\pi r} * |\left(\mathbb{I} - \Pi^\varepsilon\right)\psi^{3D}|^2; \quad R_b^\varepsilon = \frac{2}{4\pi r} * \mathcal{R}e\left(\overline{\Pi^\varepsilon \psi^{3D}}\left(\mathbb{I} - \Pi^\varepsilon\right)\psi^{3D}\right).$$

Estimating the remainders R^ε and R^ε. On the one hand, estimates (5.3), (B.3), and (B.5) lead to

$$\|R_a^\varepsilon\|_{1,q,\infty} \leq C\varepsilon^2 \qquad \forall q \in (2,\infty]. \tag{5.5}$$

On the other hand, by orthogonality we have $\langle \overline{\Pi^\varepsilon \psi^{3D}}\left(\mathbb{I} - \Pi^\varepsilon\right)\psi^{3D}\rangle = 0$. Consequently, (B.9) implies for any $q \in (2,\infty)$ and pointwise in time

$$\|R_b^\varepsilon\|_{L^\infty(\mathbb{R}^3)} \leq C\,\|z\chi^\varepsilon\|_{L^2(\mathbb{R})}^{1-2/q}\,\left\|\psi^{3D}\right\|_{2q,2}\left\|\left(\mathbb{I} - \Pi^\varepsilon\right)\psi^{3D}\right\|_{2q,2}.$$

Additionally, we deduce from (5.2) and the Sobolev embedding $H^1(\mathbb{R}^2) \hookrightarrow L^q(\mathbb{R}^2)$ that

$$\left\|\psi^{3D}\right\|_{\infty,q,2} \leq \left\|\psi^{3D}\right\|_{L^\infty((0,T),H^1(\mathbb{R}^2,L^2(\mathbb{R})))} \leq C. \tag{5.6}$$

Moreover, by (2.1) we have $\|z\chi^\varepsilon\|_{L^2(\mathbb{R})}^{1-2/q} = \mathcal{O}(\varepsilon^{1-2/q})$; therefore

$$\|R_b^\varepsilon\|_{L^\infty(\mathbb{R}^3)}(t) \leq C\varepsilon^{1-2/q}\left\|\left(\mathbb{I} - \Pi^\varepsilon\right)\psi^{3D}\right\|_{2q,2}(t).$$

Similarly, by (B.8), we have for any $\alpha \in (0,1)$ and $q \in [2,\infty)$

$$\begin{aligned}\|R_b^\varepsilon\|_{q,\infty}(t) &\leq C\,\|z\chi^\varepsilon\|_{L^2(\mathbb{R})}^{1-\alpha}\,\left\|\psi^{3D}\right\|_{\frac{4q}{2+\alpha q},2}(t)\left\|\left(\mathbb{I} - \Pi^\varepsilon\right)\psi^{3D}\right\|_{\frac{4q}{2+\alpha q},2}(t)\\ &\leq C\varepsilon^{1-\alpha}\,\left\|\left(\mathbb{I} - \Pi^\varepsilon\right)\psi^{3D}\right\|_{\frac{4q}{2+\alpha q},2}(t).\end{aligned}$$

By (5.3), we finally get

$$\forall \alpha \in (0,1), \quad \forall q \in [2,\infty], \qquad \|R_b^\varepsilon\|_{1,q,\infty} \leq C\varepsilon^{2-\alpha}, \tag{5.7}$$

where the constant C depends only on α.

Estimating the first term in the right-hand side of (5.4). We shall estimate the difference

$$w^\varepsilon := \Pi^\varepsilon\psi^{3D} - \chi^\varepsilon\phi^{2.5D}e^{-iE^\varepsilon t}.$$

To this aim, we notice that

$$\begin{cases} i\partial_t w^\varepsilon = -\dfrac{1}{2}\Delta_x w^\varepsilon + A^\varepsilon w^\varepsilon + \langle V^{3D}|\chi^\varepsilon|^2\rangle w^\varepsilon + f^\varepsilon + g^\varepsilon, \\ \omega^\varepsilon(0,x,z) = 0, \end{cases} \tag{5.8}$$

where

$$f^\varepsilon = \langle\left(V^{3D} - V^{2.5D}\right)|\chi^\varepsilon|^2\rangle\chi^\varepsilon\phi^{2.5D}e^{-iE^\varepsilon t}; \quad g^\varepsilon = \Pi^\varepsilon V^{3D}(\mathbb{I} - \Pi^\varepsilon)\psi^{3D}.$$

Standard L^2 estimates for a Schrödinger equation with a source term then imply

$$\|w^\varepsilon\|_{\infty,2,2} \leq \|f^\varepsilon\|_{1,2,2} + \|g^\varepsilon\|_{1,2,2}.$$

Remarking that

$$\Pi^\varepsilon V^{3D}(\mathbb{I}-\Pi^\varepsilon)=\Pi^\varepsilon[\Pi^\varepsilon,V^{3D}](\mathbb{I}-\Pi^\varepsilon),$$

we deduce from Lemma 2.3, (5.1), and (5.3) that

$$\|g^\varepsilon\|_{1,2,2}\le C\varepsilon\|\partial_z V^{3D}\|_{4/3,4,\infty}\|(\mathbb{I}-\Pi^\varepsilon)\psi^{3D}\|_{4,4,2}=\mathcal{O}(\varepsilon^2).$$

Additionally, in the same spirit as the proof of (5.6), by applying (4.9) and standard Sobolev embeddings, we get

$$\left\|\phi^{2.5D}\right\|_{L^\infty((0,T),L^q(\mathbb{R}^2))}\le C\qquad\forall q\in[2,\infty). \tag{5.9}$$

Therefore, it can be easily seen that for any $q\in(2,\infty]$ we have

$$\|f^\varepsilon\|_{1,2,2}\le C\|V^{3D}-V^{2.5D}\|_{1,q,\infty},$$

and we finally obtain

$$\|w^\varepsilon\|_{\infty,2,2}\le C\|V^{3D}-V^{2.5D}\|_{1,q,\infty}+\mathcal{O}(\varepsilon^2). \tag{5.10}$$

Applying the Strichartz inequality (A.5) to (5.8) after having noticed estimate (5.1), we obtain for any $q\in(2,\infty]$, $s\in[2,\infty)$,

$$\|w^\varepsilon\|_{s^*,s,2}\le C\|V^{3D}-V^{2.5D}\|_{1,q,\infty}+\mathcal{O}(\varepsilon^2). \tag{5.11}$$

This gives the following estimate for the first term of the right-hand side of (5.4), for any $q\in(2,\infty)$:

$$\begin{aligned}\left\||\Pi^\varepsilon\psi^{3D}|^2-|\chi^\varepsilon\phi^{2.5D}|^2\right\|_{\frac{2q}{2+q},1}(t)&\le\left(\|\psi^{3D}\|_{\infty,q,2}+\|\phi^{2.5D}\|_{L^\infty((0,t),L^q(\mathbb{R}^2))}\right)\|w^\varepsilon\|_{\infty,2,2}\\&\le C\int_0^t\|V^{3D}-V^{2.5D}\|_{q,\infty}(\tau)\,d\tau+\mathcal{O}(\varepsilon^2),\end{aligned}$$

where we used (5.6), (5.9), and (5.10).

End of the proof. By applying (B.3), we deduce

$$\left\|\frac{1}{r}*\left(|\Pi^\varepsilon\psi^{3D}|^2-|\chi^\varepsilon\phi^{2.5D}|^2\right)\right\|_{q,\infty}(t)\le C\int_0^t\|V^{3D}-V^{2.5D}\|_{q,\infty}(\tau)\,d\tau+\mathcal{O}(\varepsilon^2),$$

where $q\in(2,\infty)$. Consequently, (5.4) yields

$$\begin{aligned}\|(V^{3D}-V^{2.5D})\|_{q,\infty}(t)\le\;& C\int_0^t\|(V^{3D}-V^{2.5D})\|_{q,\infty}(\tau)\,d\tau\\&+\|R_a^\varepsilon\|_{q,\infty}(t)+\|R_b^\varepsilon\|_{q,\infty}(t)+O(\varepsilon^2).\end{aligned}$$

Recalling estimates (5.5) and (5.7) for the remainders, a Gronwall argument leads to the bound

$$\|(V^{3D}-V^{2.5D})\|_{\infty,q,\infty}\le C\varepsilon^{2-\alpha}\qquad\forall q\in(2,\infty),\quad\forall\alpha\in(0,1).$$

To conclude the proof, we insert this estimate into (5.11) and obtain

$$\|w^\varepsilon\|_{s^*,s,2}\le C\varepsilon^{2-\alpha}\qquad\forall s\in[2,\infty),\quad\forall\alpha\in(0,1).$$

Then we now have, for any $q \in [2,\infty)$ and $s < q^*$,

$$\left\| |\Pi^\varepsilon \psi^{3D}|^2 - |\chi^\varepsilon \phi^{2.5D}|^2 \right\|_{s,q,1} \leq C\varepsilon^{2-\alpha} \qquad \forall \alpha \in (0,1)$$

and we apply (B.5). By using again (5.4), (5.5), and (5.7), we find (2.8).

In order to prove (2.7), we simply remark that

$$\|\psi^{3D} - \phi^{2.5D}\chi^\varepsilon\, e^{-itE/\varepsilon^2}\|_{q^*,q,2} \leq \|w_\varepsilon\|_{q^*,q,2} + \|(\mathbb{I} - \Pi^\varepsilon)\psi^{3D}\|_{q^*,q,2},$$

then use (5.3) and (5.11). To prove (2.9), we remark that

$$n_s^{3D} - n_s^{2.5D} = |\Pi^\varepsilon \psi^{3D}|^2 - |\chi^\varepsilon \phi^{2.5D}|^2 + |\left(\mathbb{I} - \Pi^\varepsilon\right)\psi^{3D}|^2.$$

6. The 2D surface density model is a first-order approximation. In this section we prove Theorem 2.7, which gives estimates from below, showing that the accuracy of the limit model is exactly $\mathcal{O}(\varepsilon)$. We denote by $\phi^{2.5D}$, $V^{2.5D}$ and by ϕ^{2D}, V^{2D}, respectively, the solutions of (2.5), (2.6) and (2.2), (2.4). For notational simplicity, we denote

$$V_0^{2.5D}(t,x) = V^{2.5D}(t,x,0); \quad V_0^{2D} = V^{2D}(t,x,0).$$

Since we assume that the initial data ϕ_0 belongs to $H^2(\mathbb{R}^2)$, an application of Lemma 4.4 gives

$$\|\phi^{2.5D}\|_{L^\infty((0,T),H^2(\mathbb{R}^2))} + \|\phi^{2D}\|_{L^\infty((0,T),H^2(\mathbb{R}^2))} \leq C. \tag{6.1}$$

Moreover, with (2.10) and the Sobolev embedding $H^1(\mathbb{R}^2) \hookrightarrow L^{2q}(\mathbb{R}^2)$, we obtain

$$\|n_s^{2.5D} - n_s^{2D}\|_{L^\infty((0,T),L^q(\mathbb{R}^2))} \leq C\varepsilon \qquad \forall q \in [1,\infty). \tag{6.2}$$

Now we recall that

$$V_0^{2.5D} - V_0^{2D} = \frac{1}{4\pi|x|} *_x \left(n_s^{2.5D} - n_s^{2D}\right) + \widetilde{H^\varepsilon}(\cdot,0) *_x |\phi^{2.5D}|^2,$$

where $\widetilde{H^\varepsilon}$ is defined in (4.26). Hence, pointwise in time we get

$$\begin{aligned} \|V_0^{2.5D} - V_0^{2D}\|_{L^\infty(\mathbb{R}^2)} + \left\| \frac{1}{4\pi|x|} *_x \left(n_s^{2.5D} - n_s^{2D}\right)\right\|_{L^\infty(\mathbb{R}^2)} \\ \geq \left\| \widetilde{H^\varepsilon}(\cdot,0) *_x |\phi^{2.5D}|^2 \right\|_{L^\infty(\mathbb{R}^2)}. \end{aligned} \tag{6.3}$$

Additionally, a straightforward calculation leads to

$$i\partial_t(x\phi^{2.5D}) = -\frac{1}{2}\Delta_x(x\phi^{2.5D}) + V^{2.5D}(x\phi^{2.5D}) + \nabla_x \phi^{2.5D};$$

thus

$$\|x\phi^{2.5D}\|_{L^\infty((0,T),L^2(\mathbb{R}^2))} \leq \|x\phi_0\|_{L^2(\mathbb{R}^2)} + \|\nabla_x \phi^{2.5D}\|_{L^1((0,T),L^2(\mathbb{R}^2))} \leq C,$$

where we used (4.9). For any $R > 0$, let us denote $\mathcal{B}_R = \{x \in \mathbb{R}^2, |x| < R\}$. We have

$$\begin{aligned} \|\phi^{2.5D}\|^2_{L^\infty((0,T),L^2(\mathcal{B}_R))} &\geq \|\phi_0\|^2_{L^2(\mathbb{R}^2)} - \frac{1}{R^2}\|x\phi^{2.5D}\|^2_{L^\infty((0,T),L^2(\mathbb{R}^2)} \\ &\geq \|\phi_0\|^2_{L^2(\mathbb{R}^2)} - \frac{C}{R^2}. \end{aligned}$$

Since by assumption we have $\|\phi_0\|_{L^2(\mathbb{R}^2))} = 2\eta > 0$, by choosing R large enough we have

$$\|\phi^{2.5D}\|_{L^\infty((0,T),L^2(\mathcal{B}_R))} > \eta;$$

then

$$\forall t \in [0,T], \qquad \max_{\mathcal{B}_R} |\phi^{2.5D}(t,\cdot)|^2 > \frac{\eta^2}{\pi R^2}.$$

By using (6.1) and the Sobolev embedding $H^2(\mathbb{R}^2) \hookrightarrow C^{0,1/2}(\mathbb{R}^2)$, we deduce finally that there exists $r_0 > 0$, $\alpha > 0$ and $x_0(t) \in \mathbb{R}^2$ defined almost everywhere such that, for a.e. $t \in [0,T]$, we have

$$|\phi^{2.5D}|^2(t,x) > \alpha \quad \forall x \in \mathbb{R}^2 \quad \text{such that } |x - x_0(t)| < r_0. \tag{6.4}$$

For $t \in [0,T]$, we have

$$\begin{aligned}
&\left|\widetilde{H^\varepsilon}(\cdot,0) *_x |\phi^{2.5D}|^2\right|(x_0(t)) \\
&= \int_{\mathbb{R}^2}\int_{\mathbb{R}}\int_0^{\varepsilon|z'|} \frac{\xi}{(|x'|^2+\xi^2)^{3/2}} |\chi(z')|^2 |\phi^{2.5D}(x_0(t)-x')|^2 \, d\xi dz' dx' \\
&\geq 2\pi\alpha \int_{\mathbb{R}}\int_0^{\varepsilon|z'|}\int_{r=0}^{r_0} \frac{r\xi}{(r^2+\xi^2)^{3/2}} |\chi(z')|^2 \, dr d\xi dz' \\
&= 2\pi\alpha \int_{\mathbb{R}} \varepsilon|z'| \left(1 - \frac{\varepsilon|z'|}{r_0^2 + (r_0^2+\varepsilon^2|z'|^2)^{1/2}}\right) |\chi(z')|^2 \, dz' \\
&\geq C_1\varepsilon - C_2\varepsilon^2 \geq C_0\varepsilon,
\end{aligned}$$

where $C_0 > 0$ and ε is small enough. Therefore, by applying (6.3) and using (B.2), we have for $t \in [0,T]$,

$$\|V_0^{2.5D} - V_0^{2D}\|_{L^\infty(\mathbb{R}^2)} + \left\|n_s^{2.5D} - n_s^{2D}\right\|_{L^q(\mathbb{R}^2)}^{\theta} \left\|n_s^{2.5D} - n_s^{2D}\right\|_{L^1(\mathbb{R}^2)}^{1-\theta} \geq C\varepsilon,$$

with any $2 < q < \infty$ and $\theta = \frac{q}{2q-2}$. Bounding $\left\|n_s^{2.5D} - n_s^{2D}\right\|_{L^1(\mathbb{R}^2)}$ by $C\varepsilon$ in view of (6.2), one deduces for any $q \in (2,\infty)$

$$\frac{\|V_0^{2.5D} - V_0^{2D}\|_{L^\infty(\mathbb{R}^2)}}{\varepsilon} + \left(\frac{\left\|n_s^{2.5D} - n_s^{2D}\right\|_{L^q(\mathbb{R}^2)}}{\varepsilon}\right)^{\theta} \geq C_0'.$$

Proceeding analogously, we obtain

$$\frac{\|V_0^{2.5D} - V_0^{2D}\|_{L^\infty(\mathbb{R}^2)}}{\varepsilon} + \left(\frac{\left\|n_s^{2.5D} - n_s^{2D}\right\|_{L^1(\mathbb{R}^2)}}{\varepsilon}\right)^{1-\theta} \geq C_0'.$$

Consequently, we deduce that

$$\|V_0^{2.5D} - V_0^{2D}\|_{L^\infty(\mathbb{R}^2)} + \left\|n_s^{2.5D} - n_s^{2D}\right\|_{L^q(\mathbb{R}^2)} \geq C\varepsilon \quad \forall q \in (2,+\infty) \tag{6.5}$$

and

$$\|V_0^{2.5D} - V_0^{2D}\|_{L^\infty(\mathbb{R}^2)} + \left\|n_s^{2.5D} - n_s^{2D}\right\|_{L^1(\mathbb{R}^2)} \geq C\varepsilon.$$

The last inequality implies, by a simple interpolation argument, that (6.5) actually holds for $q \in [1, +\infty)$, which finishes the proof.

Appendix A. Strichartz estimates in $L^{q^*}L^q_xL^2_z$. For any $q \in [2, \infty)$ we recall the notation $q^* = 2q/(q-2)$: in the usual terminology for the Strichartz estimates, the pair (q^*, q) is said to be admissible. The space $L^{q^*}_t L^q_x L^2_z$ was defined in section 2. Let us first state an extension of the standard Strichartz estimate for Schrödinger equations on $\mathbb{R}^2$ with values in a Hilbert space [6, 7, 11, 20, 22].

LEMMA A.1. *Let $T > 0$ and let $\mathcal{H}$ be a separable Hilbert space. For $\psi_0 \in L^2(\mathbb{R}^2, \mathcal{H})$ and $g \in L^1((0,T), L^2(\mathbb{R}^2, \mathcal{H}))$, we consider the solution $\psi(t,x) \in L^\infty((0,T), L^2(\mathbb{R}^2, \mathcal{H}))$ of*

$$\text{(A.1)} \qquad \begin{cases} i\partial_t \psi = -\dfrac{1}{2}\Delta_x \psi + g, \\ \psi(0,\cdot) = \psi_0\,. \end{cases}$$

Then for any $q \in [2, \infty)$, the function ψ belongs to $L^{q^*}((0,T), L^q(\mathbb{R}^2, \mathcal{H}))$ and satisfies

$$\text{(A.2)} \qquad \|\psi\|_{L^{q^*}((0,T),L^q(\mathbb{R}^2,\mathcal{H}))} \leq C\|\psi_0\|_{L^2(\mathbb{R}^2,\mathcal{H})} + C\|g\|_{L^1((0,T),L^2(\mathbb{R}^2,\mathcal{H}))},$$

where $C > 0$ denotes a constant.

Proof. Let $(\cdot,\cdot)_{\mathcal{H}}$ denote the scalar product on $\mathcal{H}$ and let $(\chi_p)_{p\in\mathbb{N}^*}$ be a Hilbertian basis of $\mathcal{H}$. We shall use the Strichartz estimate for mixed quantum states proved in [5]. For this, let us introduce the following functional space:

$$\widetilde{L^q}(\mathbb{R}^2, \mathcal{H}) = \left\{ \psi \in L^q(\mathbb{R}^2, \mathcal{H}) \,:\quad \|\psi\|^2_{\widetilde{L^q}(\mathbb{R}^2,\mathcal{H})} = \sum_{p\geq 1} \|\psi_p\|^2_{L^q(\mathbb{R}^2)} < +\infty \right\},$$

where we have denoted $\psi_p = (\psi, \chi_p)_{\mathcal{H}}$ (note that this functional space a priori depends on the choice of the Hilbertian basis χ_p). This space is continuously embedded in $L^q(\mathbb{R}^2, \mathcal{H})$; indeed we have

$$\text{(A.3)} \qquad \|\psi\|_{L^q(\mathbb{R}^2,\mathcal{H})} = \left\| \sum_{p\geq 1} |\psi_p|^2 \right\|^{1/2}_{L^{q/2}(\mathbb{R}^2)} \leq \left(\sum_{p\geq 1} \left\| |\psi_p|^2 \right\|_{L^{q/2}(\mathbb{R}^2)} \right)^{1/2} = \|\psi\|_{\widetilde{L^q}(\mathbb{R}^2,\mathcal{H})}.$$

This inequality becomes an equality in the special case $q = 2$ and we have the identification $\widetilde{L^2}(\mathbb{R}^2, \mathcal{H}) = L^2(\mathbb{R}^2, \mathcal{H})$.

This functional space $\widetilde{L^q}(\mathbb{R}^2, \mathcal{H})$ can be identified with the space $L^q(\lambda)$ introduced in [5, Definition 2.1] (in dimension 2 instead of dimension 3), with the choice $\lambda = (1, 1, 1, \ldots)$ and if ψ is identified with the sequence of its components $(\psi_p)_{p\in\mathbb{N}^*}$.

Each component ψ_p satisfies the equation

$$\begin{cases} i\partial_t \psi_p = -\dfrac{1}{2}\Delta_x \psi_p + g_p, \\ \psi_p(0,\cdot) = \psi_{0,p}, \end{cases}$$

where $g_p = (g, \chi_p)_{\mathcal{H}}$ and $\psi_{0,p} = (\psi_0, \chi_p)_{\mathcal{H}}$. Therefore, an application of [5, Theorem 2.1] (adapted to dimension 2) gives

$$\begin{aligned}\|\psi\|_{L^{q^*}((0,T),\widetilde{L^q}(\mathbb{R}^2,\mathcal{H}))} &\le C\|\psi_0\|_{\widetilde{L^2}(\mathbb{R}^2,\mathcal{H})} + C\|g\|_{L^1((0,T),\widetilde{L^2}(\mathbb{R}^2,\mathcal{H}))}\\ &= C\|\psi_0\|_{L^2(\mathbb{R}^2,\mathcal{H})} + C\|g\|_{L^1((0,T),L^2(\mathbb{R}^2,\mathcal{H}))}.\end{aligned}$$

We conclude the proof by using (A.3). ∎

Now let $\mathbb{A}$ be an unbounded operator on $\mathcal{H} = L^2(\mathbb{R})$ with the domain $\mathcal{D}(\mathbb{A})$. We assume that the operator $\mathbb{A}$ is self-adjoint and denote by $e^{it\mathbb{A}}$ the unitary group generated by $i\mathbb{A}$ on $\mathcal{H}$. Throughout this paper, the results of the appendix are applied to the operator $\mathbb{A} = -\frac{1}{2}\frac{d^2}{dz^2} + V_c^\varepsilon$. The operator $i(\frac{1}{2}\Delta_x - \mathbb{A})$, defined with an abuse of notation as $i(\frac{1}{2}\Delta_x \otimes \mathbb{I}_{\mathcal{H}} - \mathbb{I}_{L^2(\mathbb{R}^2)} \otimes \mathbb{A})$ on $H^2(\mathbb{R}^2,\mathcal{H}) \cap L^2(\mathbb{R}^2,\mathcal{D}(\mathbb{A}))$, generates a group of isometries on $L^2(\mathbb{R}^2,\mathcal{H}) = L^2(\mathbb{R}^3)$. Let us now consider the problem

$$\text{(A.4)}\qquad \begin{cases} i\partial_t\psi = -\dfrac{1}{2}\Delta_x\psi + \mathbb{A}\psi + f,\\ \psi(0,x,z) = \psi_0, \end{cases}$$

where the source term $f(t,x,z)$ is given. The following result holds.

LEMMA A.2. *Let $\psi_0 \in L^2(\mathbb{R}^3)$ and $f \in L^1((0,T),L^2(\mathbb{R}^3))$. Then for any $q \in [2,\infty)$, the solution ψ of the Schrödinger equation* (A.4) *belongs to $L_t^{q^*}L_x^qL_z^2((0,T)\times\mathbb{R}^3)$ and satisfies*

$$\text{(A.5)}\qquad \|\psi\|_{q^*,q,2} \le C\|\psi_0\|_{L^2(\mathbb{R}^3)} + C\|f\|_{L^1((0,T),L^2(\mathbb{R}^3))},$$

where C denotes a constant independent of the operator $\mathbb{A}$.

Proof. This lemma is a consequence of Lemma A.1 above. Let us denote $\phi(t,x,z) = e^{i\mathbb{A}t}\,\psi(t,x;z)$. Since $\mathbb{A}$ commutes with ∂_t and Δ_x, we clearly have

$$\begin{cases} i\partial_t\phi = -\Delta_x\phi + e^{i\mathbb{A}t}\,f,\\ \phi(0,x,z) = \psi_0. \end{cases}$$

Therefore ϕ satisfies (A.1) with $g = e^{i\mathbb{A}t}\,f$. We conclude the proof by using (A.2) since $e^{i\mathbb{A}t}$ is an isometry on $L^2(\mathbb{R})$. ∎

Appendix B. The Poisson equation with L^1 densities. This section deals with the convolution product

$$u = \frac{1}{r} * f,$$

where, $r = \sqrt{|x|^2+z^2}$ and $f \in L_x^pL_z^1$. We recall that throughout this paper $x \in \mathbb{R}^2$, $z \in \mathbb{R}$, and $L_x^pL_z^q = L^p(\mathbb{R}^2,L^q(\mathbb{R}))$. We first prove the following result in $\mathbb{R}^2$ with a convolution kernel more singular than the kernel of the Poisson equation.

LEMMA B.1. (i) *Let $f \in L^p(\mathbb{R}^2)$ with $1 < p < 2$. Then*

$$\text{(B.1)}\qquad \left\| \frac{1}{|x|} *_x f \right\|_{L^{p^\#}(\mathbb{R}^2)} \le C_p\,\|f\|_{L^p(\mathbb{R}^2)},$$

where $p^\# = \frac{2p}{2-p}$.

(ii) *Let $f \in L^p(\mathbb{R}^2) \cap L^1(\mathbb{R}^2)$ with $2 < p \le +\infty$. Then*

$$\text{(B.2)}\qquad \left\| \frac{1}{|x|} *_x f \right\|_{L^\infty(\mathbb{R}^2)} \le C_p\,\|f\|_{L^p(\mathbb{R}^2)}^{\theta}\|f\|_{L^1(\mathbb{R}^2)}^{1-\theta},$$

where $\theta = \frac{p}{2p-2}$.

Proof. The first part of the lemma is a straightforward consequence of the generalized Young's formula [18]. Indeed, the function $x \mapsto \frac{1}{|x|}$ belongs to $L^2_w(\mathbb{R}^2)$, and the function f is in $L^p(\mathbb{R}^2)$; thus $\frac{1}{|x|} *_x f$ belongs to $L^{p^\#}(\mathbb{R}^2)$, with $\frac{1}{p} + \frac{1}{2} = 1 + \frac{1}{p^\#}$.

In order to prove item (ii), for any $R > 0$ we separate the integral into two parts:

$$\begin{aligned}\left|\frac{1}{|x|} *_x f\right| &\leq \int_{|x-x'|<R} \frac{|f(x')|}{|x-x'|}dx' + \frac{1}{R}\|f\|_{L^1(\mathbb{R}^2)} \\ &\leq CR^{\frac{p-2}{p}}\|f\|_{L^p(\mathbb{R}^2)} + \frac{1}{R}\|f\|_{L^1(\mathbb{R}^2)},\end{aligned}$$

where we used the Hölder's inequality to estimate the first integral. The value of θ is obtained after optimization of R. ∎

LEMMA B.2. (i) *Let* $f \in L^p_x L^1_z$ *with* $1 < p < 2$. *Then we have*

$$\left\|\frac{1}{r} * f\right\|_{p^\#,\infty} + \left\|\nabla_{x,z}\left(\frac{1}{r} * f\right)\right\|_{p^\#,1} \leq C_p \|f\|_{p,1}, \tag{B.3}$$

where $p^\# = \frac{2p}{2-p}$. *If in addition* $\nabla_x f \in L^p_x L^1_z$, *then*

$$\left\|\nabla_{x,z}\left(\frac{1}{r} * f\right)\right\|_{p^\#,\infty} \leq C_p \|\nabla_x f\|_{p,1}. \tag{B.4}$$

(ii) *Let* $f \in L^p_x L^1_z \cap L^1(\mathbb{R}^3)$ *with* $2 < p \leq +\infty$. *Then we have*

$$\left\|\frac{1}{r} * f\right\|_{L^\infty(\mathbb{R}^3)} + \left\|\nabla_{x,z}\left(\frac{1}{r} * f\right)\right\|_{\infty,1} \leq C_p \|f\|^\theta_{p,1} \|f\|^{1-\theta}_{L^1(\mathbb{R}^3)}, \tag{B.5}$$

where $\theta = \frac{p}{2p-2}$. *If in addition* $\nabla_x f \in L^p_x L^1_z \cap L^1(\mathbb{R}^3)$, *then*

$$\left\|\nabla_{x,z}\left(\frac{1}{r} * f\right)\right\|_{L^\infty(\mathbb{R}^3)} \leq C_p \|\nabla_x f\|^\theta_{p,1} \|\nabla_x f\|^{1-\theta}_{L^1(\mathbb{R}^3)}. \tag{B.6}$$

Proof. Items (i) and (ii) can be proved similarly by using, respectively, items (i) and (ii) of Lemma B.1. We shall only prove here item (i). Denoting $u = \frac{1}{r} * f$, we have

$$\|u(x,\cdot)\|_{L^\infty(\mathbb{R})} \leq \frac{1}{|x|} *_x \|f(x,\cdot)\|_{L^1(\mathbb{R})},$$

and the first part of (B.3) is a consequence of (B.1) since $x \mapsto \|f(x,\cdot)\|_{L^1(\mathbb{R})}$ belongs to $L^p(\mathbb{R}^2)$. Now we have

$$\begin{aligned}\int_{\mathbb{R}} |\nabla_x u(x,z)|\, dz &\leq \iiint_{\mathbb{R}^4} \frac{|x-x'|}{(|x-x'|^2 + (z-z')^2)^{3/2}} |f(x',z')|\, dx'\, dz'\, dz. \\ &= 2\int_{\mathbb{R}^2} \frac{1}{|x-x'|}\|f(x',\cdot)\|_{L^1(\mathbb{R})}\, dx' \\ &= \frac{2}{|x|} *_x \|f(x,\cdot)\|_{L^1(\mathbb{R})},\end{aligned}$$

where we have just evaluated the integral

$$\int_{\mathbb{R}} \frac{|x-x'|}{\left(|x-x'|^2+(z-z')^2\right)^{3/2}}\,dz = \frac{2}{|x-x'|}.$$

Then by again using (B.1) we conclude with the estimate of $\|\nabla_x u\|_{p^\#,1}$. We estimate $\|\partial_z u\|_{p^\#,1}$ similarly:

$$\begin{aligned}\int_{\mathbb{R}} |\partial_z u(x,z)|\,dz &\le \iiint_{\mathbb{R}^4} \frac{|z-z'|}{\left(|x-x'|^2+(z-z')^2\right)^{3/2}} |f(x',z')|\,dx'\,dz'\,dz. \\ &= 2\int_{\mathbb{R}^2} \frac{1}{|x-x'|}\|f(x',\cdot)\|_{L^1(\mathbb{R})}\,dx'.\end{aligned}$$

This proves (B.3). Next, for $i=1,2$ and to prove (B.4) we write

$$\|\nabla_{x,z}\,\partial_{x_i} u\|_{p^\#,1} \le \left\|\nabla_{x,z}\left(\frac{1}{r} * (\partial_{x_i} f)\right)\right\|_{p^\#,1} \le C_p\,\|\partial_{x_i} f\|_{p,1}. \tag{B.7}$$

Together with (B.3) this implies that $\partial_{x_i} u$ belongs to $L^{p^\#}(\mathbb{R}^2, W^{1,1}(\mathbb{R}))$. Note that $W^{1,1}(\mathbb{R}) \hookrightarrow L^\infty(\mathbb{R})$. Therefore, $\partial_{x_i} u$ is in $L_x^{p^\#} L_z^\infty$ and satisfies

$$\|\partial_{x_i} u\|_{p^\#,\infty} \le C\|\partial_z \partial_{x_i} u\|_{p^\#,1} \le C_p\,\|\partial_{x_i} f\|_{p,1}.$$

To prove (B.4), it remains to estimate $\partial_z u$. We recall that $-\Delta_{x,z} u = f$. Additionally, we remark that

$$x \mapsto \|f(x,\cdot)\|_{L^1(\mathbb{R})}$$

belongs to $W^{1,p}(\mathbb{R}^2) \hookrightarrow L^{p^\#}(\mathbb{R}^2)$. Consequently,

$$\|f\|_{p^\#,1} \le C_p \|\nabla_x f\|_{p,1},$$

and applying (B.7) we get

$$\|\partial_{zz} u\|_{p^\#,1} \le \|f\|_{p^\#,1} + \|\Delta_x u\|_{p^\#,1} \le C_p\,\|\nabla_x f\|_{p,1}.$$

Therefore, as above, $\partial_z u$ is bounded in $L^{p^\#}(\mathbb{R}^2, W^{1,1}(\mathbb{R}))$, and thus in $L_x^{p^\#} L_z^\infty$. ∎

LEMMA B.3. (i) *Let $f \in L_x^p L_z^1$, with $1 < p < \infty$ be such that $\int_{\mathbb{R}} f(x,z)dz = 0$, x a.e., and $z\,f \in L_x^p L_z^1$. Then for any $\alpha \in (0, \min(1, 2/p))$ we have*

$$\left\|\frac{1}{r} * f\right\|_{q,\infty} \le C\,\|zf\|_{p,1}^{1-\alpha}\|f\|_{p,1}^{\alpha}, \tag{B.8}$$

where $q = \frac{2p}{2-\alpha p}$.

(ii) *Let $f \in L_x^p L_z^1$, with $2 < p < \infty$ be such that $\int_{\mathbb{R}} f(x,z)dz = 0$, x a.e. and $z\,f \in L_x^p L_z^1$. Then*

$$\left\|\frac{1}{r} * f\right\|_{L^\infty(\mathbb{R}^3)} \le C\,\|zf\|_{p,1}^{1-2/p}\|f\|_{p,1}^{2/p}. \tag{B.9}$$

Proof. Denote $u = \frac{1}{r} * f$. Since $\int_{\mathbb{R}} f(x,z)dz = 0$, we have

$$u(x,z) = \iint_{\mathbb{R}^3} \left(\frac{1}{\left(|x-x'|^2 + (z-z')^2\right)^{1/2}} - \frac{1}{\left(|x-x'|^2 + z^2\right)^{1/2}} \right) f(x',z')\, dx'dz'$$

$$= \iint_{\mathbb{R}^3} \int_0^{z'} \left(\frac{z-\xi}{\left(|x-x'|^2 + (z-\xi)^2\right)^{3/2}} \right) f(x',z')\, d\xi dx'dz'. \tag{B.10}$$

Then we remark that for any z, ξ, and $x \neq x'$, we have

$$\frac{|z-\xi|}{\left(|x-x'|^2 + (z-\xi)^2\right)^{3/2}} \leq \frac{2}{3\sqrt{3}} \frac{1}{|x-x'|^2} \tag{B.11}$$

and that

$$\int_0^{z'} \frac{|z-\xi|}{\left(|x-x'|^2 + (z-\xi)^2\right)^{3/2}} d\xi \leq \int_{\mathbb{R}} \frac{|z-\xi|}{\left(|x-x'|^2 + (z-\xi)^2\right)^{3/2}} d\xi = \frac{2}{|x-x'|}. \tag{B.12}$$

Let us first prove (B.8). By (B.11) and (B.12) we have for any $0 \leq \alpha \leq 1$

$$\begin{aligned} \int_{\mathbb{R}} \int_0^{z'} \frac{|z-\xi|}{\left(|x-x'|^2 + (z-\xi)^2\right)^{3/2}} |f(x',z')| d\xi dz' &\leq C \int_{\mathbb{R}} \frac{|z' f(x',z')|^{1-\alpha}}{|x-x'|^{2(1-\alpha)}} \frac{|f(x',z')|^{\alpha}}{|x-x'|^{\alpha}} dz' \\ &\leq \frac{C}{|x-x'|^{2-\alpha}}\, g(x), \end{aligned}$$

where

$$g(x) = \left(\int |z f(x,z)| dz \right)^{1-\alpha} \left(\int |f(x,z)| dz \right)^{\alpha}.$$

Hence from (B.10) we deduce that

$$\|u(x,\cdot)\|_{L^\infty(\mathbb{R})} \leq C \left(\frac{1}{|x|^{2-\alpha}} * g \right)(x).$$

From the assumptions on f, we deduce that g belongs to $L^p(\mathbb{R}^2)$. Since the function $x \mapsto \frac{1}{|x|^{2-\alpha}}$ belongs to $L_w^{2/(2-\alpha)}(\mathbb{R}^2)$, the generalized Young's inequality gives (B.8).

In order to prove (B.9), the right-hand side of (B.10) is separated into two parts:

$$\int_{\mathbb{R}^3} = \int_{|x-x'|>R} + \int_{|x-x'|<R}.$$

By (B.11), the first part is controlled by

$$C \int_{|x-x'|>R} \frac{\|z' f(x',\cdot)\|_{L^1(\mathbb{R})}}{|x-x'|^2} dx' \leq \frac{C}{R^{2/p}} \|z' f\|_{p,1},$$

while the second integral is estimated, through (B.12), by

$$C \int_{|x-x'|<R} \frac{\|f(x',\cdot)\|_{L^1(\mathbb{R})}}{|x-x'|} dx' \leq R^{1-2/p} \|f\|_{p,1}.$$

Optimization of R leads to (B.9). ∎

REFERENCES

[1] T. Ando, B. Fowler, and F. Stern, *Electronic properties of two-dimensional systems*, Rev. Modern Phys., 54 (1982), pp. 437–672.

[2] G. Bastard, *Wave Mechanics Applied to Semiconductor Heterostructures*, Les Éditions de Physique, EDP Sciences, Les Ulis Cedex, France, 1992.

[3] N. Ben Abdallah and F. Méhats, *Semiclassical analysis of the Schrödinger equation with a partially confining potential*, J. Math. Pures Appl., to appear.

[4] F. Brezzi and P. A. Markowich, *The three dimensional Wigner-Poisson problem: Existence, uniqueness and approximation*, Math. Methods Appl. Sci., 14 (1991), pp. 35–61.

[5] F. Castella, *L^2 solutions to the Schrödinger-Poisson system: Existence, uniqueness, time behaviour, and smoothing effects*, Math. Models Methods Appl. Sci., 7 (1997), pp. 1051–1083.

[6] T. Cazenave and F. Weissler, *The Cauchy problem for the nonlinear Schrödinger equation in H^1*, Manuscripta Math., 61 (1988), pp. 477–494.

[7] T. Cazenave, *An Introduction to Nonlinear Schrödinger Equations*, 3rd ed., Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil, 1996.

[8] R. C. T. da Costa, *Quantum mechanics for a constraint particle*, Phys. Rev. A, 23 (1981), pp. 1982–1987.

[9] D. K. Ferry and S. M. Goodnick, *Transport in Nanostructures*, Cambridge University Press, Cambridge, UK, 1997.

[10] R. Froese and I. Herbst, *Realizing holonomic constraints in classical and quantum mechanics*, Comm. Math. Phys., 220 (2001), pp. 489–535.

[11] J. Ginibre and G. Velo, *The global Cauchy problem for the nonlinear Schrödinger equation revisited*, Ann. Inst. H. Poincaré, Anal. Non Linéaire, 2 (1985), pp. 309–327.

[12] G. A. Hagedorn and A. Joye, *A time-dependent Born-Oppenheimer approximation with exponentially small error estimates*, Comm. Math. Phys., 223 (2001), pp. 583–626.

[13] R. Illner, P. F. Zweifel, and H. Lange, *Global existence, uniqueness and asymptotic behaviour of solutions of the Wigner-Poisson and Schrödinger-Poisson systems*, Math. Methods Appl. Sci., 17 (1994), pp. 349-376.

[14] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, Heidelberg, 1966.

[15] K. A. Mitchell, *Geometric phase, Curvature, and Extrapotentials in Constrained Quantum Systems*, Preprint quant-ph/0001059, Cornell University, New York, 2000. Available online at http://www.arxiv.org/list/quant-ph/0001.

[16] O. Pinaud, *Adiabatic approximation of the Schrödinger-Poisson system with a partial confinement: The stationary case*, J. Math. Phys., 45 (2004), pp. 2029–2050.

[17] E. Polizzi and N. Ben Abdallah, *Self-consistent three dimensional model for quantum ballistic transport in open systems*, Phys. Rev. B, 66 (2002), pp. 245301–245309.

[18] M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, Academic Press, New York, San Francisco, London, 1975.

[19] H. Spohn and S. Teufel, *Adiabatic decoupling and time-dependent Born-Oppenheimer theory*, Comm. Math. Phys., 224 (2001), pp. 113–132.

[20] R. S. Strichartz, *Restriction of Fourier transform to quadratic surfaces and decay of solutions of wave equations*, Duke Math. J., 44 (1977), pp. 705–714.

[21] S. Teufel, *Adiabatic Perturbation Theory in Quantum Dynamics*, Lecture Notes in Math. 1821, Springer-Verlag, Berlin, Heidelberg, New York, 2003.

[22] K. Yajima, *Existence of solutions for Schrödinger evolution equations*, Comm. Math. Phys., 110 (1987), pp. 415–426.

SIAM J. MATH. ANAL.
Vol. 36, No. 3, pp. 1014–1030

GLOBAL SOLUTIONS OF THE 2D DISSIPATIVE QUASI-GEOSTROPHIC EQUATION IN BESOV SPACES*

JIAHONG WU†

Abstract. The two-dimensional (2D) quasi-geostrophic (QG) equation is a 2D model of the 3D incompressible Euler equations, and its dissipative version includes an extra term bearing the operator $(-\Delta)^\alpha$ with $\alpha \in [0,1]$. Existing research appears to indicate the criticality of $\alpha = \frac{1}{2}$ in the sense that the issue of global existence for the 2D dissipative QG equation becomes extremely difficult when $\alpha \le \frac{1}{2}$. It is shown here that for any $\alpha \le \frac{1}{2}$ the 2D dissipative QG equation with an initial datum in the Besov space $B^r_{2,\infty}$ or $B^r_{p,\infty}$ $(p > 2)$ possesses a unique global solution if the norm of the datum in these spaces is comparable to κ, the diffusion coefficient. Since the Sobolev space H^r is embedded in $B^r_{2,\infty}$, a special consequence is the global existence of small data solutions in H^r for any $r > 2 - 2\alpha$.

Key words. 2D quasi-geostrophic equation, Besov spaces, global existence

AMS subject classifications. 76U05, 76B03, 35Q35

DOI. 10.1137/S0036141003435576

1. Introduction.

This paper is concerned with global existence results for the two-dimensional (2D) dissipative quasi-geostrophic (QG) equation

$$\begin{cases} \partial_t \theta + u \cdot \nabla \theta + \kappa(-\Delta)^\alpha \theta = 0, \\ u = (u_1, u_2) = \nabla^\perp \psi, \quad (-\Delta)^{\frac{1}{2}} \psi = \theta \end{cases} \tag{1.1}$$

supplemented with the initial condition

$$\theta(x,0) = \theta_0(x). \tag{1.2}$$

In (1.1), $x \in \mathbb{R}^2$, $t \ge 0$, $\kappa > 0$ is the diffusion coefficient and $\alpha \in [0,1]$ is a parameter, $\theta = \theta(x,t)$ is a scalar representing the temperature, u is the velocity field, and ψ is the usual stream function. Besides its geophysical applications [11], [12], the 2D dissipative QG equation serves as a 2D model of the 3D Navier–Stokes equations and has recently been extensively investigated (see [1], [2], [3], [5], [6], [7], [8], [9], [10], [13], [14], [15], [16]).

Prior work on the issue of global existence concerning the 2D dissipative QG equation (1.1) appears to indicate that $\alpha = \frac{1}{2}$ is a critical index. In the subcritical case, namely, $\alpha > \frac{1}{2}$, solutions at several regularity levels, including solutions in the classical sense, have been shown to be global in time [7], [13], [16]. The theory of global existence and regularity for this case is thus in a satisfactory state. In the critical case $\alpha = \frac{1}{2}$, classical solutions are known to be global if their initial L^∞-norms are comparable to κ [6]. For initial data of arbitrary size, the global existence of classical solutions has not been established. It is hoped that the resolution of this problem will shed light on the millennium prize problem on the 3D Navier–Stokes equations. The supercritical case $\alpha < \frac{1}{2}$ seems even harder to deal with, and work on this case has just started to appear. For $\alpha \le \frac{1}{2}$, Chae and Lee [3] established a global existence

*Received by the editors September 29, 2003; accepted for publication (in revised form) May 28, 2004; published electronically January 5, 2005.
http://www.siam.org/journals/sima/36-3/43557.html
†Department of Mathematics, Oklahoma State University, Stillwater, OK 74078 (jiahong@math.okstate.edu).

result under the assumption that θ_0 is small in the Besov space $B^{2-2\alpha}_{2,1}$. In a recent work [9], A Córdoba and D. Córdoba obtained for any $\alpha \in [0,1]$ a local existence result for $\theta_0 \in H^s$ with $s+\alpha > 2$ and a global result for small θ_0 in H^s with $s > 2$ or in $H^{3/2}$ in the case of $\alpha = \frac{1}{2}$.

This paper is devoted to establishing global existence results for (1.1) with θ_0 in the Besov space $B^r_{2,\infty}$ or in $B^r_{p,\infty}$ with $p > 2$. For any $\alpha \le \frac{1}{2}$ and $\theta_0 \in B^r_{2,\infty}$ with $r > 2-2\alpha$, we prove that the 2D QG equation (1.1) has a unique global solution provided that the norm of θ_0 in $B^r_{2,\infty}$ is comparable to κ. Because of the embeddings $B^r_{2,1} \hookrightarrow H^r \hookrightarrow B^r_{2,\infty}$, a special consequence is the global existence result for small data in $B^r_{2,1}$ or H^r with $r > 2-2\alpha$. We defer the precise statement and many more details to section 3.

The situation for $\theta_0 \in B^r_{p,\infty}$ with $p > 2$ is more sophisticated and the major difficulty lies in how to obtain suitable lower bounds for terms generated from the dissipative part. Thanks to the L^p-decay estimate of A. Córdoba and D. Córdoba [9], we are able to establish a global existence result for solutions in the Besov space $B^r_{p,\infty}$ with $r > 1 + \frac{2}{p}$. Appropriate smallness conditions are imposed on the initial datum θ_0 here. This is accomplished in section 4, which consists of two subsections. The first subsection provides an a priori bound and the second proves the global existence result.

2. Preliminaries. This section provides a precise characterization of the Besov space $B^r_{p,q}$ through the Littlewood–Paley decomposition and gathers several important estimates involving $B^r_{p,\infty}$. First, we recall two commutator estimates established in a previous work [17]. Then follows the tame estimate for the usual product of two functions. Finally, a logarithmic bound for the L^∞-norm of a function in terms of its norms in Besov spaces is stated and proven. We shall also reproduce here the L^p-decay estimate of A. Córdoba and D. Córdoba for the dissipative QG equation [9].

We start with a dyadic decomposition of $\mathbb{R}^d$, where $d > 0$ is an integer. It is a classical result that there exist two radial functions $\chi \in C_0^\infty(\mathbb{R}^d)$ and $\phi \in C_0^\infty(\mathbb{R}^d \backslash \{0\})$ satisfying

$$\operatorname{supp}\chi \subset \{\xi : |\xi| \le 4/3\}, \qquad \operatorname{supp}\phi \subset \{\xi : 3/4 < |\xi| < 8/3\},$$

$$\chi(\xi) + \sum_{j\ge 0} \phi(2^{-j}\xi) = 1 \quad \text{for all } \xi \in \mathbb{R}^d.$$

For the purpose of isolating different Fourier frequencies, define the operators Δ_i for $i \in \mathbb{Z}$ as follows:

$$\Delta_i u = \begin{cases} 0 & \text{if } i \le -2, \\ \chi(D)u = \int h(y)u(x-y)dy & \text{if } i = -1, \\ \phi(2^{-i}D)u = 2^{id}\int g(2^i y)u(x-y)dy & \text{if } i \ge 0, \end{cases} \tag{2.1}$$

where $h = \chi^\vee$ and $g = \phi^\vee$ are the inverse Fourier transforms of χ and ϕ, respectively. We note that Δ_i in (2.1) can be defined in other ways. For example, by further requiring $\chi(\xi) = 1$ for $|\xi| \le \frac{3}{8}$ and writing

$$g(x) = 2^d h(2x) - h(x), \quad g_j(x) = 2^{dj} g(2^j x),$$

one can define $\Delta_{-1} = h*$ and $\Delta_j = g_j*$ for $j \ge 0$.

For $i \in \mathbb{Z}$, S_i is the sum of Δ_j with $j \le i-1$, i.e.,

$$S_i u = \Delta_{-1}u + \Delta_0 u + \Delta_1 u + \cdots + \Delta_{i-1}u = \int_{\mathbb{R}^d} h(2^i y)u(x-y)dy.$$

It can be shown for any tempered distribution f that $S_i f \to f$ in the distributional sense, as $i \to \infty$.

For any $r \in \mathbb{R}$ and $p, q \in [1, \infty]$, the Besov space $B^r_{p,q}$ consists of all tempered distributions f such that the sequence $\{2^{jr}\|\Delta_j f\|_{L^p}\}_{j\in\mathbb{Z}}$ belongs to $L^q(\mathbb{Z})$. In particular, $B^r_{p,\infty}$ contains any function f satisfying

$$\|f\|_{B^r_{p,\infty}} \equiv \sup_{j\in\mathbb{Z}} 2^{jr}\|\Delta_j f\|_{L^p} < \infty. \tag{2.2}$$

It is easy to check that $B^r_{p,\infty}$ endowed with the norm (2.2) is a Banach space.

The following version of Bernstein's lemma can be found in [4].

LEMMA 2.1 (Bernstein's lemma). *Let $d > 0$ be an integer and $R_2 > R_1 > 0$ be two real numbers. If $p \in [1, \infty]$ and $\operatorname{supp}\widehat{f} \subset \{\xi \in \mathbb{R}^d : R_1 2^j \le |\xi| \le R_2 2^j\}$, then*

$$C^{-1}2^{jk}\|f\|_{L^p(\mathbb{R}^d)} \le \max_{|\alpha|=k} \|\partial^\alpha f\|_{L^p(\mathbb{R}^d)} \le C2^{jk}\|f\|_{L^p(\mathbb{R}^d)},$$

where $C > 0$ is a constant depending on k, R_1, and R_2 only.

We now recall two commutator estimates previously established in [17].

PROPOSITION 2.2. *Let $j \ge -1$ be an integer, let $r \in \mathbb{R}$, and let $p \in [1, \infty]$. Then,*

$$\|[u\cdot\nabla, \Delta_j]\theta\|_{L^p} \le C2^{-jr}\left(\|\nabla\theta\|_{L^\infty}\|u\|_{B^r_{p,\infty}} + \|\nabla u\|_{L^\infty}\|\theta\|_{B^r_{p,\infty}}\right), \tag{2.3}$$

where C is a pure constant and the brackets $[\,,\,]$ represent the commutator, namely,

$$[u\cdot\nabla, \Delta_j]\theta = u\cdot\nabla(\Delta_j\theta) - \Delta_j(u\cdot\nabla\theta).$$

Inequality (2.3) is suitable for situations when u and θ are equally regular. If $\nabla\theta$ is not known to be bounded in L^∞, then (2.3) fails. The following proposition provides a new estimate which needs no information about $\nabla\theta$. As a trade-off, u is required to be in $B^{r+1}_{p,\infty}$. The importance of this estimate will be seen in the proofs of Theorems 3.1 and 4.1.

PROPOSITION 2.3. *Let $j \ge -1$, let $r \in \mathbb{R}$, and let $p \in [1, \infty]$. Then, for some pure constant C,*

$$\|[u\cdot\nabla, \Delta_j]\theta\|_{L^p} \le C2^{-jr}\left(\|\nabla u\|_{L^\infty}\|\theta\|_{B^r_{p,\infty}} + \|\theta\|_{L^\infty}\|u\|_{B^{r+1}_{p,\infty}}\right). \tag{2.4}$$

Estimates for the product uv of two functions u and v are handy in dealing with the quadratic nonlinear term in many partial differential equations. In the context of Besov spaces, we have the following estimate.

PROPOSITION 2.4. *Let $r > 0$ be a real number and let $p \in [1, \infty]$. Then*

$$\|u\,v\|_{B^r_{p,\infty}} \le C\big(\|u\|_{L^\infty}\|v\|_{B^r_{p,\infty}} + \|u\|_{B^r_{p,\infty}}\|v\|_{L^\infty}\big),$$

where C is constant depending on r and p only.

In the course of establishing existence results for the QG equation, very often we need to bound the L^∞-norm of a function in terms of its norm in $B^r_{p,\infty}$. The following logarithmic estimate is very helpful.

PROPOSITION 2.5. *Let $p \in [1,\infty]$, let $r_c = \frac{d}{p}$, and let $r > \frac{d}{p}$. Then there exists a constant C depending on p and r only such that*

$$\|f\|_{L^\infty(\mathbb{R}^d)} \le C\|f\|_{B^{r_c}_{p,\infty}(\mathbb{R}^d)} \log_2\left(e + \frac{\|f\|_{B^r_{p,\infty}(\mathbb{R}^d)}}{\|f\|_{B^{r_c}_{p,\infty}(\mathbb{R}^d)}}\right), \tag{2.5}$$

which, in particular, implies

$$\|f\|_{L^\infty(\mathbb{R}^d)} \le C\|f\|_{B^r_{p,\infty}(\mathbb{R}^d)}. \tag{2.6}$$

Proof. According to the definition of Δ_i's in (2.1), $\Delta_k\Delta_j = 0$ if $|k-j| \ge 2$. For $j \ge 0$,

$$\begin{aligned}\|\Delta_j f\|_{L^\infty} &\le \sum_{|k-j|\le 1} \|\Delta_k\Delta_j f\|_{L^\infty} = \sum_{|k-j|\le 1} \|2^{kd} g(2^k\cdot) * (\Delta_j f)\|_{L^\infty} \\ &\le \sum_{|k-j|\le 1} \|2^{kd} g(2^k\cdot)\|_{L^q} \|\Delta_j f\|_{L^p} = \sum_{|k-j|\le 1} 2^{kd\,\frac{1}{p}} \|g\|_{L^q}\, \|\Delta_j f\|_{L^p},\end{aligned}$$

where q is the conjugate of p, or $1/p + 1/q = 1$. Thus,

$$\|\Delta_j f\|_{L^\infty} \le C\, 2^{jr_c}\, \|\Delta_j f\|_{L^p}$$

for a pure constant C. A similar estimate for the case $j = -1$ leads to the same bound. Using this bound, we have

$$\begin{aligned}\|f\|_{L^\infty} &\le \sum_{j\ge -1} \|\Delta_j f\|_{L^\infty} = \sum_{j=-1}^{N-1} \|\Delta_j f\|_{L^\infty} + \sum_{j\ge N} \|\Delta_j f\|_{L^\infty} \\ &\le C\,(N+1)\|f\|_{B^{r_c}_{p,\infty}} + C\,\|f\|_{B^r_{p,\infty}} \sum_{j\ge N} 2^{-j(r-r_c)} \\ &= C(N+1)\|f\|_{B^{r_c}_{p,\infty}} + C\,\frac{2^{-N(r-r_c)}}{1-2^{-(r-r_c)}}\,\|f\|_{B^r_{p,\infty}}.\end{aligned}$$

The desired inequality (2.5) is then obtained by letting

$$N = 1 + \left[\frac{1}{r-r_c}\log_2 \frac{\|f\|_{B^r_{p,\infty}}}{\|f\|_{B^{r_c}_{p,\infty}}}\right].$$

Inequality (2.6) is true because of (2.5) and the fact that $x \to x\log_2(e + M/x)$ with a fixed constant M is an increasing function for $x > 0$. This completes the proof. ∎

As seen in (1.1) of the introduction, the components of the velocity field u are Riesz transforms of θ, namely,

$$u = \mathcal{R}^\perp(\theta) \equiv (-\mathcal{R}_2(\theta), \mathcal{R}_1(\theta)),$$

where $\mathcal{R}_k = \partial_{x_k}\Lambda^{-1}$ for $k = 1, 2$ and $\Lambda \equiv (-\Delta)^{1/2}$. It is a classical result in the Calderon–Zygmund theory that for any $p \in (1,\infty)$ and $r \in \mathbb{R}$

$$\|u\|_{B^r_{p,\infty}} \le C\|\theta\|_{B^r_{p,\infty}}, \tag{2.7}$$

where C is a constant depending only on p and r.

Finally, we recall the L^p-decay result of A. Córdoba and D. Córdoba. In a recent work [9], A. Córdoba and D. Córdoba skillfully proved a pointwise inequality involving the operator $\Lambda^{2\alpha}$ with $\alpha \in [0,1]$ and then derived the L^p-decay result as a special consequence.

PROPOSITION 2.6. *Let $\alpha \in [0,1]$ and let $\theta \in \mathbb{S}$, the Schwartz class. Then,*

$$2\theta\,\Lambda^{2\alpha}\theta(x) \geq \Lambda^{2\alpha}\,\theta^2(x)$$

for any $x \in \mathbb{R}^2$.

The estimate in the following proposition is slightly different from the corresponding L^p-decay result derived in [9].

PROPOSITION 2.7. *Let $p = 2^k$ for an integer $k \geq 1$. If θ solves (1.1) with an initial data $\theta_0 \in L^p$, then the L^p-norm of θ decays algebraically in time. More precisely,*

$$\|\theta(\cdot,t)\|_{L^p} \leq \frac{\|\theta_0\|_{L^p}}{\left(1+\kappa C_p \gamma t \|\theta_0\|_{L^2}^{-\gamma p}\|\theta_0\|_{L^p}^{\gamma p}\right)^{\frac{1}{\gamma p}}},$$

where $\gamma = \frac{\alpha}{p-2}$ and C_p is a constant depending on p and α only.

3. Global existence in $B_{2,\infty}^r$. We shall assume in this section that θ_0 is in the Besov space $B^r_{2,\infty}$. Consider the solution of the 2D dissipative QG equation

$$\partial_t\theta + u\cdot\nabla\theta + \kappa\Lambda^{2\alpha}\theta = 0, \quad u = \mathcal{R}^{\perp}(\theta) \tag{3.1}$$

with $\theta(x,0) = \theta_0(x)$. Assuming $r > 2-2\alpha$, our major result states that (3.1) has a unique global solution if the norm of θ_0 in $B^r_{2,\infty}$ is comparable to κ.

THEOREM 3.1. *Let $\kappa > 0$ and let $0 \leq \alpha \leq \frac{1}{2}$. Assume the initial datum θ_0 is in the Besov space $B^r_{2,\infty}$ with $r > 2-2\alpha$. There exists a constant C_0 depending on α and r only such that if*

$$\|\theta_0\|_{B^r_{2,\infty}} \leq C_0\kappa, \tag{3.2}$$

then the 2D dissipative QG equation (3.1) with $\theta(x,0) = \theta_0(x)$ has a unique global solution θ satisfying

$$\theta \in L^\infty([0,\infty); B^r_{2,\infty}) \cap L^1([0,\infty); B^{r+2\alpha}_{2,\infty}) \cap \mathrm{Lip}([0,\infty); B^{r-1}_{2,\infty}) \cap C([0,\infty); B^{\delta}_{2,\infty})$$

for any $\delta \in [r-1, r)$, and

$$\|\theta(\cdot,t)\|_{B^r_{2,\infty}} \leq C_0\kappa \quad \text{for any } t \geq 0.$$

Remark. Because of the embeddings

$$B^s_{2,1} \hookrightarrow H^s_2 \hookrightarrow B^s_{2,\infty},$$

this theorem also implies that (3.1) has global solutions for small data in $B^s_{2,1}$ or H^s with any $s > 2-2\alpha$.

Before proving Theorem 3.1, we first establish an a priori estimate.

PROPOSITION 3.2. *Assume that θ solves the 2D dissipative QG equation (3.1) with $\kappa > 0$ and $0 \leq \alpha \leq 1$. Let $r \in \mathbb{R}$ and let $s > 2$. Then*

$$\frac{d}{dt}\|\theta\|_{B^r_{2,\infty}} + C_1\,\kappa\|\theta\|_{B^{r+2\alpha}_{2,\infty}} \leq C_2\,\|\theta\|_{B^s_{2,\infty}}\|\theta\|_{B^r_{2,\infty}}, \tag{3.3}$$

where C_1 and C_2 are constants depending on r only.

If $r > 2 - 2\alpha$, we can choose $s = r + 2\alpha$. Then, (3.3) reduces to

$$\frac{d}{dt}\|\theta\|_{B^r_{2,\infty}} + C_1\,\kappa\|\theta\|_{B^{r+2\alpha}_{2,\infty}} \le C_2\,\|\theta\|_{B^{r+2\alpha}_{2,\infty}}\|\theta\|_{B^r_{2,\infty}}.$$

This inequality bears two consequences, which we state as a corollary.

COROLLARY 3.3. *Assume that θ solves the 2D dissipative QG equation* (3.1) *with $\kappa > 0$ and $0 \le \alpha \le 1$. Let $r > 2 - 2\alpha$ be a real number. There exists a constant C_0 depending on α and r only such that if*

$$\|\theta_0\|_{B^r_{2,\infty}} \le C_0\kappa,$$

then, for any $t \ge 0$,

$$\|\theta(\cdot,t)\|_{B^r_{2,\infty}} \le \|\theta_0\|_{B^r_{2,\infty}} \le C_0\kappa.$$

In addition, θ also satisfies the inequality

$$\|\theta(\cdot,t)\|_{B^r_{2,\infty}} + C_1\,\kappa\int_0^t \|\theta(\cdot,\tau)\|_{B^{r+2\alpha}_{2,\infty}}d\tau \le \|\theta_0\|_{B^r_{2,\infty}}\exp\left(C_2\int_0^t \|\theta(\cdot,\tau)\|_{B^{r+2\alpha}_{2,\infty}}d\tau\right).$$

Proof of Proposition 3.2. Let $j \ge -1$. Applying Δ_j to (3.1), we obtain

$$\partial_t\Delta_j\theta + u\cdot\nabla\Delta_j\theta + \kappa\,\Lambda^{2\alpha}\Delta_j\theta = [u\cdot\nabla,\Delta_j]\theta.$$

Multiplying both sides by $2\Delta_j\theta$ and integrating over $\mathbb{R}^2$ yields

$$\frac{d}{dt}\int|\Delta_j\theta|^2dx + 2\kappa\int|\Lambda^\alpha\Delta_j\theta|^2dx = 2\int\Delta_j\theta\,[u\cdot\nabla,\Delta_j]\theta\,dx.$$

Applying Lemma 2.1 to the dissipative term and Hölder's inequality to the right-hand side, we find that

$$\frac{d}{dt}\|\Delta_j\theta\|_{L^2} + C\kappa 2^{2\alpha j}\|\Delta_j\theta\|_{L^2} \le \|[u\cdot\nabla,\Delta_j]\theta\|_{L^2}.$$

In the above inequality, we have used the fact that Lemma 2.1 is valid for fractional derivatives when $p = 2$. For any $r \in \mathbb{R}$, Proposition 2.2 applied to the term on the right-hand side yields

$$\frac{d}{dt}\|\theta\|_{B^r_{2,\infty}} + C\kappa\|\theta\|_{B^{r+2\alpha}_{2,\infty}} \le C\big(\|\nabla u\|_{L^\infty}\|\theta\|_{B^r_{2,\infty}} + \|\nabla\theta\|_{L^\infty}\|u\|_{B^r_{2,\infty}}\big). \tag{3.4}$$

Furthermore, Proposition 2.5 applied to ∇u and $\nabla\theta$ asserts that for any $s > 2$,

$$\|\nabla u\|_{L^\infty} \le C\|u\|_{B^s_{2,\infty}}, \quad \|\nabla\theta\|_{L^\infty} \le C\|\theta\|_{B^s_{2,\infty}}.$$

Inserting these estimates in (3.4) and noticing (2.7), we obtain

$$\frac{d}{dt}\|\theta\|_{B^r_{2,\infty}} + C_1\,\kappa\|\theta\|_{B^{r+2\alpha}_{2,\infty}} \le C_2\,\|\theta\|_{B^s_{2,\infty}}\|\theta\|_{B^r_{2,\infty}}. \qquad \square$$

Proof of Theorem 3.1. We start with a successive approximation sequence $\{\theta^{(n)}\}$ satisfying

$$\begin{cases} \theta^{(1)} = S_2\theta_0, \\ \partial_t\theta^{(n+1)} + u^{(n)}\cdot\nabla\theta^{(n+1)} + \kappa\Lambda^{2\alpha}\theta^{(n+1)} = 0, \\ u^{(n)} = \mathcal{R}^{\perp}(\theta^{(n)}), \\ \theta^{(n+1)}(x,0) = \theta_0^{(n+1)}(x) = S_{n+2}\theta_0. \end{cases}$$

The rest of the proof is divided into two major parts. The first part establishes that $\{\theta^{(n)}\}$ is bounded uniformly in $L^\infty([0,\infty); B^r_{2,\infty})$. The second part verifies that $\{\theta^{(n)}\}$ is a Cauchy sequence in $L^\infty([0,\infty); B^{r-1}_{2,\infty})$.

Noticing that $r > 2-2\alpha$, we proceed as in the proof of Proposition 3.2 to obtain

$$\begin{aligned} \frac{d}{dt}\|\theta^{(n+1)}\|_{B^r_{2,\infty}} &+ C_1\kappa\|\theta^{(n+1)}\|_{B^{r+2\alpha}_{2,\infty}} \\ &\le C\big(\|\nabla u^{(n)}\|_{L^\infty}\|\theta^{(n+1)}\|_{B^r_{2,\infty}} + \|\nabla\theta^{(n+1)}\|_{L^\infty}\|u^{(n)}\|_{B^r_{2,\infty}}\big) \\ &\le C\big(\|u^{(n)}\|_{B^{r+2\alpha}_{2,\infty}}\|\theta^{(n+1)}\|_{B^r_{2,\infty}} + \|\theta^{(n+1)}\|_{B^{r+2\alpha}_{2,\infty}}\|u^{(n)}\|_{B^r_{2,\infty}}\big) \\ &\le C_2\big(\|\theta^{(n)}\|_{B^{r+2\alpha}_{2,\infty}}\|\theta^{(n+1)}\|_{B^r_{2,\infty}} + \|\theta^{(n+1)}\|_{B^{r+2\alpha}_{2,\infty}}\|\theta^{(n)}\|_{B^r_{2,\infty}}\big), \end{aligned} \tag{3.5}$$

where C_1 and C_2 are constants with dependence on α and r only. Now, we choose $C_0 < C_1/(4C_2)$. Further restrictions will be imposed on C_0 in the second part. We show that if

$$\|\theta_0\|_{B^r_{2,\infty}} \le C_0\,\kappa,$$

then for any integer n and any $t \ge 0$,

$$\sup_{\tau\in[0,t]}\|\theta^{(n)}(\cdot,\tau)\|_{B^r_{2,\infty}} + C_1\,\kappa\int_0^t\|\theta^{(n)}(\cdot,\tau)\|_{B^{r+2\alpha}_{2,\infty}}d\tau \le 2C_0\,\kappa. \tag{3.6}$$

We proceed by induction. If (3.6) holds for $n = k$, namely,

$$\|\theta^{(k)}(\cdot,t)\|_{B^r_{2,\infty}} + C_1\,\kappa\int_0^t\|\theta^{(k)}(\cdot,\tau)\|_{B^{r+2\alpha}_{2,\infty}}d\tau \le 2C_0\,\kappa,$$

then, according to (3.5),

$$\begin{aligned} &\sup_{\tau\in[0,t]}\|\theta^{(k+1)}(\cdot,\tau)\|_{B^r_{2,\infty}} + C_1\,\kappa\int_0^t\|\theta^{(k+1)}(\cdot,\tau)\|_{B^{r+2\alpha}_{2,\infty}}d\tau \\ &\quad\le \|\theta_0^{(k+1)}\|_{B^r_{2,\infty}} + C_2\sup_{\tau\in[0,t]}\|\theta^{(k+1)}(\cdot,\tau)\|_{B^r_{2,\infty}}\int_0^t\|\theta^{(k)}(\cdot,\tau)\|_{B^{r+2\alpha}_{2,\infty}}d\tau \\ &\qquad + C_2\sup_{\tau\in[0,t]}\|\theta^{(k)}(\cdot,\tau)\|_{B^r_{2,\infty}}\int_0^t\|\theta^{(k+1)}(\cdot,\tau)\|_{B^{r+2\alpha}_{2,\infty}}d\tau \\ &\quad\le \|\theta_0\|_{B^r_{2,\infty}} + \frac{2C_0C_2}{C_1}\sup_{\tau\in[0,t]}\|\theta^{(k+1)}(\cdot,\tau)\|_{B^r_{2,\infty}} \\ &\qquad + 2C_0C_2\kappa\int_0^t\|\theta^{(k+1)}(\cdot,\tau)\|_{B^{r+2\alpha}_{2,\infty}}d\tau. \end{aligned}$$

Since $2C_0C_2 \le \frac{1}{2}C_1$, the inequality above becomes

$$\sup_{\tau\in[0,t]} \|\theta^{(k+1)}(\cdot,\tau)\|_{B^r_{2,\infty}} + C_1\,\kappa \int_0^t \|\theta^{(k+1)}(\cdot,\tau)\|_{B^{r+2\alpha}_{2,\infty}} d\tau \le 2\|\theta_0\|_{B^r_{2,\infty}} \le 2C_0\kappa.$$

Thus, (3.6) is verified. This completes the first part of the proof.

Next, we consider the difference

$$\eta^{(n+1)} = \theta^{(n+1)} - \theta^{(n)}.$$

The sequence $\{\eta^{(n)}\}$ satisfies

$$\begin{cases} \eta^{(1)} = S_2\theta_0 - \theta_0, \\ \partial_t\eta^{(n+1)} + u^{(n)}\cdot\nabla\eta^{(n+1)} + \kappa\Lambda^{2\alpha}\eta^{(n+1)} = w^{(n)}\cdot\nabla\theta^{(n)}, \\ w^{(n)} = \mathcal{R}^{\perp}(\eta^{(n)}), \\ \eta^{(n+1)}(x,0) = \eta_0^{(n+1)}(x) = \Delta_{n+1}\theta_0. \end{cases}$$

Starting with the equation for $\eta^{(n+1)}$ and proceeding as above, we are led to the following inequality:

$$\begin{aligned} &\frac{d}{dt}\|\eta^{(n+1)}\|_{B^{r-1}_{2,\infty}} + \ C_1\kappa\|\eta^{(n+1)}\|_{B^{r-1+2\alpha}_{2,\infty}} \\ &\quad \le 2^{(r-1)j}\|[u^{(n)}\cdot\nabla,\Delta_j]\eta^{(n+1)}\|_{L^2} + \|w^{(n)}\cdot\nabla\theta^{(n)}\|_{B^{r-1}_{2,\infty}}. \end{aligned} \tag{3.7}$$

Applying Propositions 2.3 and 2.5 to the first term on the right leads to

$$\begin{aligned} &2^{(r-1)j}\|[u^{(n)}\cdot\nabla,\Delta_j]\eta^{(n+1)}\|_{L^2} \\ &\quad \le C\big(\|\nabla u^{(n)}\|_{L^\infty}\|\eta^{(n+1)}\|_{B^{r-1}_{2,\infty}} + \|\eta^{(n+1)}\|_{L^\infty}\|u^{(n)}\|_{B^r_{2,\infty}}\big) \\ &\quad \le C\left(\|u^{(n)}\|_{B^{r+2\alpha}_{2,\infty}}\|\eta^{(n+1)}\|_{B^{r-1}_{2,\infty}} + \|\eta^{(n+1)}\|_{B^{r-1+2\alpha}_{2,\infty}}\|u^{(n)}\|_{B^r_{2,\infty}}\right) \\ &\quad \le C\left(\|\theta^{(n)}\|_{B^{r+2\alpha}_{2,\infty}}\|\eta^{(n+1)}\|_{B^{r-1}_{2,\infty}} + \|\eta^{(n+1)}\|_{B^{r-1+2\alpha}_{2,\infty}}\|\theta^{(n)}\|_{B^r_{2,\infty}}\right). \end{aligned}$$

Since $\alpha \le \frac{1}{2},\ r-1 > 1-2\alpha > 0$ and the same estimate in Proposition 2.4 applies. Consequently,

$$\begin{aligned} \|w^{(n)}\cdot\nabla\theta^{(n)}\|_{B^{r-1}_{2,\infty}} &\le C\big(\|w^{(n)}\|_{L^\infty}\|\nabla\theta^{(n)}\|_{B^{r-1}_{2,\infty}} + \|w^{(n)}\|_{B^{r-1}_{2,\infty}}\|\nabla\theta^{(n)}\|_{L^\infty}\big) \\ &\le C\big(\|w^{(n)}\|_{B^{r-1+2\alpha}_{2,\infty}}\|\theta^{(n)}\|_{B^r_{2,\infty}} + \|w^{(n)}\|_{B^{r-1}_{2,\infty}}\|\theta^{(n)}\|_{B^{r+2\alpha}_{2,\infty}}\big) \\ &\le C\big(\|\eta^{(n)}\|_{B^{r-1+2\alpha}_{2,\infty}}\|\theta^{(n)}\|_{B^r_{2,\infty}} + \|\eta^{(n)}\|_{B^{r-1}_{2,\infty}}\|\theta^{(n)}\|_{B^{r+2\alpha}_{2,\infty}}\big). \end{aligned}$$

Inserting these estimates in (3.7) yields

$$\begin{aligned} &\frac{d}{dt}\|\eta^{(n+1)}\|_{B^{r-1}_{2,\infty}} + C_1\kappa\|\eta^{(n+1)}\|_{B^{r-1+2\alpha}_{2,\infty}} \\ &\quad \le C_3\big(\|\theta^{(n)}\|_{B^{r+2\alpha}_{2,\infty}}\|\eta^{(n+1)}\|_{B^{r-1}_{2,\infty}} + \|\eta^{(n+1)}\|_{B^{r-1+2\alpha}_{2,\infty}}\|\theta^{(n)}\|_{B^r_{2,\infty}}\big) \\ &\qquad + C_3\big(\|\eta^{(n)}\|_{B^{r-1+2\alpha}_{2,\infty}}\|\theta^{(n)}\|_{B^r_{2,\infty}} + \|\eta^{(n)}\|_{B^{r-1}_{2,\infty}}\|\theta^{(n)}\|_{B^{r+2\alpha}_{2,\infty}}\big). \end{aligned}$$

Integrating over $[0,t]$, we obtain

$$\sup_{\tau\in[0,t]}\|\eta^{(n+1)}(\cdot,\tau)\|_{B^{r-1}_{2,\infty}}+C_1\kappa\int_0^t\|\eta^{(n+1)}(\cdot,\tau)\|_{B^{r-1+2\alpha}_{2,\infty}}\,d\tau$$

$$\le\|\theta_0^{(n+1)}\|_{B^{r-1}_{2,\infty}}+C_3\left(\sup_{\tau\in[0,t]}\|\eta^{(n+1)}\|_{B^{r-1}_{2,\infty}}+\sup_{\tau\in[0,t]}\|\eta^{(n)}\|_{B^{r-1}_{2,\infty}}\right)\int_0^t\|\theta^{(n)}\|_{B^{r+2\alpha}_{2,\infty}}d\tau$$

$$+C_3\sup_{\tau\in[0,t]}\|\theta^{(n)}\|_{B^r_{2,\infty}}\int_0^t\left(\|\eta^{(n)}\|_{B^{r-1+2\alpha}_{2,\infty}}+\|\eta^{(n+1)}\|_{B^{r-1+2\alpha}_{2,\infty}}\right)d\tau. \tag{3.8}$$

We now show by induction that for any $t\ge 0$,

$$\sup_{\tau\in[0,t]}\|\eta^{(n)}(\cdot,\tau)\|_{B^{r-1}_{2,\infty}}+C_1\kappa\int_0^t\|\eta^{(n)}(\cdot,\tau)\|_{B^{r-1+2\alpha}_{2,\infty}}\,d\tau\le\|\theta_0\|_{B^r_{2,\infty}}\,2^{-(n-3)}. \tag{3.9}$$

First, we notice that

$$\|\theta_0^{(n+1)}\|_{B^{r-1}_{2,\infty}}=\|\Delta_{n+1}\theta_0\|_{B^{r-1}_{2,\infty}}\le\|\theta_0\|_{B^r_{2,\infty}}\,2^{-n}.$$

Now, we require that C_0 further satisfy

$$2C_0C_3/C_1\le 1/4.$$

According to (3.6), we have the uniform bounds

$$C_3\sup_{\tau\in[0,t]}\|\theta^{(n)}(\cdot,\tau)\|_{B^r_{2,\infty}}\le 2C_0C_3\kappa,\quad C_3\int_0^t\|\theta^{(n)}(\cdot,\tau)\|_{B^{r+2\alpha}_{2,\infty}}\,d\tau\le 2C_0C_3/C_1.$$

If (3.9) is satisfied by $n=k$, then it follows from (3.8) that

$$\begin{aligned}&\frac{3}{4}\left(\max_{\tau\in[0,t]}\|\eta^{(k+1)}(\cdot,\tau)\|_{B^{r-1}_{2,\infty}}+C_1\kappa\int_0^t\|\eta^{(k+1)}(\cdot,\tau)\|_{B^{r-1+2\alpha}_{2,\infty}}\,d\tau\right)\\&\le\|\theta_0\|_{B^r_{2,\infty}}\,2^{-k}+\frac{1}{4}\left(\max_{\tau\in[0,t]}\|\eta^{(k)}(\cdot,\tau)\|_{B^{r-1}_{2,\infty}}+C_1\kappa\int_0^t\|\eta^{(k)}(\cdot,\tau)\|_{B^{r-1+2\alpha}_{2,\infty}}\,d\tau\right)\\&\le 3\|\theta_0\|_{B^r_{2,\infty}}\,2^{-k}.\end{aligned}$$

Thus, (3.9) is true for $n=k+1$. In other words, $\{\eta^{(n)}\}=\{\theta^{(n)}-\theta^{(n-1)}\}$ is a Cauchy sequence in $L^\infty([0,\infty);B^{r-1}_{2,\infty})$.

Therefore, there exists a $\theta\in L^\infty([0,\infty);B^r_{2,\infty})\cap L^1([0,\infty);B^{r+2\alpha}_{2,\infty})$ such that

$$\theta^{(n)}\to\theta\quad\text{in}\quad L^\infty([0,\infty);B^{r-1}_{2,\infty})\cap L^1([0,\infty);B^{r-1+2\alpha}_{2,\infty}).$$

Furthermore, for $0\le\alpha\le\frac{1}{2}$,

$$\begin{aligned}\|\partial_t\theta^{(n)}(\cdot,t)\|_{B^{r-1}_{2,\infty}}&\le\|u^{(n-1)}\cdot\nabla\theta^{(n)}(\cdot,t)\|_{B^{r-1}_{2,\infty}}+\kappa\|\Lambda^{2\alpha}\theta^{(n)}(\cdot,t)\|_{B^{r-1}_{2,\infty}}\\&\le C_3\|\theta^{(n)}(\cdot,t)\|_{B^r_{2,\infty}}\|\theta^{(n-1)}(\cdot,t)\|_{B^{r-1}_{2,\infty}}+\kappa\|\theta^{(n)}(\cdot,t)\|_{B^r_{2,\infty}}\\&\le C_3(C_0\kappa)^2+C_0\kappa^2=(C_3C_0+1)C_0\kappa^2.\end{aligned}$$

Therefore, $\theta \in \mathrm{Lip}([0,\infty); B^{r-1}_{2,\infty})$. Another consequence is $\theta \in C([0,\infty); B^{\delta}_{2,\infty})$ for any $\delta \in [r-1, r)$. Finally, the a priori estimates in Proposition 3.2 and Corollary 3.3 allow us to conclude that

$$\|\theta(\cdot,t)\|_{B^r_{2,\infty}} \le C_0\kappa.$$

This completes the proof. □

4. Global existence in $B_{,\infty}$ with $p > 2$. Attention is now turned to the 2D dissipative QG equation

$$\partial_t\theta + u\cdot\nabla\theta + \kappa\Lambda^{2\alpha}\theta = 0 \tag{4.1}$$

with $\theta(x,0) = \theta_0(x)$ in the Besov space $B^r_{p,\infty}$. We have the following theorem.

THEOREM 4.1. *Let $\kappa > 0$ and let $0 \le \alpha \le \frac{1}{2}$. Consider the solution of the dissipative QG equation* (4.1) *corresponding to $\theta_0 \in B^r_{2,\infty} \cap B^r_{p,\infty}$ with $p = 2^N (N > 1)$. Assume that*

$$\begin{cases} r > 1 + \frac{2}{p} \text{ and } \|\theta_0\|_{B^r_{2,\infty}} \le C_0\kappa \quad \text{if } (1-2\alpha)p \le 2, \\ r > 2 - 2\alpha,\ \|\theta_0\|_{B^r_{2,\infty}} \le C_0\kappa, \text{ and } \|\theta_0\|_{B^r_{p,\infty}} \le C_0\kappa \quad \text{if } (1-2\alpha)p > 2, \end{cases} \tag{4.2}$$

where C_0 is a suitably chosen constant with dependence on α, r, and p only. Then the 2D QG equation (4.1) *has a unique global solution θ satisfying*

$$\theta \in L^\infty([0,\infty); B^r_{p,\infty}) \cap L^1([0,\infty); B^{r+2\alpha}_{p,\infty}) \cap \mathrm{Lip}([0,\infty); B^{r-1}_{p,\infty}) \cap C([0,\infty); B^{\delta}_{p,\infty})$$

for any $\delta \in [r-1, r)$, and

$$\|\theta(\cdot,t)\|_{B^r_{p,\infty}} \le \max\{\|\theta_0\|_{B^r_{p,\infty}}, \tilde{C}_0\kappa\}$$

for any $t \ge 0$ and some constant $\tilde{C}_0$ depending on α, r, and p only.

The rest of this section revolves around the proof of Theorem 4.1 and is divided into two subsections. The first subsection presents an a priori estimate and the second subsection proves Theorem 4.1.

4.1. An a priori bound. We state and prove a global a priori bound.

PROPOSITION 4.2. *Assume that θ solves the 2D dissipative QG equation* (4.1) *with $\kappa > 0$ and $0 \le \alpha \le 1$. Let $r \in \mathbb{R}$, let $p = 2^N$ for an integer $N > 1$, and let $s > 1 + \frac{2}{p}$. Then*

$$\frac{d}{dt}\|\theta\|_{B^r_{p,\infty}} + C_4\, p^{-1}\,\kappa\,\|\theta\|^{1+\beta p}_{B^r_{p,\infty}}\,\|\theta\|^{-\beta p}_{B^r_{2,\infty}} \le C_5\,\|\theta\|_{B^r_{p,\infty}}\,\|\theta\|_{B^s_{p,\infty}}, \tag{4.3}$$

where $\beta = \frac{2\alpha}{p-2}$, and C_4 and C_5 are constants with possible dependence on α and p only.

Remark. The case $p = 2$ is excluded here since this case has been dealt with in the previous section. The assumption $p = 2^N$ is made in order to use Proposition 2.7.

Proof of Proposition 4.2. Applying Δ_j to (4.1), multiplying by $p|\Delta_j\theta|^{p-2}\Delta_j\theta$, and integrating over $\mathbb{R}^2$, we obtain

$$\frac{d}{dt}\|\Delta_j\theta\|^p_{L^p} + I = II, \tag{4.4}$$

where I and II represent the terms

$$I = \kappa p \int |\Delta_j\theta|^{p-2}(\Delta_j\theta)\Lambda^{2\alpha}\Delta_j\theta\, dx,$$

$$II = p \int |\Delta_j\theta|^{p-2}(\Delta_j\theta)\,[u\cdot\nabla, \Delta_j]\theta dx.$$

To estimate II, we first apply Hölder's inequality and then Proposition 2.2 to obtain

$$\begin{aligned} |II| &\le p\,\|\Delta_j\theta\|_{L^p}^{p-1}\,\|[u\cdot\nabla,\Delta_j]\theta\|_{L^p} \\ &\le p\,\|\Delta_j\theta\|_{L^p}^{p-1}\,\big[2^{-jr}\big(\|\nabla u\|_{L^\infty}\|\theta\|_{B^r_{p,\infty}} + \|\nabla\theta\|_{L^\infty}\,\|u\|_{B^r_{p,\infty}}\big)\big]. \end{aligned}$$

For $s > 1+\frac{2}{p}$, Proposition 2.5 asserts that

$$\|\nabla\theta\|_{L^\infty} \le C\,\|\theta\|_{B^s_{p,\infty}}, \quad \|\nabla u\|_{L^\infty} \le C\,\|u\|_{B^s_{p,\infty}} \le C\,\|\theta\|_{B^s_{p,\infty}}.$$

Therefore, for some constant C,

$$|II| \le C\,p\,2^{-jr}\,\|\Delta_j\theta\|_{L^p}^{p-1}\,\|\theta\|_{B^r_{p,\infty}}\,\|\theta\|_{B^s_{p,\infty}}. \tag{4.5}$$

To obtain a lower bound for I, we use Proposition 2.6 and a basic embedding inequality,

$$I \ge C\,\kappa\int\left|\Lambda^\alpha\left(|\Delta_j\theta|^{\frac{p}{2}}\right)\right|^2 dx \ge C\,\kappa\left(\int|\Delta_j\theta|^{\frac{p}{1-\alpha}}dx\right)^{1-\alpha} = C\,\kappa\|\Delta_j\theta\|^p_{L^{\frac{p}{1-\alpha}}},$$

where the assumption $p = 2^N$ is used in the first inequality. Applying the interpolation inequality

$$\|f\|_{L^p} \le C\,\|f\|_{L^2}^{\frac{2\alpha}{p+2\alpha-2}}\,\|f\|_{L^{\frac{p}{1-\alpha}}}^{\frac{p-2}{p+2\alpha-2}}$$

with $f = \Delta_j\theta$, we finally obtain the lower bound

$$I \ge C\,\kappa\|\Delta_j\theta\|_{L^p}^{(1+\beta)p}\,\|\Delta_j\theta\|_{L^2}^{-\beta p}, \tag{4.6}$$

where we have set $\beta = \frac{2\alpha}{p-2}$. Combining (4.4), (4.5), and (4.6) yields

$$\frac{d}{dt}\|\theta\|_{B^r_{p,\infty}} + C\,p^{-1}\,\kappa\,2^{jr}\,\|\Delta_j\theta\|_{L^p}^{1+\beta\,p}\,\|\Delta_j\theta\|_{L^2}^{-\beta p} \le C\,\|\theta\|_{B^r_{p,\infty}}\,\|\theta\|_{B^s_{p,\infty}}$$

or, equivalently,

$$\frac{d}{dt}\|\theta\|_{B^r_{p,\infty}} + C\,p^{-1}\,\kappa\,\|\theta\|_{B^r_{p,\infty}}^{1+\beta\,p}\,\|\theta\|_{B^r_{2,\infty}}^{-\beta\,p} \le C\,\|\theta\|_{B^r_{p,\infty}}\,\|\theta\|_{B^s_{p,\infty}}. \qquad \square$$

We now explore several consequences of Proposition 4.2. If $(1-2\alpha)p \le 2$, then $2\alpha + 2/p \ge 1$ or $\beta p \ge 1$. In addition, $r > 1 + 2/p$ implies $r > 2 - 2\alpha$. It thus follows from Corollary 3.3 that $\|\theta_0\|_{B^r_{2,\infty}} \le C_0\kappa$ implies

$$\|\theta(\cdot,t)\|_{B^r_{2,\infty}} \le C_0\kappa$$

for all $t > 0$. Consequently, (4.3) can be reduced to

$$\frac{d}{dt}\|\theta\|_{B^r_{p,\infty}} \le C_5\,\|\theta\|^2_{B^r_{p,\infty}}\big(1 - C_5^{-1}\,C_4\,(p\,C_0^{\beta p})^{-1}\,\kappa^{1-\beta p}\,\|\theta\|^{\beta\,p-1}_{B^r_{p,\infty}}\big). \tag{4.7}$$

For $\beta p > 1$, (4.7) indicates that

$$\begin{cases} \|\theta(\cdot,t)\|_{B^r_{p,\infty}} \text{ decreases as a function of } t \text{ for a big initial norm } \|\theta_0\|_{B^r_{p,\infty}}, \\ \|\theta(\cdot,t)\|_{B^r_{p,\infty}} \text{ increases up to } \big(C_4/(p\,C_5\,C_0^{\beta\,p})\big)^{\frac{1}{\beta p-1}}\,\kappa \text{ for small } \|\theta_0\|_{B^r_{p,\infty}}. \end{cases}$$

In other words,

$$\|\theta(\cdot,t)\|_{B^r_{p,\infty}} \le \max\Big\{\|\theta_0\|_{B^r_{p,\infty}},\,\big(C_4/(p\,C_5\,C_0^{\beta\,p})\big)^{\frac{1}{\beta p-1}}\,\kappa\Big\}.$$

For $\beta p = 1$ and $C_0 \le C_4/(pC_5)$, (4.7) indicates that $\|\theta(\cdot,t)\|_{B^r_{p,\infty}}$ is a decreasing function of t and thus

$$\|\theta(\cdot,t)\|_{B^r_{p,\infty}} \le \|\theta_0\|_{B^r_{p,\infty}}$$

for any $t \ge 0$.

If $(1-2\alpha)p > 2$, then $1 + \beta p < 2$ and $r > 2 - 2\alpha$ implies that $r > 1 + \frac{2}{p}$. In this case, (4.3) becomes

$$\frac{d}{dt}\|\theta\|_{B^r_{p,\infty}} \le \|\theta\|^{1+\beta\,p}_{B^r_{p,\infty}}\big(C_5\|\theta\|^{1-\beta\,p}_{B^r_{p,\infty}} - C_4\,(p\,C_0^{\beta p})^{-1}\,\kappa^{1-\beta p}\big).$$

If θ_0 satisfies

$$\|\theta_0\|_{B^r_{p,\infty}} \le \big(C_4/(p\,C_5\,C_0^{\beta\,p})\big)^{\frac{1}{1-\beta p}}\,\kappa, \tag{4.8}$$

then $\|\theta(\cdot,t)\|_{B^r_{p,\infty}}$ decreases as a function of t and thus

$$\|\theta(\cdot,t)\|_{B^r_{p,\infty}} \le \|\theta_0\|_{B^r_{p,\infty}}$$

for any $t \ge 0$.

In summary, we have established the following corollary.

COROLLARY 4.3. *Let $\kappa > 0$ and let $0 \le \alpha \le 1$. Assume that θ solves the 2D dissipative QG equation* (4.1) *corresponding to θ_0 in $B^r_{2,\infty} \cap B^r_{p,\infty}$ with $p = 2^N (N > 1)$. If r and θ_0 satisfy* (4.2)*, then we have the global bounds*

$$\|\theta(\cdot,t)\|_{B^r_{2,\infty}} \le \tilde{C}_0\kappa \quad \textit{and} \quad \|\theta(\cdot,t)\|_{B^r_{p,\infty}} \le \max\{\|\theta_0\|_{B^r_{p,\infty}}, \tilde{C}_0\kappa\}$$

for some constant $\tilde{C}_0$ depending on α, r, and p only.

It is worth mentioning that the argument leading to the above corollary can be replaced by utilizing explicit formulas given in the following lemma.

LEMMA 4.4. *Let $\sigma > 0$. Assume that $y = y(t)$ satisfies*

$$\frac{d}{dt}y + g(t)y^{1+\sigma} \le h(t)y \tag{4.9}$$

for real-valued functions g and h. Then $y = y(t)$ is bounded pointwise according to

$$y(t) \le \frac{y(0)\,\exp\left(\int_0^t h(\tau)d\tau\right)}{\left(1+\sigma\, y^\sigma(0)\,\int_0^t g(\tau)\exp\left(\sigma\int_0^\tau h(s)ds\right)d\tau\right)^{\frac{1}{\sigma}}}. \tag{4.10}$$

Proof. It follows easily from (4.9) that $z = y\,\exp\left(-\int_0^t h(\tau)d\tau\right)$ satisfies

$$\frac{d}{dt}z \le -g(t)\,\exp\left(\sigma\int_0^t h(\tau)d\tau\right)\,z^{1+\sigma}.$$

Dividing both sides by $z^{1+\sigma}$ and integrating over $[0,t]$, we obtain

$$z^{-\sigma}(t) \ge z^{-\sigma}(0) + \sigma\int_0^t g(\tau)\exp\left(\sigma\int_0^\tau h(s)ds\right)d\tau,$$

which can be converted into the following inequality for y:

$$y^\sigma \le \frac{y^\sigma(0)\,\exp\left(\sigma\int_0^t h(\tau)d\tau\right)}{1+\sigma y^\sigma(0)\,\int_0^t g(\tau)\exp\left(\sigma\int_0^\tau h(s)ds\right)d\tau}.$$

Raising both sides to $\frac{1}{\sigma}$ yields (4.10). ∎

When an extra term $f(t)$ is added to (4.9), the method of variation of constants still allows us to obtain a formal bound involving a function $C(t)$, which satisfies an additional ordinary differential equation.

LEMMA 4.5. *Let $\sigma > 0$. Assume that $y = y(t)$ satisfies*

$$\frac{d}{dt}y + g(t)y^{1+\sigma} \le h(t)y + f(t) \tag{4.11}$$

for real-valued functions g, h, and f. Then y obeys the bound

$$y(t) \le \frac{\exp\left(\int_0^t h(\tau)d\tau\right)}{\left(-\sigma C(t)+\sigma\,\int_0^t g(\tau)\exp\left(\sigma\int_0^\tau h(s)ds\right)d\tau\right)^{\frac{1}{\sigma}}}, \tag{4.12}$$

where $C(t)$ satisfies the following ordinary differential equation:

$$\begin{aligned}\frac{d}{dt}C(t) &= f(t)\,\exp\left(-\int_0^t h(\tau)d\tau\right)\\ &\quad\times\left(-\sigma C(t)+\sigma\,\int_0^t g(\tau)\exp\left(\sigma\int_0^\tau h(s)ds\right)d\tau\right)^{1+\frac{1}{\sigma}}\end{aligned} \tag{4.13}$$

with the initial datum $C(0) = -1/(\sigma y^\sigma(0))$.

Remark. When $f = 0$, $C(t) = C(0) = -1/(\sigma y^\sigma(0))$ and (4.12) becomes (4.10).

4.2. Proof of Theorem 4.1. Assume that $\{\theta^{(n)}\}$ is a successive approximation sequence satisfying the equations

$$\begin{cases} \theta^{(1)} = S_2\theta_0, \\ \partial_t\theta^{(n+1)} + u^{(n)}\cdot\nabla\theta^{(n+1)} + \kappa\Lambda^{2\alpha}\theta^{(n+1)} = 0, \\ u^{(n)} = \mathcal{R}^{\perp}(\theta^{(n)}), \\ \theta^{(n+1)}(x,0) = \theta_0^{(n+1)}(x) = S_{n+2}\theta_0. \end{cases}$$

Following the same procedure as in the proof of Proposition 4.2 leads to

$$\frac{d}{dt}\|\theta^{(n+1)}\|_{B^r_{p,\infty}} + C_4\, p^{-1}\,\kappa\,\|\theta^{(n+1)}\|^{1+\beta\,p}_{B^r_{p,\infty}}\,\|\theta^{(n+1)}\|^{-\beta\,p}_{B^r_{2,\infty}} \le C_5\,\|\theta^{(n)}\|_{B^r_{p,\infty}}\,\|\theta^{(n+1)}\|_{B^r_{p,\infty}}. \tag{4.14}$$

If the conditions in (4.2) are met, we know from the proof of Theorem 3.1 that

$$\|\theta^{(n)}(\cdot,t)\|_{B^r_{2,\infty}} \le C_0\kappa$$

for any integer n and any $t \ge 0$. Inequality (4.14) can then be rewritten as

$$\frac{d}{dt}\|\theta^{(n+1)}\|_{B^r_{p,\infty}} \le \|\theta^{(n+1)}\|_{B^r_{p,\infty}}\left(C_5\,\|\theta^{(n)}\|_{B^r_{p,\infty}} - C_4(pC_0^{\beta p})^{-1}\kappa^{1-\beta p}\,\|\theta^{(n+1)}\|^{\beta\,p}_{B^r_{p,\infty}}\right).$$

When (4.2) is satisfied, we can argue similarly as in the previous subsection and conclude that

$$\|\theta^{(n)}(\cdot,t)\|_{B^r_{p,\infty}} \le \begin{cases} \|\theta_0\|_{B^r_{p,\infty}} & \text{if } \beta p \le 1, \\ \max\left\{\|\theta_0\|_{B^r_{p,\infty}},\ \left(C_4/(p\,C_5\,C_0^{\beta\,p})\right)^{\frac{1}{\beta p-1}}\ \kappa\right\} & \text{if } \beta p > 1. \end{cases} \tag{4.15}$$

An alternative argument using the explicit formula in Lemma 4.9 also leads to the same bound.

We now show that $\{\eta^{(n)}\} = \{\theta^{(n)} - \theta^{(n-1)}\}$ is a Cauchy sequence in $C([0,\infty); B^{r-1}_{p,\infty})$. The sequence $\{\eta^{(n)}\}$ satisfies

$$\begin{cases} \eta^{(1)} = S_2\theta_0 - \theta_0, \\ \partial_t\eta^{(n+1)} + u^{(n)}\cdot\nabla\eta^{(n+1)} + \kappa\Lambda^{2\alpha}\eta^{(n+1)} = w^{(n)}\cdot\nabla\theta^{(n)}, \\ w^{(n)} = \mathcal{R}^{\perp}(\eta^{(n)}), \\ \eta^{(n+1)}(x,0) = \eta_0^{(n+1)}(x) = \Delta_{n+1}\theta_0. \end{cases}$$

Following the procedures as in the proof of Theorem 3.1 as well as in the first part of this proof, we obtain

$$\frac{d}{dt}\|\eta^{(n+1)}\|_{B^{r-1}_{p,\infty}} + C_4\,p^{-1}\,\kappa\,\|\eta^{(n+1)}\|^{1+\beta\,p}_{B^{r-1}_{p,\infty}}\,\|\eta^{(n+1)}\|^{-\beta\,p}_{B^{r-1}_{2,\infty}} \le K_1 + K_2, \tag{4.16}$$

where K_1 and K_2 represent

$$K_1 = 2^{(r-1)j}\|[u^{(n)}\cdot\nabla, \Delta_j]\eta^{(n+1)}\|_{L^p}, \quad K_2 = \|w^{(n)}\cdot\nabla\theta^{(n)}\|_{B^{r-1}_{p,\infty}}.$$

To estimate K_1 and K_2, we assume that (4.2) is satisfied. By Proposition 2.3,

$$\begin{aligned} K_1 &\le C\left(\|\nabla u^{(n)}\|_{L^\infty}\|\eta^{(n+1)}\|_{B^{r-1}_{p,\infty}} + \|\eta^{(n+1)}\|_{L^\infty}\|u^{(n)}\|_{B^r_{p,\infty}}\right) \\ &\le 2C\,\|u^{(n)}\|_{B^r_{p,\infty}}\,\|\eta^{(n+1)}\|_{B^{r-1}_{2,\infty}} \\ &\le 2C\,\|\theta^{(n)}\|_{B^r_{p,\infty}}\,\|\eta^{(n+1)}\|_{B^{r-1}_{p,\infty}}. \end{aligned} \tag{4.17}$$

By Proposition 2.4,

$$\begin{aligned} K_2 &\le C\left(\|w^{(n)}\|_{L^\infty}\|\nabla\theta^{(n)}\|_{B^{r-1}_{p,\infty}} + \|w^{(n)}\|_{B^{r-1}_{p,\infty}}\|\nabla\theta^{(n)}\|_{L^\infty}\right) \\ &\le 2C\|\theta^{(n)}\|_{B^r_{p,\infty}}\|w^{(n)}\|_{B^{r-1}_{p,\infty}} \\ &\le 2C\|\theta^{(n)}\|_{B^r_{p,\infty}}\|\eta^{(n)}\|_{B^{r-1}_{p,\infty}}. \end{aligned} \tag{4.18}$$

Inserting (4.17) and (4.18) in (4.16), we obtain

$$\begin{aligned} &\frac{d}{dt}\|\eta^{(n+1)}\|_{B^{r-1}_{p,\infty}} + C_4\,p^{-1}\,\kappa\,\|\eta^{(n+1)}\|^{1+\beta\,p}_{B^{r-1}_{p,\infty}}\,\|\eta^{(n+1)}\|^{-\beta\,p}_{B^{r-1}_{2,\infty}} \\ &\quad\le C_7\,\|\theta^{(n)}\|_{B^r_{p,\infty}}\,\left(\|\eta^{(n+1)}\|_{B^{r-1}_{p,\infty}} + \|\eta^{(n)}\|_{B^{r-1}_{p,\infty}}\right). \end{aligned} \tag{4.19}$$

According to the proof of Theorem 3.1 and the first part of this proof,

$$\|\eta^{(n+1)}\|_{B^{r-1}_{2,\infty}} \le \|\theta_0\|_{B^{r-1}_{2,\infty}}\,2^{-(n-2)}, \qquad \|\theta^{(n)}\|_{B^r_{p,\infty}} \le \max\{\|\theta_0\|_{B^r_{p,\infty}},\,\tilde{C}\kappa\},$$

where $\tilde{C}$ is a constant. We are now ready to show that

$$\|\eta^{(n+1)}(\cdot,t)\|_{B^{r-1}_{p,\infty}} \le \bar{C}\,2^{-(n-2-1/(\beta p))},$$

where $\bar{C}$ is given explicitly by

$$\bar{C} = \max\left\{\frac{1}{2},\ \left(2C_7\,\max\{\|\theta_0\|_{B^r_{p,\infty}},\tilde{C}\kappa\}/(C_4 p^{-1}\kappa)\right)^{1/\sigma}\right\}\,\|\theta_0\|_{B^r_{p,\infty}}.$$

To simplify the notation, we set

$$\sigma = \beta p, \quad y(t) = \|\eta^{(n+1)}(\cdot,t)\|_{B^{r-1}_{p,\infty}}, \quad g = C_4 p^{-1}\,\kappa\,2^{(n-1)}\,\|\theta_0\|^{-\sigma}_{B^{r-1}_{2,\infty}},$$

$$h = C_7\,\max\{\|\theta_0\|_{B^r_{p,\infty}},\,\tilde{C}\kappa\}, \quad f = C_7\,\max\{\|\theta_0\|_{B^r_{p,\infty}},\,\tilde{C}\kappa\}\,\bar{C}\,2^{-(n-3-1/(\beta p))}.$$

Inequality (4.19) then becomes

$$\frac{d}{dt}y \le -g\,y^{\sigma+1} + h\,y + f.$$

We further write $z(t)$ for the right-hand side of the inequality above. If $y(0)$ is sufficiently large such that $z(0) \le 0$, then

$$N \equiv \sup_{t\ge 0} y(t) \le y(0).$$

If, on the other hand, $y(0)$ is small and $z(0) \geq 0$, then $y(t)$ initially grows as t increases. But its growth stops as soon as $z(t)$ becomes zero. Therefore, N obeys

$$-g\,N^{1+\sigma} + h\,N + f = 0 \quad \text{or} \quad N^{1+\sigma} - \frac{h}{g}\,N = \frac{f}{g}. \tag{4.20}$$

The discussion is then divided into two cases: i) $N^\sigma \leq 2h/g$ and ii) $N^\sigma > 2h/g$. In the first case,

$$N \leq \left(\frac{2h}{g}\right)^{\frac{1}{\sigma}}.$$

In the second case, (4.20) implies that

$$N^{1+\sigma} < \frac{2f}{g} \quad \text{or} \quad N \leq \left(\frac{2f}{g}\right)^{\frac{1}{1+\sigma}}.$$

In summary, we have obtained

$$\sup_{t\geq 0} y(t) \leq \max\left\{y(0),\, \left(\frac{2h}{g}\right)^{\frac{1}{\sigma}},\, \left(\frac{2f}{g}\right)^{\frac{1}{1+\sigma}}\right\}. \tag{4.21}$$

Returning to the original variable, we find

$$y(0) = \|\Delta_{n+1}\theta_0\|_{B^{r-1}_{p,\infty}} \leq \|\theta_0\|_{B^r_{p,\infty}}\, 2^{-n}, \quad \left(\frac{2h}{g}\right)^{\frac{1}{\sigma}} \leq \bar{C}\, 2^{-n+2},$$

$$\left(\frac{2f}{g}\right)^{\frac{1}{1+\sigma}} \leq \bar{C}\left(2^{-n+3+1/\sigma}\; 2^{-\sigma(n-2)}\right)^{\frac{1}{1+\sigma}} \leq \bar{C}\, 2^{-(n-2-1/\sigma)}.$$

As a consequence, (4.21) yields the desired bound

$$\sup_{t\geq 0} \|\eta^{(n+1)}(\cdot,t)\|_{B^{r-1}_{p,\infty}} \leq \bar{C}\, 2^{-(n-2-1/(\beta p))}.$$

After a similar argument as in the proof of Theorem 3.1, the proof of Theorem 4.1 is then completed. □

REFERENCES

[1] L. Berselli, *Vanishing viscosity limit and long-time behavior for 2D quasi-geostrophic equations*, Indiana Univ. Math. J., 51 (2002), pp. 905–930.

[2] D. Chae, *The quasi-geostrophic equation in the Triebel-Lizorkin spaces*, Nonlinearity, 16 (2003), pp. 479–495.

[3] D. Chae and J. Lee, *Global well-posedness in the super-critical dissipative quasi-geostrophic equations*, Comm. Math. Phys., 233 (2003), pp. 297–311.

[4] J.-Y. Chemin, *Perfect Incompressible Fluids*, Clarendon Press, Oxford, UK, 1998.

[5] P. Constantin, A. Majda, and E. Tabak, *Formation of strong fronts in the 2-D quasi-geostrophic thermal active scalar*, Nonlinearity, 7 (1994), pp. 1495–1533.

[6] P. Constantin, D. Córdoba, and J. Wu, *On the critical dissipative quasi-geostrophic equation*, Indiana Univ. Math. J., 50 (2001), pp. 97–107.

[7] P. Constantin and J. Wu, *Behavior of solutions of 2D quasi-geostrophic equations*, SIAM J. Math. Anal., 30 (1999), pp. 937–948.

[8] D. Córdoba, *Nonexistence of simple hyperbolic blow-up for the quasi-geostrophic equation*, Ann. of Math. (2), 148 (1998), pp. 1135–1152.

[9] A. Córdoba and D. Córdoba, *A maximum principle applied to quasi-geostrophic equations*, Comm. Math. Phys., 249 (2004), pp. 511–528.
[10] D. Córdoba and C. Fefferman, *Growth of solutions for QG and 2D Euler equations*, J. Amer. Math. Soc., 15 (2002), pp. 665–670.
[11] I. Held, R. Pierrehumbert, S. Garner, and K. Swanson, *Surface quasi-geostrophic dynamics*, J. Fluid Mech., 282 (1995), pp. 1–20.
[12] J. Pedlosky, *Geophysical Fluid Dynamics*, Springer-Verlag, New York, 1987.
[13] S. Resnick, *Dynamical Problem in Nonlinear Advective Partial Differential Equations*, Ph.D. thesis, University of Chicago, Chicago, 1995.
[14] M. E. Schonbek and T. P. Schonbek, *Asymptotic behavior to dissipative quasi-geostrophic flows*, SIAM J. Math. Anal., 35 (2003), pp. 357–375.
[15] J. Wu, *Inviscid limits and regularity estimates for the solutions of the 2D dissipative quasi-geostrophic equations*, Indiana Univ. Math. J., 46 (1997), pp. 1113–1124.
[16] J. Wu, *Dissipative quasi-geostrophic equations with L^p data*, Electron. J. Differential Equations, 2001 (2001), pp. 1–13.
[17] J. Wu, *Solutions of the 2D Quasi-geostrophic Equation in Hölder Spaces*, preprint, Oklahoma State University, Stillwater, OK, 2003.

SIAM J. MATH. ANAL.
Vol. 36, No. 4, pp. 1031–1056

LINEARIZED STABILITY ANALYSIS OF STATIONARY SOLUTIONS FOR SURFACE DIFFUSION WITH BOUNDARY CONDITIONS*

HARALD GARCKE†, KAZUO ITO‡, AND YOSHIHITO KOHSAKA§

Abstract. The linearized stability of stationary solutions to the surface diffusion flow with angle conditions and no-flux conditions as boundary conditions is studied. We perform a linearized stability analysis in which the H^{-1}-gradient flow structure plays a key role. As a byproduct our analysis also gives a criterion for the stability of critical points of the length functional of curves which come into contact with the outer boundary. Finally, we study the linearized stability of several examples.

Key words. surface diffusion, gradient flow, stability of stationary solutions, eigenvalues, isoperimetric problems

AMS subject classifications. 35G30, 35B35, 35R35, 80A

DOI. 10.1137/S0036141003437939

1. Introduction.

The geometrical evolution law

$$V = -\Delta\kappa$$

was derived by Mullins [20] to model the motion of interfaces in the case that the motion of interfaces is governed purely by mass diffusion within the interfaces (for simplicity we set the diffusion constant to 1). Here V is the normal velocity of the evolving interface, Δ is the Laplace–Beltrami operator, and κ is the mean curvature of the interface where we use the sign convention that a sphere with the normal pointing to the inside has positive curvature. We also refer to work by Davi and Gurtin [7], who derived the above law from balance laws in conjunction with an appropriate version of the second law of thermodynamics, and to work by Cahn, Elliott, and Novick-Cohen [4], who derived this evolution law as the sharp interface limit of a Cahn–Hilliard equation with degenerate mobility. This evolution law has the property that for closed embedded hypersurfaces the enclosed volume is preserved and the surface area decreases in time (see, e.g., [10], [11]). An existence result for curves in the plane and stability of spheres—which are stationary under the flow—has been shown by Elliott and Garcke [10]. This result was generalized to the higher-dimensional case by Escher, Mayer, and Simonett [11].

In general, interfaces will meet an outer boundary or they might intersect at triple or multiple junctions. In this case boundary conditions have to hold, and they were derived by Garcke and Novick-Cohen [14] as the asymptotic limit of a Cahn–Hilliard system with a degenerate mobility matrix. At the outer boundary and at triple junctions, angle conditions and a balance condition for the mass fluxes have to

*Received by the editors November 26, 2003; accepted for publication (in revised form) April 9, 2004; published electronically January 27, 2005. This work was supported by the 2001 Canon Foundation Research Fellowship, the Research Fellowship of the Japan Society for the Promotion of Young Scientists, and the Regensburger Universitätsstiftung Hans Vielberth.

http://www.siam.org/journals/sima/36-4/43793.html

†NWF I–Mathematik, Universität Regensburg, 93040 Regensburg, Germany (harald.garcke@mathematik.uni-regensburg.de).

‡Graduate School of Mathematical Sciences Kyushu University, Fukuoka 812-8581, Japan (k-ito@math.kyushu-u.ac.jp).

§Muroran Institute of Technology, 27-1 Mizumoto-cho, Muroran 050-8585, Japan (kohsaka@mmm.muroran-it.ac.jp).

hold. In addition, at triple junctions a continuity condition for chemical potentials has to hold. Numerical simulations for the degenerate Cahn–Hilliard systems have been performed by Barrett, Blowey, and Garcke [2]. An existence result for surface diffusion of curves that intersect the outer boundary and meet at triple junctions has been given by Garcke and Novick-Cohen [14]. The stability problem for stationary solutions for the surface diffusion flow with triple junctions was addressed by Ito and Kohsaka [18], by Escher, Garcke, and Ito [12] in the case of a geometry with a mirror symmetry, and by Ito and Kohsaka [19] in a triangular domain. The general case is still open. This is partly due to the fact that the stability depends in a nontrivial way on the geometry of the boundary.

For motion by mean curvature, which is given by the law

$$V = \kappa, \tag{1.1}$$

the stability of stationary interfaces with boundaries was studied by Rubinstein, Sternberg, and Keller [21] in the case where the evolving curves intersect an outer boundary with a 90° angle. For stability results for the stationary solutions of (1.1) in the presence of triple junctions, we refer to Sternberg and Ziemer [22] and Ikota and Yanagida [16]. The last authors developed a linear stability criterion that is based on ideas of Ei and Yanagida [9] and Ei, Sato, and Yanagida [8].

One main difference between motion by mean curvature and motion by surface diffusion is that the former does not preserve volume, whereas the latter does. This implies that the stationary solutions are different. For motion by surface diffusion, spherical arcs that intersect the outer boundary perpendicular are stationary. It is the goal of this paper to study the stability of such stationary solutions under surface diffusion. More precisely we study the following problem. Given an open bounded domain Ω, we look for evolving curves $\Gamma = (\Gamma_t)_{t>0}$ (for a definition see Gurtin [15]) lying in Ω with the following properties (for a precise definition of the flow see section 2):

$$\begin{cases} V = -\kappa_{ss} & \text{for all points on the curve,} \\ \partial\Gamma_t \subset \partial\Omega & \text{at all times,} \\ \sphericalangle(\partial\Omega, \Gamma) = \pi/2 & \text{at the boundary,} \\ \kappa_s = 0 & \text{at the boundary.} \end{cases} \tag{1.2}$$

Here a subscript s denotes differentiation with respect to arc-length. The second and third conditions imply that the boundary of the curves at all times intersects the outer boundary perpendicularly. The last condition says that there is no mass flux at the outer boundary (see [14]). It is not difficult to show that (see [14])

$$\frac{d}{dt}\mathrm{Area}_\Gamma(t) = 0, \qquad \frac{d}{dt}\mathrm{Length}_\Gamma(t) \le 0$$

under surface diffusion with the above boundary conditions. Here we denote by $\mathrm{Area}_\Gamma(t)$ the area enclosed by the curve and $\partial\Omega$ at time t (for definiteness we take the side of Γ toward which the normal points) and by $\mathrm{Length}_\Gamma(t)$ the length of Γ at time t.

We will introduce a linear stability criterion based on the work of [9], [8], [16], which studied the mean curvature flow. The analysis in the case of surface diffusion is more difficult because the surface diffusion flow is the gradient flow with respect to the H^{-1} inner product (see [5], [13], [23]) in contrast to the case of motion by mean curvature, which is a gradient flow with respect to the L^2 inner product. We

want to emphasize that the observation that also the linearized problem is an H^{-1}-gradient flow of the bilinearized area functional is an important ingredient of our analysis (see section 4). Indeed, the zero solution is an asymptotically stable solution of the linearized equation $\rho_t = \mathcal{A}\rho$ ($\mathcal{A}$ being the linearized operator) if and only if all eigenvalues of $\mathcal{A}$ are negative, and it will turn out that this is equivalent to the fact that the bilinearized area functional is positive definite. We refer to Bates and Fife [3], who studied a linearized stability analysis for the Cahn–Hilliard equation, and also to Alikakos et al. [1], who considered the Mullins–Sekerka motion—which is also volume conserving—of small droplets on a fixed boundary.

The stability of stationary arcs that are attached perpendicular to the outer boundary depends on their curvature, their length, and the curvature of the outer boundary in a nontrivial way. The reader is advised to have a look at section 7, where we illustrate the stability behavior with the help of several examples. Taking advantage of the gradient flow property of the evolution and using variational arguments, we are able to analyze the linear stability behavior, i.e., the stability of the zero solution of the linearized operator (see section 6).

It would remain to show that the principle of linearized stability holds, which means, except for the critical (or neutral) case of stability, the zero solution of the linearized problem has the same stability as the stationary solution of the nonlinear problem around which we linearized the equation. For the nonlinear boundary conditions appearing in our problem, abstract semigroup theory cannot be applied directly. However, we refer to [12] for a result in this direction. Note that the principle of linearized stability in [12] is shown in spaces consisting of functions which are $(4+\gamma)$-Hölder continuous with respect to the space variable for $\gamma \in (0,1)$. We remark that it is a nontrivial task to apply the method of [12] to our problem. Indeed, in the setting in [12] the boundary conditions are simpler and the difficulties arising from the nonlinear boundary conditions could be resolved. But in our case the boundary conditions keep a highly nonlinear form, which makes the analysis of the principle of linearized stability more difficult. Furthermore, the area-preserving property leads to additional difficulties. In the linearization, a zero eigenvalue appears if we do not take the area-preserving property into account. The zero eigenvalue corresponds to equilibria of (1.2), which in general enclose a different area. In conclusion, if we do not take the mass conservation into account, none of the equilibria is isolated, which further complicates the nonlinear stability analysis (see [11], [12] for similar difficulties arising from an area-preserving property). Nonlinear stability analysis is the topic of ongoing research.

Finally, we remark that our results also have some relevance for isoperimetric problems, as they give stability results for critical points of the length functional, which is restricted to curves that enclose a fixed area. Since the surface diffusion flow reduces length conserving area at the same time, the study of critical points of the length functional (given an area constraint) is what the stability analysis for the evolution problem can be reduced to.

2. Parameterization. In this section we give a precise definition of the flow (1.2), and in particular we introduce a parameterization of an evolving curve that will be convenient for our analysis. For a smooth function $\psi : \mathbb{R}^2 \to \mathbb{R}$ with $\nabla\psi(x) \neq 0$, if $\psi(x) = 0$, set

$$\Omega = \{x \in \mathbb{R}^2 \mid \psi(x) < 0\}, \quad \partial\Omega = \{x \in \mathbb{R}^2 \mid \psi(x) = 0\}.$$

Let Γ_* be a stationary solution and let σ be the arc-length parameter of Γ_*. Then we denote an arc-length parameterization of Γ_* as

$$\Gamma_* = \{\Phi_*(\sigma) \mid \sigma \in [-l, l]\}.$$

Note that we can extend Γ_* naturally either to the full circle when Γ_* is a part of the circle or to the straight line when Γ_* is a line segment. Also note that the curvature κ_* of Γ_* is a constant. We denote

$$\bar{l} := \begin{cases} \pi/|\kappa_*|, & \kappa_* \neq 0, \\ +\infty, & \kappa_* = 0; \end{cases}$$

i.e., $\bar{l}$ is the length of the extension of Γ_* to a full circle (if $\kappa_* \neq 0$). Define

$$\begin{cases} \xi_+(q) = \max\{\sigma \in (-\bar{l}, \bar{l}) \mid \Phi_*(\sigma) + qN_*(\sigma) \in \Omega\}, \\ \xi_-(q) = \min\{\sigma \in (-\bar{l}, \bar{l}) \mid \Phi_*(\sigma) + qN_*(\sigma) \in \Omega\}, \end{cases}$$

where $q \in [-d, d]$ for a small $d > 0$, and $N_*(\sigma)$ is a unit normal vector of Γ_* at σ and is obtained by rotating the unit tangent vector $T_*(\sigma)$ of Γ_* with $\pi/2$. Then it holds that $\psi(\Phi_*(\xi_\pm(q)) + qN_*(\xi_\pm(q))) = 0$. In addition, we have $\xi_\pm(0) = \pm l$. Using the implicit function theorem, we see that $\xi_+(q)$ and $\xi_-(q)$ are smooth. Let

$$\Psi(\sigma, q) := \Phi_*(\xi(\sigma, q)) + qN_*(\xi(\sigma, q))$$

with

$$\xi(\sigma, q) := \xi_-(q) + \frac{\sigma + l}{2l}(\xi_+(q) - \xi_-(q)).$$

Note that $\xi(\pm l, q) = \xi_\pm(q)$ and $\xi(\sigma, 0) = \sigma$.

Let Γ be curves in the neighborhood of Γ_*, which touch the boundary $\partial\Omega$ and are contained in Ω. For some functions $\rho : [-l, l] \to [-d, d]$, we define $\Phi(\sigma) := \Psi(\sigma, \rho(\sigma))$ for $\sigma \in [-l, l]$, which denotes a parameterization of such curves Γ. Thus we set

$$\Gamma(t) := \{\Phi(\sigma, t) \mid \sigma \in [-l, l]\} \tag{2.1}$$

with $\Phi(\sigma, t) := \Psi(\sigma, \rho(\sigma, t))$ for a function ρ depending on σ and t. We remark that $\rho \equiv 0$ means that curves Γ coincide with a stationary curve Γ_*.

Let us derive the representation of (1.2) to the parameterization (2.1). For the arc-length parameter s of Γ, we have

$$\frac{ds}{d\sigma} = |\Phi_\sigma| = \sqrt{|\Psi_\sigma|^2 + 2(\Psi_\sigma, \Psi_q)_{\mathbb{R}^2}\rho_\sigma + |\Psi_q|^2\rho_\sigma^2} \ (=: J(\rho)). \tag{2.2}$$

Here and hereafter $(\cdot, \cdot)_{\mathbb{R}^2}$ denotes the inner product in $\mathbb{R}^2$. Then we find

$$T = \frac{1}{J(\rho)}\Phi_\sigma, \quad N = \frac{1}{J(\rho)}R\Phi_\sigma,$$

where T and N are the unit tangent and normal vector of Γ, respectively, and R is the rotation matrix with $\pi/2$. The normal velocity V of $\Gamma(t)$ is denoted by

$$V = (\Phi_t, N)_{\mathbb{R}^2} = \frac{1}{J(\rho)}(\Phi_t, R\Phi_\sigma)_{\mathbb{R}^2} = \frac{1}{J(\rho)}(\Psi_q, R\Psi_\sigma)_{\mathbb{R}^2}\rho_t.$$

Moreover, since (2.2) gives

$$\partial_s^2 = \frac{1}{J(\rho)}\partial_\sigma\left(\frac{1}{J(\rho)}\partial_\sigma\right) = \frac{1}{(J(\rho))^2}\partial_\sigma^2 + \frac{1}{J(\rho)}\left(\partial_\sigma\frac{1}{J(\rho)}\right)\partial_\sigma \ (=: \Delta(\rho)), \tag{2.3}$$

the curvature κ of $\Gamma(t)$ is written by

$$\begin{aligned}\kappa(\rho) &= (\Delta(\rho)\Phi, N)_{\mathbb{R}^2}\\ &= \frac{1}{(J(\rho))^3}(\Phi_{\sigma\sigma}, R\Phi_\sigma)_{\mathbb{R}^2}\\ &= \frac{1}{(J(\rho))^3}\Big[(\Psi_q, R\Psi_\sigma)_{\mathbb{R}^2}\rho_{\sigma\sigma} + \big\{2(\Psi_{\sigma q}, R\Psi_\sigma)_{\mathbb{R}^2} + (\Psi_{\sigma\sigma}, R\Psi_q)_{\mathbb{R}^2}\big\}\rho_\sigma\\ &\quad + \big\{(\Psi_{qq}, R\Psi_\sigma)_{\mathbb{R}^2} + 2(\Psi_{\sigma q}, R\Psi_q)_{\mathbb{R}^2} + (\Psi_{qq}, R\Psi_q)_{\mathbb{R}^2}\rho_\sigma\big\}\rho_\sigma^2\\ &\quad + (\Psi_{\sigma\sigma}, R\Psi_\sigma)_{\mathbb{R}^2}\Big].\end{aligned} \tag{2.4}$$

Thus the surface diffusion flow equation is described by

$$\rho_t = -L(\rho)\Delta(\rho)\kappa(\rho), \tag{2.5}$$

where

$$L(\rho) := \frac{1}{(\Psi_q, R\Psi_\sigma)_{\mathbb{R}^2}}J(\rho). \tag{2.6}$$

Let us derive the representation of the boundary conditions, which are the Neumann boundary condition and the no-flux condition $\kappa_s = 0$ on $\partial\Omega$ (the second condition in (1.2) is automatically fulfilled). Since the Neumann boundary condition $(\Phi_\sigma, T_{\partial\Omega})_{\mathbb{R}^2} = 0$ on $\partial\Omega$ is equivalent to $(R\Phi_\sigma, \nabla\psi(\Phi))_{\mathbb{R}^2} = 0$ on $\partial\Omega$, we have

$$(R\Psi_\sigma + R\Psi_q\rho_\sigma, \nabla\psi(\Psi))_{\mathbb{R}^2} = 0 \quad \text{at} \ \ \sigma = \pm l.$$

By (2.2) and (2.4) the no-flux condition $\kappa_s = 0$ on $\partial\Omega$ is denoted by

$$\partial_\sigma\kappa(\rho) = 0 \quad \text{at} \ \ \sigma = \pm l.$$

Consequently we have the following proposition.

PROPOSITION 2.1. *For a parameterization* (2.1)*, problem* (1.2) *is represented by*

$$\begin{cases} \rho_t = -L(\rho)\Delta(\rho)\kappa(\rho) & \text{for} \quad \sigma \in (-l, l), \ t > 0,\\ (R\Psi_\sigma + R\Psi_q\rho_\sigma, \nabla\psi(\Psi))_{\mathbb{R}^2} = 0 & \text{at} \quad \sigma = \pm l,\\ \partial_\sigma\kappa(\rho) = 0 & \text{at} \quad \sigma = \pm l, \end{cases} \tag{2.7}$$

where $L(\rho)$*,* $\Delta(\rho)$*, and* $\kappa(\rho)$ *are defined by* (2.6)*,* (2.3)*, and* (2.4)*, respectively.*

3. Linearization. To study the linearized stability of a stationary solution Γ_*, the curvature κ_* of which is a constant, we linearize (2.7) around $\rho \equiv 0$. For this purpose we need the following properties of Ψ at $q = 0$.

LEMMA 3.1. *For the parameterization of section* 2*, the following hold:*

(i) $\Psi(\sigma, 0) = \Phi_*(\sigma)$.

(ii) $\Psi_\sigma(\sigma, 0) = T_*(\sigma)$ *and* $\Psi_q(\sigma, 0) = N_*(\sigma)$.

(iii) $\Psi_{\sigma\sigma}(\sigma, 0) = \kappa_* N_*(\sigma)$ *and* $\Psi_{\sigma q}(\sigma, 0) = -\kappa_* T_*(\sigma)$.

(iv) $\Psi_{\sigma\sigma q}(\sigma,0) = -\kappa_*^2 N_*(\sigma)$.

Proof. By the definition of Ψ, (i) is obvious. Using (i), we readily derive $\Psi_\sigma(\sigma,0) = T_*(\sigma)$. To derive $\Psi_q(\sigma,0) = N_*(\sigma)$, we first prove $\xi_+'(0) = \xi_-'(0) = 0$. Note that $\xi(\pm l, q) = \xi_\pm(q)$ and $\xi_q(\pm l, q) = \xi_\pm'(q)$. Since it follows from the Frenet–Serret formula that

$$\Psi_q(\sigma,q) = \xi_q(\sigma,q)(1-q\kappa_*)T_*(\xi(\sigma,q)) + N_*(\xi(\sigma,q)), \tag{3.1}$$

we are led to

$$\begin{aligned} 0 &= \frac{d}{dq}\psi(\Psi(\pm l,q)) \\ &= (1-q\kappa_*)(\nabla\psi(\Psi(\pm l,q)), T_*(\xi_\pm(q)))_{\mathbb{R}^2}\xi_\pm'(q) + (\nabla\psi(\Psi(\pm l,q)), N_*(\xi_\pm(q)))_{\mathbb{R}^2}. \end{aligned}$$

Putting $q=0$, we have $(\nabla\psi(\Phi_*(\pm l)), T_*(\pm l))_{\mathbb{R}^2}\xi_\pm'(0) = 0$, so that $\xi_+'(0) = \xi_-'(0) = 0$. Then this implies

$$\xi_q(\sigma,0) = \xi_-'(0) + \frac{\sigma+l}{2l}(\xi_+'(0) - \xi_-'(0)) = 0.$$

Putting $q=0$ in (3.1), we derive $\Psi_q(\sigma,0) = N_*(\sigma)$. By virtue of (ii) and the Frenet–Serret formula, we readily derive (iii). Finally, by differentiating $\Psi_{\sigma q}(\sigma,0) = -\kappa_* T_*(\sigma)$ with respect to σ and applying the Frenet–Serret formula, we are led to (iv). ∎

We define the operator $G(\rho) := -L(\rho)\Delta(\rho)\kappa(\rho)$, which maps a function $\rho \in C^4[-l,l]$ to a function in $C^0[-l,l]$. Then we can compute this Fréchet derivative as follows.

LEMMA 3.2. *The operator $G : C^4[-l,l] \to C^0[-l,l]$ is Fréchet differentiable with derivative $\mathcal{A}_0 := \partial G(0)$, where*

$$\mathcal{A}_0\rho = -\partial_\sigma^2(\partial_\sigma^2 + \kappa_*^2)\rho.$$

Proof. Since $G(\rho) = -L(\rho)\Delta(\rho)\kappa(\rho)$, we have

$$\mathcal{A}_0\rho = \partial G(0)\rho = -(\partial L(0)\rho)\Delta(0)\kappa(0) - L(0)(\partial\Delta(0)\rho)\kappa(0) - L(0)\Delta(0)\partial\kappa(0)\rho. \tag{3.2}$$

By virtue of Lemma 3.1 and the definition of $L(\rho)$, $\Delta(\rho)$, and $\kappa(\rho)$, we observe

$$L(0) \equiv 1, \quad \Delta(0) = \partial_\sigma^2, \quad \kappa(0) \equiv \kappa_*. \tag{3.3}$$

Then, since κ_* is a constant, we have $\Delta(0)\kappa(0) = 0$ and

$$\begin{aligned} &(\partial\Delta(0)\rho)\kappa(0) \\ &\quad = \left(\frac{d}{d\varepsilon}(J(\varepsilon\rho))^{-2}\Big|_{\varepsilon=0}\right)\partial_\sigma^2\kappa(0) + \left(\frac{d}{d\varepsilon}(J(\varepsilon\rho))^{-1}\partial_\sigma(J(\varepsilon\rho))^{-1}\Big|_{\varepsilon=0}\right)\partial_\sigma\kappa(0) = 0. \end{aligned}$$

Let us derive $\partial\kappa(0)\rho$. Set

$$\begin{cases} a_1(\rho) = (\Psi_q, R\Psi_\sigma)_{\mathbb{R}^2}, \\ a_2(\rho) = 2(\Psi_{\sigma q}, R\Psi_\sigma)_{\mathbb{R}^2} + (\Psi_{\sigma\sigma}, R\Psi_q)_{\mathbb{R}^2}, \\ a_3(\rho) = (\Psi_{qq}, R\Psi_\sigma)_{\mathbb{R}^2} + 2(\Psi_{\sigma q}, R\Psi_q)_{\mathbb{R}^2} + (\Psi_{qq}, R\Psi_q)_{\mathbb{R}^2}\rho_\sigma, \\ a_4(\rho) = (\Psi_{\sigma\sigma}, R\Psi_\sigma)_{\mathbb{R}^2}. \end{cases}$$

Then $\kappa(\rho)$ is written by

$$\kappa(\rho) = (J(\rho))^{-3} a(\rho),$$

where

$$a(\rho) := a_1(\rho)\rho_{\sigma\sigma} + a_2(\rho)\rho_\sigma + a_3(\rho)\rho_\sigma^2 + a_4(\rho).$$

Thus we have

$$\begin{aligned}\partial\kappa(0)\rho &= \frac{d}{d\varepsilon}\kappa(\varepsilon\rho)\bigg|_{\varepsilon=0} \\ &= (J(0))^{-3}\frac{d}{d\varepsilon}a(\varepsilon\rho)\bigg|_{\varepsilon=0} + \left(\frac{d}{d\varepsilon}(J(\varepsilon\rho))^{-3}\bigg|_{\varepsilon=0}\right) a(0).\end{aligned}$$

By virtue of Lemma 3.1, we observe $J(0) = 1$ and $a(0) = a_4(0) = \kappa_*$. In addition, it holds that

$$\begin{aligned}&\frac{d}{d\varepsilon}a(\varepsilon\rho)\bigg|_{\varepsilon=0} = a_1(0)\rho_{\sigma\sigma} + a_2(0)\rho_\sigma + \partial a_4(0)\rho = \partial_\sigma^2\rho - 2\kappa_*^2\rho, \\ &\frac{d}{d\varepsilon}(J(\varepsilon\rho))^{-3}\bigg|_{\varepsilon=0} = -3(J(0))^{-4}\frac{d}{d\varepsilon}J(\varepsilon\rho)\bigg|_{\varepsilon=0} = 3\kappa_*\rho.\end{aligned}$$

Consequently, we are led to

$$\partial\kappa(0)\rho = (\partial_\sigma^2 + \kappa_*^2)\rho. \tag{3.4}$$

The assertion follows from (3.2)–(3.4). □

Let us consider the boundary condition. Set

$$\begin{cases} B_1(\rho) := (R\Psi_\sigma, \nabla\psi(\Psi))_{\mathbb{R}^2} + (R\Psi_q, \nabla\psi(\Psi))_{\mathbb{R}^2}\rho_\sigma, \\ B_2(\rho) := \partial_\sigma\kappa(\rho) \end{cases}$$

for $\rho \in C^3[-l, l]$, i.e., the operator B_i $(i = 1, 2)$ maps $C^3[-l, l]$ to $C^0[-l, l]$. Define

$$\mathcal{B}_0 := \begin{pmatrix} \partial B_1(0)/(\mp|\nabla\psi(x_*^\pm)|) \\ \partial B_2(0) \end{pmatrix} \quad \text{at} \quad \sigma = \pm l,$$

where $x_*^\pm := \Phi_*(\pm l) \in \partial\Omega$ and $\partial B_i(0)$ $(i = 1, 2)$ is the Fréchet derivatives of B_i at 0. Then we have the following representation of $\mathcal{B}_0$.

LEMMA 3.3. *Let ρ belong to $C^3[-l, l]$ and let $h_\pm$ be the curvatures of $\partial\Omega$ at $x_*^\pm \in \Gamma_* \cap \partial\Omega$, respectively (where we use the sign convention that $h_\pm < 0$ if Ω is convex). Then*

$$\mathcal{B}_0\rho = \begin{pmatrix} \partial_\sigma \pm h_\pm \\ \partial_\sigma(\partial_\sigma^2 + \kappa_*^2) \end{pmatrix}\rho \quad \text{at} \quad \sigma = \pm l.$$

Proof. First we derive $\partial B_1(0)$. Set

$$b_1(\rho) = (R\Psi_\sigma, \nabla\psi(\Psi))_{\mathbb{R}^2}, \quad b_2(\rho) = (R\Psi_q, \nabla\psi(\Psi))_{\mathbb{R}^2}.$$

Then we have $B_1(\rho) = b_1(\rho) + b_2(\rho)\rho_\sigma$, so that

$$\partial B_1(0)\rho = \frac{d}{d\varepsilon}B_1(\varepsilon\rho)\bigg|_{\varepsilon=0} = \frac{d}{d\varepsilon}b_1(\varepsilon\rho)\bigg|_{\varepsilon=0} + b_2(0)\rho_\sigma.$$

It follows from Lemma 3.1 that

$$\frac{d}{d\varepsilon} R\Psi_\sigma(\sigma, \varepsilon\rho)\Big|_{\varepsilon=0} = -\kappa_* N_*(\sigma)\rho,$$
$$\frac{d}{d\varepsilon} \nabla\psi(\Psi(\sigma, \varepsilon\rho))\Big|_{\varepsilon=0} = [D^2\psi(\Phi_*(\sigma))]N_*(\sigma)\rho,$$

where $D^2\psi$ is the Hessian matrix of ψ. Since $(N_*(\sigma), \nabla\psi(\Phi_*(\sigma)))_{\mathbb{R}^2} = 0$ at $\sigma = \pm l$, we are led to

$$\frac{d}{d\varepsilon} b_1(\varepsilon\rho)\Big|_{\varepsilon=0} = (N_*(\sigma), [D^2\psi(\Phi_*(\sigma))]N_*(\sigma))_{\mathbb{R}^2}\rho \quad \text{at} \quad \sigma = \pm l.$$

This implies that for $\sigma = \pm l$

$$\partial B_1(0)\rho = -(T_*(\sigma), \nabla\psi(\Phi_*(\sigma)))_{\mathbb{R}^2}\rho_\sigma + (N_*(\sigma), [D^2\psi(\Phi_*(\sigma))]N_*(\sigma))_{\mathbb{R}^2}\rho.$$

Let the arc-length parameter of $\partial\Omega$ run clockwise. Here we have

$$\kappa_{\partial\Omega} = -\frac{1}{|\nabla\psi|}([D^2\psi]T_{\partial\Omega}, T_{\partial\Omega})_{\mathbb{R}^2},$$

where $T_{\partial\Omega}$ is the unit tangent vector of $\partial\Omega$ and $\kappa_{\partial\Omega}$ is computed in the direction of the unit normal vector $N_{\partial\Omega}$ of $\partial\Omega$, which is obtained by rotating $T_{\partial\Omega}$ with $\pi/2$. Note that $h_\pm = \kappa_{\partial\Omega}(x_*^\pm)$, and denote $T_{\partial\Omega}^\pm := T_{\partial\Omega}(x_*^\pm)$ and $N_{\partial\Omega}^\pm := N_{\partial\Omega}(x_*^\pm)$. At $\sigma = \pm l$, we observe $T_*(\pm l) = \pm N_{\partial\Omega}^\pm$ and $N_*(\pm l) = \mp T_{\partial\Omega}^\pm$. This implies that for $\sigma = l$

$$\begin{aligned}
\partial B_1(0)\rho &= -(N_{\partial\Omega}^+, \nabla\psi(x_*^+))_{\mathbb{R}^2}\rho_\sigma + (-T_{\partial\Omega}^+, [D^2\psi(x_*^+)](-T_{\partial\Omega}^+))_{\mathbb{R}^2}\rho \\
&= -|\nabla\psi(x_*^+)|\left\{\rho_\sigma + \left(-\frac{1}{|\nabla\psi(x_*^+)|}(T_{\partial\Omega}^+, [D^2\psi(x_*^+)]T_{\partial\Omega}^+)_{\mathbb{R}^2}\right)\rho\right\} \\
&= -|\nabla\psi(x_*^+)|(\rho_\sigma + h_+\rho),
\end{aligned}$$

and that for $\sigma = -l$

$$\begin{aligned}
\partial B_1(0)\rho &= -(-N_{\partial\Omega}^-, \nabla\psi(x_*^-))_{\mathbb{R}^2}\rho_\sigma + (T_{\partial\Omega}^-, [D^2\psi(x_*^-)]T_{\partial\Omega}^-)_{\mathbb{R}^2}\rho \\
&= |\nabla\psi(x_*^-)|\left\{\rho_\sigma - \left(-\frac{1}{|\nabla\psi(x_*^-)|}(T_{\partial\Omega}^-, [D^2\psi(x_*^-)]T_{\partial\Omega}^-)_{\mathbb{R}^2}\right)\rho\right\} \\
&= |\nabla\psi(x_*^-)|(\rho_\sigma - h_-\rho).
\end{aligned}$$

Consequently, we have

$$\mp\frac{1}{|\nabla\psi(x_*^\pm)|}\partial B_1(0)\rho = (\partial_\sigma \pm h_\pm)\rho \quad \text{at} \quad \sigma = \pm l.$$

Let us also derive $\partial B_2(0)$. From (3.4) we have

$$\partial B_2(0)\rho = \partial_\sigma[\partial\kappa(0)\rho] = \partial_\sigma(\partial_\sigma^2 + \kappa_*^2)\rho \quad \text{at} \quad \sigma = \pm l.$$

This completes the proof. □

By Lemmas 3.2 and 3.3 we have derived the linearization of (2.7) around $\rho \equiv 0$.

THEOREM 3.4. *The linearization of* (2.7) *around* $\rho \equiv 0$ *is as follows:*

$$\begin{cases} \rho_t = -\partial_\sigma^2(\partial_\sigma^2 + \kappa_*^2)\rho & \text{for} \quad \sigma \in (-l,l),\ t>0\,, \\ (\partial_\sigma \pm h_\pm)\rho = 0 & \text{at} \quad \sigma = \pm l, \\ \partial_\sigma(\partial_\sigma^2 + \kappa_*^2)\rho = 0 & \text{at} \quad \sigma = \pm l. \end{cases} \tag{3.5}$$

Remark 3.5. The flow (1.2) has the property that the area enclosed by the curve Γ and $\partial\Omega$ is preserved. A constraint of fixed area leads to a nonlinear constraint for ρ. If we linearize this constraint, we obtain

$$\int_{-l}^{l} \rho \, d\sigma = 0. \tag{3.6}$$

Since the original problem (1.2) has the area-preserving property, we will analyze the linearized problem (3.5) for functions ρ satisfying (3.6). In fact, the linearized operator will have one eigenvalue zero if we do not take the constraint (3.6) into account. But the eigenfunction related to the eigenvalue zero corresponds to solutions with a different mean value and is therefore not relevant for the stability (see [11], [12] for similar difficulties arising from an area constraint).

4. Gradient flow structure. The surface diffusion flow can be interpreted as the H^{-1}-gradient flow of the area functional (see [5], [13], [23]). In this section we demonstrate that the linearization (3.5) derived in section 3 can also be interpreted as a gradient flow. This observation will be important for our stability analysis.

In what follows we will need the duality pairing $\langle \cdot, \cdot \rangle$ between $(H^1(-l,l))'$ and $(H^1(-l,l))$, and we will need the following weak formulation. We denote by $\| \cdot \|_s$ the norm on $H^s(-l,l)$ where $H^0(-l,l) = L^2(-l,l)$.

DEFINITION 4.1. *We say that* $u_v \in H^1(-l,l)$ *for a given* $v \in (H^1(-l,l))'$ *with* $\langle v, 1\rangle = 0$ *is a weak solution of*

$$\begin{cases} -\partial_\sigma^2 u_v = v & \text{for} \quad \sigma \in (-l,l)\,, \\ \partial_\sigma u_v = 0 & \text{at} \quad \sigma = \pm l \end{cases} \tag{4.1}$$

if u_v *satisfies*

$$\langle v, \xi \rangle = \int_{-l}^{l} \partial_\sigma u_v \partial_\sigma \xi$$

for all $\xi \in H^1(-l,l)$.

DEFINITION 4.2. *For a given* $v \in (H^1(-l,l))'$ *with* $\langle v, 1\rangle = 0$, *we say that* $\rho \in H^3(-l,l)$ *with* $\int_{-l}^{l} \rho = 0$ *is a weak solution of the boundary value problem*

$$\begin{cases} v = -\partial_\sigma^2(\partial_\sigma^2 + \kappa_*^2)\rho & \text{for} \quad \sigma \in (-l,l)\,, \\ (\partial_\sigma \pm h_\pm)\rho = 0 & \text{at} \quad \sigma = \pm l\,, \\ \partial_\sigma(\partial_\sigma^2 + \kappa_*^2)\rho = 0 & \text{at} \quad \sigma = \pm l \end{cases} \tag{4.2}$$

if ρ *satisfies*

$$\langle v, \xi \rangle = \int_{-l}^{l} \partial_\sigma(\partial_\sigma^2 + \kappa_*^2)\rho \, \partial_\sigma \xi \quad \text{and} \quad (\partial_\sigma \pm h_\pm)\rho = 0 \ \text{ at } \ \sigma = \pm l$$

for all $\xi \in H^1(-l,l)$.

In the case that $v \in L^2(-l,l)$ we obtain that $v = -\partial_\sigma^2(\partial_\sigma^2 + \kappa_*^2)\rho$ is fulfilled almost everywhere in $(-l,l)$ and $\partial_\sigma(\partial_\sigma^2 + \kappa_*^2)\rho = 0$ is fulfilled for $\sigma = \pm l$.

In addition we also need the symmetric bilinear form on $H^1(-l,l)$,

$$I(\rho_1,\rho_2) := \int_{-l}^{l} \{\partial_\sigma\rho_1\partial_\sigma\rho_2 - \kappa_*^2\rho_1\rho_2\}d\sigma + h_+\rho_1(l)\rho_2(l) + h_-\rho_1(-l)\rho_2(-l),$$

and the inner product

$$(\rho_1,\rho_2)_{-1} := \int_{-l}^{l} \partial_\sigma u_{\rho_1}\partial_\sigma u_{\rho_2},$$

where $u_{\rho_i} \in H^1(-l,l)$ for a given $\rho_i \in (H^1(-l,l))'$ with $\langle \rho_i, 1\rangle = 0$ is defined as the weak solution of (4.1). The bilinear form I is defined on $H^1(-l,l)$ and the inner product $(\cdot\,,\cdot)_{-1}$ is defined for all pairs of elements in $(H^1(-l,l))'$ with $\langle \rho_i, 1\rangle = 0$. We remark that by Definition 4.1

$$(\rho_1,\rho_2)_{-1} = \langle \rho_1, u_{\rho_2}\rangle \tag{4.3}$$

holds for $\rho_i \in (H^1(-l,l))'$ with $\langle \rho_i, 1\rangle = 0$.

Now we are going to show that the linearized problem (3.5) is the gradient flow of $E(\rho) := I(\rho,\rho)/2$ with respect to the H^{-1} inner product $(\cdot,\cdot)_{-1}$. Let us review the concept of gradient flows. For a given functional E on a linear space X and an inner product $(\cdot,\cdot)_X$ on X, we say that a time-dependent function ρ with values in X is a solution of the gradient flow equation to E and $(\cdot,\cdot)_X$ if and only if

$$(\rho_t(t),\xi)_X = -\partial E(\rho(t))(\xi)$$

holds for all $\xi \in X$ and all t. Here $\partial E(\rho(t))(\xi)$ denotes the derivative of E at the point $\rho(t)$ in the direction ξ. The fact that the linearized problem (3.5) is the gradient flow of $I(\rho,\rho)/2$ with respect to the $(\cdot,\cdot)_{-1}$ inner product follows from the following lemma. This is true since the derivative of $E(\rho) = I(\rho,\rho)/2$ in a direction ξ is given by $I(\rho,\xi)$.

LEMMA 4.3. *Let* $v \in (H^1(-l,l))'$ *with* $\langle v, 1\rangle = 0$ *be given. Then a function* $\rho \in H^3(-l,l)$ *with* $\int_{-l}^{l}\rho = 0$ *is a weak solution of* (4.2) *if and only if*

$$(v,\xi)_{-1} = -I(\rho,\xi)$$

holds for all $\xi \in H^1(-l,l)$ *with* $\int_{-l}^{l}\xi = 0$.

Proof. Let $\rho \in H^3(-l,l)$ be a weak solution of (4.2). By (4.3) and Definition 4.2, we have

$$(v,\xi)_{-1} = \langle v, u_\xi\rangle = \int_{-l}^{l} \partial_\sigma(\partial_\sigma^2 + \kappa_*^2)\rho\,\partial_\sigma u_\xi$$

for all $\xi \in H^1(-l,l)$ with $\int_{-l}^{l}\xi = 0$. Note that $u_\xi \in H^1(-l,l)$ is a weak solution of (4.1) with $\xi \in H^1(-l,l)$. Then, by virtue of $(\partial_\sigma^2 + \kappa_*^2)\rho \in H^1(-l,l)$, we see

$$\int_{-l}^{l} \partial_\sigma(\partial_\sigma^2 + \kappa_*^2)\rho\,\partial_\sigma u_\xi = \int_{-l}^{l} (\partial_\sigma^2 + \kappa_*^2)\rho\,\xi.$$

This implies that

$$\begin{aligned}(v,\xi)_{-1} &= \int_{-l}^{l} (\partial_\sigma^2 + \kappa_*^2)\rho\,\xi \\ &= -\int_{-l}^{l} (\partial_\sigma \rho\,\partial_\sigma \xi - \kappa_*^2 \rho\,\xi) + [\partial_\sigma \rho\,\xi]_{\sigma=-l}^{\sigma=l} \\ &= -I(\rho,\xi)\,.\end{aligned}$$

The last equality is shown by using $(\partial_\sigma \pm h_\pm)\rho = 0$ at $\sigma = \pm l$.

Conversely, assume that $\rho \in H^1(-l,l)$ with $\int_{-l}^{l} \rho = 0$ satisfies

$$(v,\xi)_{-1} = -I(\rho,\xi) \tag{4.4}$$

for all $\xi \in H^1(-l,l)$ with $\int_{-l}^{l} \xi = 0$. Choose $\xi = -\partial_\sigma^2 \eta$ in (4.4) for a given function $\eta \in H^3(-l,l)$ with $\partial_\sigma \eta = 0$ at $\sigma = \pm l$. Then it holds that

$$\begin{aligned}\langle v,\eta\rangle &= (v,\xi)_{-1} \\ &= -I(\rho,\xi) \\ &= -\int_{-l}^{l} (\partial_\sigma \rho\,\partial_\sigma \xi - \kappa_*^2 \rho\,\xi) - \{h_+\rho(l)\xi(l) + h_-\rho(-l)\xi(-l)\} \\ &= -\int_{-l}^{l} (-\partial_\sigma \rho\,\partial_\sigma^3 \eta + \kappa_*^2 \rho\,\partial_\sigma^2 \eta) + \{h_+\rho(l)\partial_\sigma^2\eta(l) + h_-\rho(-l)\partial_\sigma^2\eta(-l)\}\,.\end{aligned}$$

Since $v \in (H^1(-l,l))'$, we deduce from the above identity that $\rho \in H^3(-l,l)$. Integration by parts gives

$$\begin{aligned}\langle v,\eta\rangle &= \int_{-l}^{l} (-\partial_\sigma^2 \rho\,\partial_\sigma^2 \eta + \kappa_*^2 \partial_\sigma \rho\,\partial_\sigma \eta) + [(\partial_\sigma \rho \pm h_\pm \rho)\partial_\sigma^2 \eta]_{\sigma=-l}^{\sigma=l} \\ &= \int_{-l}^{l} \partial_\sigma(\partial_\sigma^2 + \kappa_*^2)\rho\,\partial_\sigma \eta + [(\partial_\sigma \rho \pm h_\pm \rho)\partial_\sigma^2 \eta]_{\sigma=-l}^{\sigma=l}\,,\end{aligned} \tag{4.5}$$

where $[(\partial_\sigma \rho \pm h_\pm \rho)\partial_\sigma^2 \eta]_{\sigma=-l}^{\sigma=l} = (\partial_\sigma \rho + h_+\rho)\partial_\sigma^2\eta|_{\sigma=l} - (\partial_\sigma \rho - h_-\rho)\partial_\sigma^2\eta|_{\sigma=-l}$. Since $\partial_\sigma^2 \eta$ can be chosen arbitrarily at $\sigma = \pm l$ and v is a bounded linear functional on $H^1(-l,l)$, we can deduce that the first boundary condition $(\partial_\sigma \pm h_\pm)\rho = 0$ at $\sigma = \pm l$ holds. The remaining identity in (4.5) then is a weak formulation of $v = -\partial_\sigma^2(\partial_\sigma^2 + \kappa_*^2)\rho$ for $\sigma \in (-l,l)$ together with $\partial_\sigma(\partial_\sigma^2 + \kappa_*^2)\rho = 0$ at $\sigma = \pm l$ (see Definition 4.2). ∎

5. Self-adjointness of the linearized operator. It is the aim of this section to show that the linearized operator is self-adjoint and to study its spectrum. A nonlinear stability analysis will most likely involve spaces of functions that are differentiable in a classical sense. For analyzing the spectrum it will be more appropriate to use spaces involving functions that are differentiable in a weak sense. Since eigenfunctions will be smooth, the spectrum will not depend on the domain of definition as long as the boundary conditions are incorporated correctly. Therefore, choosing an appropriate domain of definition, the linearized operator of (3.5) is given by

$$\mathcal{A} : \mathcal{D}(\mathcal{A}) \to H$$

with

$$\begin{cases} \mathcal{D}(\mathcal{A}) = \left\{ \rho \in H^3(-l,l) \mid (\partial_\sigma \pm h_\pm)\rho = 0 \ \text{ at } \ \sigma = \pm l \ \text{ and } \ \displaystyle\int_{-l}^{l} \rho = 0 \right\}, \\ H = \{\rho \in (H^1(-l,l))' \mid \langle \rho, 1 \rangle = 0\} \end{cases}$$

and

$$\langle \mathcal{A}\rho, \xi \rangle := \int_{-l}^{l} \partial_\sigma (\partial_\sigma^2 + \kappa_*^2)\rho \, \partial_\sigma \xi \, . \tag{5.1}$$

Then the boundary value problem (4.2) corresponds to the problem of finding a $\rho \in \mathcal{D}(\mathcal{A})$ with

$$\mathcal{A}\rho = v \, .$$

We also remark that this definition gives, for all $\xi \in H^1(-l,l)$ with $\int_{-l}^{l} \xi = 0$,

$$(\mathcal{A}\rho, \xi)_{-1} = -I(\rho, \xi).$$

For this operator $\mathcal{A}$, we have the following lemma.

LEMMA 5.1. *The operator $\mathcal{A}$ is symmetric with respect to the inner product $(\cdot,\cdot)_{-1}$.*

Proof. For all $\rho, \xi \in \mathcal{D}(\mathcal{A})$ we have

$$(\mathcal{A}\rho, \xi)_{-1} = -I(\rho, \xi) = -I(\xi, \rho) = (\mathcal{A}\xi, \rho)_{-1} = (\rho, \mathcal{A}\xi)_{-1},$$

so that $\mathcal{A}$ is symmetric. ∎

We need to analyze the spectrum of $\mathcal{A}$ in order to decide on the stability behavior of the linearized problem (3.5). Using classical principles of the variational calculus, we can describe the spectrum of $\mathcal{A}$ with the help of the inner product $(\cdot\,,\cdot)_{-1}$ and I. In fact, if ρ is an eigenfunction to the eigenvalue λ, it holds that

$$\lambda(\rho, \xi)_{-1} = (\mathcal{A}\rho, \xi)_{-1} = -I(\rho, \xi) \, .$$

We remark that eigenvalues $\lambda \neq 0$ always correspond to eigenfunctions that have the mean value zero. This follows by integrating the identity

$$-\partial_\sigma^2(\partial_\sigma^2 + \kappa_*^2)\rho = \lambda\rho$$

and using the boundary conditions. In what follows we will study only eigenvalues which have eigenfunctions with mean value zero. This is a natural request for the linearized problem. It follows when we take the mass constraint in the nonlinear problem into account. This makes sense because the surface diffusion flow is mass preserving (cf. [14]).

Therefore we define $V = \{\rho \in H^1(-l,l) \mid \int_{-l}^{l} \rho = 0\}$. The following two lemmas will be needed to show the boundness of the eigenvalue from above.

LEMMA 5.2. *For all $\delta > 0$ there exists a C_δ such that for all functions $\rho \in V$ the inequality*

$$\rho(l)^2 \leq \delta \|\partial_\sigma \rho\|_0^2 + C_\delta \|\rho\|_{-1}^2$$

holds. The same inequality holds for $\rho(-l)^2$ instead of $\rho(l)^2$.

Proof. We prove the assertion by contradiction. Assume that there exists a $\delta > 0$ such that for all $n \in \mathbb{N}$, $\rho_n \in V$ with $\rho_n(l)^2 = 1$ satisfy

$$1 = \rho_n(l)^2 > \delta \|\partial_\sigma \rho_n\|_0^2 + n\|\rho_n\|_{-1}^2 .$$

This implies

$$\|\rho_n\|_{-1}^2 < \frac{1}{n} \to 0 \qquad \text{as } n \to \infty$$

and

$$\|\partial_\sigma \rho_n\|_0^2 < \frac{1}{\delta} .$$

Since $\int_{-l}^{l} \rho_n = 0$, we conclude from Poincaré's inequality that ρ_n is bounded uniformly in $H^1(-l, l)$. This gives

$$\rho_n \to 0 \quad \text{weakly in } H^1(-l, l)$$

and therefore (since the embedding $H^1(-l, l)$ into $C^0([-l, l])$ is compact)

$$\rho_n(l) \to 0 .$$

This is a contradiction, and therefore the lemma is shown. □

LEMMA 5.3. *There exist positive constants c_1 and c_2 such that*

$$\|\rho\|_1^2 \le c_1 \|\rho\|_{-1}^2 + c_2 I(\rho, \rho) \ \textit{for all } \rho \in V .$$

Proof. Since the embedding $H^1(-l, l) \hookrightarrow L^2(-l, l)$ is compact, we obtain that for all $\delta > 0$ there exists a $\hat{C}_\delta > 0$ such that

$$\|\rho\|_0^2 \le \delta \|\partial_\sigma \rho\|_0^2 + \hat{C}_\delta \|\rho\|_{-1}^2 .$$

This can, for example, be shown in exactly the same manner as in the proof of the preceding lemma. Therefore we obtain, with the help of Lemma 5.2 and the above inequality,

$$\begin{aligned} I(\rho, \rho) &= \int_{-l}^{l} |\partial_\sigma \rho|^2 - \kappa_*^2 \int_{-l}^{l} \rho^2 + h_+ \rho(l)^2 + h_- \rho(-l)^2 \\ &\ge \int_{-l}^{l} |\partial_\sigma \rho|^2 - \kappa_*^2 \int_{-l}^{l} \rho^2 - |h_+| \rho(l)^2 - |h_-| \rho(-l)^2 \\ &\ge (1 - \varepsilon) \int_{-l}^{l} |\partial_\sigma \rho|^2 - C_\varepsilon \|\rho\|_{-1}, \end{aligned}$$

which holds for suitable ε and C_ε. The above inequality proves the lemma. □

COROLLARY 5.4. *The largest eigenvalue of $\mathcal{A}$ is bounded from above by c_1/c_2.*

Proof. Let λ be an eigenvalue of $\mathcal{A}$. Then there exists a $\rho \neq 0$ such that

$$\lambda(\rho, \rho)_{-1} = -I(\rho, \rho) .$$

Assume $\lambda > c_1/c_2$. This implies

$$0 = I(\rho, \rho) + \lambda(\rho, \rho)_{-1} > I(\rho, \rho) + c_1/c_2 (\rho, \rho)_{-1} \ge 1/c_2 \|\rho\|_1^2 > 0,$$

which is a contradiction. □

By virtue of Lemma 5.1 and Corollary 5.4, we have following theorem.

THEOREM 5.5. (i) *The operator $\mathcal{A}$ is self-adjoint with respect to the inner product $(\cdot,\cdot)_{-1}$.*

(ii) *The spectrum of $\mathcal{A}$ contains a countable system of real eigenvalues.*

(iii) *The initial value problem* (3.5) *is solvable for initial data in H.*

(iv) *The zero solution is an asymptotically stable solution of* (3.5) *if and only if the largest eigenvalue of $\mathcal{A}$ is negative.*

Proof. First we show that the resolvent $(\mathcal{A}-\omega)^{-1}$ exists for some $\omega \in \mathbb{R}$. Choosing $\omega > c_1/c_2$ and using Corollary 5.4, we know that $\mathcal{A}-\omega$ is injective. It remains to show that $\mathcal{A}-\omega$ is surjective. For a given $f \in H$ we need to prove that there exists a weak solution ρ of the boundary value problem

$$\left\{\begin{array}{lll} -\partial_\sigma^2\{-(\partial_\sigma^2+\kappa_*^2)\}\rho+\omega\rho=f & \text{for} & \sigma\in(-l,l),\\ (\partial_\sigma \pm h_\pm)\rho=0 & \text{at} & \sigma=\pm l,\\ \partial_\sigma(\partial_\sigma^2+\kappa_*^2)\rho=0 & \text{at} & \sigma=\pm l. \end{array}\right. \tag{5.2}$$

To obtain a solution to (5.2) we use the fact that the minimizing problem

$$F(\rho) := \int_{-l}^{l}\Big(\frac{1}{2}|\partial_\sigma\rho|^2-\frac{1}{2}\kappa_*^2\rho^2\Big)+\frac{1}{2}h_+\rho^2(l)+\frac{1}{2}h_-\rho^2(-l)+\frac{\omega}{2}\|\rho\|_{-1}^2-\int_{-l}^{l}u_f\rho\to\min$$

under all $\rho\in H^1(-l,l)$ with $\int_{-l}^{l}\rho=0$ admits as solutions $\tilde\rho$. This holds since F is coercive, which follows from Lemmas 5.2 and 5.3. Taking the first variation of F, we observe that

$$\left\{\begin{array}{lll} -(\partial_\sigma^2+\kappa_*^2)\tilde\rho+\omega u_{\tilde\rho}=u_f & \text{for} & \sigma\in(-l,l),\\ (\partial_\sigma\pm h_\pm)\tilde\rho=0 & \text{at} & \sigma=\pm l \end{array}\right. \tag{5.3}$$

holds in a weak sense. Since $u_{\tilde\rho}$, $u_f\in H^1(-l,l)$, we have $\tilde\rho\in H^3(-l,l)$. Furthermore, it follows from $\partial_\sigma u_{\tilde\rho}=\partial_\sigma u_f=0$ at $\sigma=\pm l$ that $\partial_\sigma(\partial_\sigma^2+\kappa_*^2)\bar\rho=0$ at $\sigma=\pm l$. Taking second derivatives of (5.3) in a weak sense, we derive that $\tilde\rho$ solves (5.2). This shows that $\mathcal{A}-\omega$ is surjective and hence $(\mathcal{A}-\omega)^{-1}$ exists.

Let us prove (i). We already know from Lemma 5.1 that $\mathcal{A}$ is symmetric. Since the self-adjointness of $\mathcal{A}$ follows from the self-adjointness of $\mathcal{A}-\omega$ for some $\omega\in\mathbb{R}$, we show the self-adjointness of $\mathcal{A}-\omega$. Suppose that there are v, $w\in H$ such that

$$((\mathcal{A}-\omega)\rho,v)_{-1}=(\rho,w)_{-1} \tag{5.4}$$

for all $\rho\in D(\mathcal{A}-\omega)$. By the above argument $\mathcal{A}-\omega$ is invertible if ω is large enough. Then there exists a $z\in D(\mathcal{A}-\omega)$ such that

$$(\mathcal{A}-\omega)z=w \tag{5.5}$$

for sufficiently large ω. By (5.4), (5.5), and Lemma 5.1, we have

$$((\mathcal{A}-\omega)\rho,v)_{-1}=(\rho,(\mathcal{A}-\omega)z)_{-1}=((\mathcal{A}-\omega)\rho,z)_{-1}.$$

Since $\mathcal{A}-\omega$ is surjective, we obtain $v=z$. This implies that $v\in D(\mathcal{A}-\omega)$ and

$$(\mathcal{A}-\omega)v=w,$$

so that $\mathcal{A}-\omega$ is self-adjoint.

Since $(\mathcal{A}-\omega)^{-1}$ exists and is compact, (ii) follows from [14, Theorem 6.29, Chapter 3] and the fact that $\mathcal{A}$ is self-adjoint.

Using the fact that $\mathcal{A}$ is a self-adjoint operator on H, the theory of semigroups is applicable to show (iii) (see, e.g., the functional calculus in sections 5.8–5.10 of [24]). Semigroup theory also gives (iv). ∎

To decide on the linearized stability, it will be important to know that the eigenvalues of $\mathcal{A}$ depend continuously on h_+, h_-, and κ_*^2 and are also monotone in each of these parameters. The following lemma ensures these properties.

LEMMA 5.6. *Let*

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots$$

be the eigenvalues of $\mathcal{A}$ (taking the multiplicity into account).

(i) *Then it holds for all $n \in \mathbf{N}$ that*

$$-\lambda_n = \inf_{W\in\Sigma_n} \sup_{u\in W\setminus\{0\}} \frac{I(u,u)}{(u,u)_{-1}},$$

$$-\lambda_n = \sup_{W\in\Sigma_{n-1}} \inf_{u\in W^\perp\setminus\{0\}} \frac{I(u,u)}{(u,u)_{-1}}.$$

Here Σ_n is the collection of n-dimensional subspaces of V and $W^\perp$ is the orthogonal complement with respect to the $(\cdot\,,\cdot)_{-1}$-scalar product.

(ii) *The eigenvalues λ_n depend continuously on h_+, h_-, and κ_*^2 and are monotone decreasing in each of the parameters h_+, h_-, and $(-\kappa_*^2)$.*

Proof. The lemma follows with the help of Courant's maximum-minimum principle, together with the fact that I depends in a monotone and continuous way on h_+, h_-, and $(-\kappa_*^2)$. The proof follows the lines of Courant and Hilbert [4, section VI.2]. ∎

6. Stability analysis. To obtain a linearized stability result for stationary solutions of (2.7), it is enough to show that $I(\rho,\rho)$ is positive for all $\rho \in V\setminus\{0\}$. Then $\lambda_1 < 0$, which implies stability. This is true since λ_1 allows the characterization

$$-\lambda_1 = \inf_{\rho\in V\setminus\{0\}} \frac{I(\rho,\rho)}{(\rho,\rho)_{-1}},$$

and the infimum is in fact a minimum; therefore it is enough to show the positivity of I pointwise.

In the following arguments we consider only the case $\kappa_* > 0$ (or $\kappa_* = 0$). We remark that the same result is derived for $\kappa_* < 0$. Also note that the stationary solution is a part of a circle with radius κ_*. The length of the stationary solution is $2l$, and therefore the restriction

$$2l < \frac{2\pi}{\kappa_*},$$

which gives $\kappa_* l < \pi$, has to hold.

Now the following lemma shows that for given κ_* the stationary solution is always stable, provided h_+, h_- are large enough.

LEMMA 6.1. *Let $\kappa_* l < \pi$. Then there exists a constant $K > 0$, such that*

$$I(\rho,\rho) > 0 \quad \textit{for all} \quad \rho \in V\setminus\{0\},$$

provided that $h_+, h_- > K$.

Proof. Using the transformation

$$u(s) = \rho\Big(\frac{2l}{\pi}s - l\Big)$$

and the fact that $\kappa_* l < \pi$, it is enough to show that there exists a constant $\overline{c} > 0$ such that

$$\|u'\|_0^2 - 4\|u\|_0^2 + \overline{c}(u(0)^2 + u(\pi)^2) \geq 0$$

for all $u \in H^1(0,\pi)$ with $\int_0^\pi u = 0$. Assume that such a constant $\overline{c}$ does not exist. Then there exists a sequence u_n (without loss of generality we assume $\|u_n\|_0^2 = 1$) such that

$$\|u_n'\|_0^2 - 4\|u_n\|_0^2 + n(u_n^2(0) + u_n^2(\pi)) < 0\,.$$

This implies

$$\|u_n'\|_0^2 \leq 4,$$

and we deduce the existence of a subsequence (which we also label by $\{u_n\}_{n\in\mathbf{N}}$) such that

$$\begin{array}{lcll} u_n' & \to & u' & \text{weakly in } L^2(0,\pi)\,, \\ u_n & \to & u & \text{strongly in } L^2(0,\pi)\,, \\ u_n & \to & u & \text{strongly in } C^0([0,\pi])\,. \end{array}$$

Then

$$u_n^2(0) + u_n^2(\pi) \leq \frac{4}{n}$$

implies

$$u(0) = u(\pi) = 0\,.$$

The lower semicontinuity of the L^2-norm under weak convergence implies

$$\|u'\|_0^2 < 4\|u\|_0^2\,,$$

which contradicts the facts that $u \in \overset{\circ}{H}{}^1(0,\pi)$ and $\int_0^\pi u = 0$ (see the following lemma). This proves the lemma. □

LEMMA 6.2. *For all* $u \in \overset{\circ}{H}{}^1(0,\pi)$ *with* $\int_0^\pi u(s)ds = 0$, *it holds that*

$$\|u\|_0^2 \leq \frac{1}{4}\|u'\|_2^2\,.$$

Proof. Each $u \in \overset{\circ}{H}{}^1(0,\pi)$ has a representation

$$u(s) = \sum_{k=1}^{\infty} a_k \sin ks\,.$$

Then we have

$$\|u\|_0^2 = \frac{\pi}{2}\sum_{k=1}^{\infty} a_k^2\,, \quad \|u'\|_0^2 = \frac{\pi}{2}\sum_{k=1}^{\infty} k^2 a_k^2.$$

In addition, the assumption $\int_0^\pi u(s)ds = 0$ implies

$$\sum_{\substack{k=1\\ k\,\text{odd}}}^{\infty} \frac{2}{k} a_k = 0\,. \tag{6.1}$$

Now we readily see

$$\sum_{\substack{k=1\\ k\,\text{even}}}^{\infty} a_k^2 \le \frac{1}{4}\sum_{\substack{k=1\\ k\,\text{even}}}^{\infty} k^2 a_k^2\,.$$

It remains to estimate the sum over all odd k, which would follow from

$$3a_1^2 \le \sum_{\substack{k=3\\ k\,\text{odd}}}^{\infty} (k^2-4) a_k^2\,. \tag{6.2}$$

The mean value constraint (6.1) implies

$$a_1 = -\sum_{\substack{k=3\\ k\,\text{odd}}}^{\infty} \frac{1}{k} a_k,$$

which gives

$$3a_1^2 \le 3\Big(\sum_{\substack{k=3\\ k\,\text{odd}}}^{\infty} \frac{1}{k} a_k\Big)^2 \le 3\Big(\sum_{\substack{k=3\\ k\,\text{odd}}}^{\infty} a_k^2\Big)\Big(\sum_{\substack{k=3\\ k\,\text{odd}}}^{\infty} \frac{1}{k^2}\Big) = 3\sum_{\substack{k=3\\ k\,\text{odd}}}^{\infty} a_k^2 \cdot \Big(\frac{\pi^2}{8} - 1\Big)\,.$$

Since $3(\pi^2/8 - 1) < k^2 - 4$ $(k = 3, 5, \dots)$, the inequality (6.2) is derived. Thus the lemma follows. □

The strategy now is as follows. We know that for large h_+ and h_- we have stability. In addition we know that the eigenvalues depend in a monotone and continuous way on h_+ and h_-. If we start with a stable situation $(h_+, h_- \gg 1)$ and decrease h_+ and, respectively, h_-, a loss of stability can therefore occur only in the case that the largest eigenvalue λ_1 passes through zero. For that reason we analyze for which values of h_+, h_-, and κ_* a zero eigenvalue is possible. To obtain a complete picture about the dimension of the unstable manifold, we also determine the multiplicity of a possible zero eigenvalue.

LEMMA 6.3. (i) *Assume that $\kappa_* \neq 0$ and $\kappa_* l < \pi$. Then the operator $\mathcal{A}$ has a zero eigenvalue if and only if*

$$\frac{a}{c} + \frac{b}{c}(h_+ + h_-) + h_+ h_- = 0, \tag{6.3}$$

where

$$\begin{aligned} a &= -2\kappa_*^2 l \sin(\kappa_* l)\cos(\kappa_* l)\,,\\ b &= \kappa_* l(\cos^2(\kappa_* l) - \sin^2(\kappa_* l)) - \sin(\kappa_* l)\cos(k_* l)\,,\\ c &= 2\Big\{-\frac{1}{\kappa_*}\sin^2(\kappa_* l) + l\sin(\kappa_* l)\cos(\kappa_* l)\Big\}\,. \end{aligned}$$

Furthermore, the following inequality holds:

$$\frac{b^2}{c^2} - \frac{a}{c} > 0\,. \tag{6.4}$$

(ii) *If $\kappa_* = 0$, then the operator $\mathcal{A}$ has a zero eigenvalue if and only if*

$$\frac{3}{l^2} + \frac{2}{l}(h_+ + h_-) + h_+h_- = 0\,. \tag{6.5}$$

(iii) *If we interpret a, b, and c as functions of κ_*, we obtain*

$$\frac{a}{c} \to \frac{3}{l^2} \quad \textit{and} \quad \frac{b}{c} \to \frac{2}{l} \quad \textit{as} \quad \kappa_* \to 0\,.$$

(iv) *The multiplicity of a possible zero eigenvalue is equal to one for all h_+, h_-, and κ_*.*

In what follows we set

$$\mathcal{D}(h_+, h_-, \kappa_*) = \frac{a}{c} + \frac{b}{c}(h_+ + h_-) + h_+h_-$$

for all h_+, h_-, and κ_*. The extension to $\kappa_* = 0$ is well defined by the preceding lemma.

Remark 6.4. (a) The equations (6.3) and (6.5) define hyperbolas in the (h_-, h_+)-plane (see Figures 1–5). The hyperbolas are symmetric with respect to the $h_- = h_+$ line, and the inequality (6.4) implies that the line defined by $h_+ = h_-$ always has two intersection points with the hyperbolas.

(b) From (iii) in the preceding lemma we can conclude that the hyperbolas obtained for the case $\kappa_* > 0$ tend to the one for $\kappa_* = 0$.

Proof of Lemma 6.3. (i) Assume that $-\partial_\sigma^2(\partial_\sigma^2 + \kappa_*^2)\rho = 0$. Then the function ρ can be denoted by

$$\rho(\sigma) = \alpha_1\sigma + \alpha_0 + \alpha_c\cos(\kappa_*\sigma) + \alpha_s\sin(\kappa_*\sigma)$$

for constants $(\alpha_1, \alpha_0, \alpha_c, \alpha_s)$. By the boundary conditions $\partial_\sigma(\partial_\sigma^2+\kappa_*^2)\rho = 0$ at $\sigma = \pm l$, we have

$$\pm\alpha_c\kappa_*^3\sin(\kappa_* l) - \alpha_s\kappa_*^3\cos(\kappa_* l) + \kappa_*^2\{\alpha_1 \mp \alpha_c\kappa_*\sin(\kappa_* l) + \alpha_s\kappa_*\cos(\kappa_* l)\} = 0.$$

This implies that $\kappa_*^2\alpha_1 = 0$, so that $\alpha_1 = 0$. Using the boundary conditions $(\partial_\sigma \pm h_\pm)\rho = 0$ at $\sigma = \pm l$, we derive

$$\begin{cases} h_+\alpha_0 + (-\kappa_*\sin(\kappa_* l) + h_+\cos(\kappa_* l))\alpha_c + (\kappa_*\cos(\kappa_* l) + h_+\sin(\kappa_* l))\alpha_s = 0, \\ -h_-\alpha_0 + (\kappa_*\sin(\kappa_* l) - h_-\cos(\kappa_* l))\alpha_c + (\kappa_*\cos(\kappa_* l) + h_-\sin(\kappa_* l))\alpha_s = 0. \end{cases}$$

Moreover, it follows from $\int_{-l}^{l}\rho = 0$ that

$$2l\alpha_0 + \Big\{\frac{2}{\kappa_*}\sin(\kappa_* l)\Big\}\alpha_c = 0.$$

Let us define the 3×3-matrix $M(h_+, h_-, \kappa_*)$ as

$$M(h_+, h_-, \kappa_*) := \begin{pmatrix} h_+ & -\kappa_*\sin(\kappa_* l) + h_+\cos(\kappa_* l) & \kappa_*\cos(\kappa_* l) + h_+\sin(\kappa_* l) \\ -h_- & \kappa_*\sin(\kappa_* l) - h_-\cos(\kappa_* l) & \kappa_*\cos(\kappa_* l) + h_-\sin(\kappa_* l) \\ l & \{\sin(\kappa_* l)\}/\kappa_* & 0 \end{pmatrix}.$$

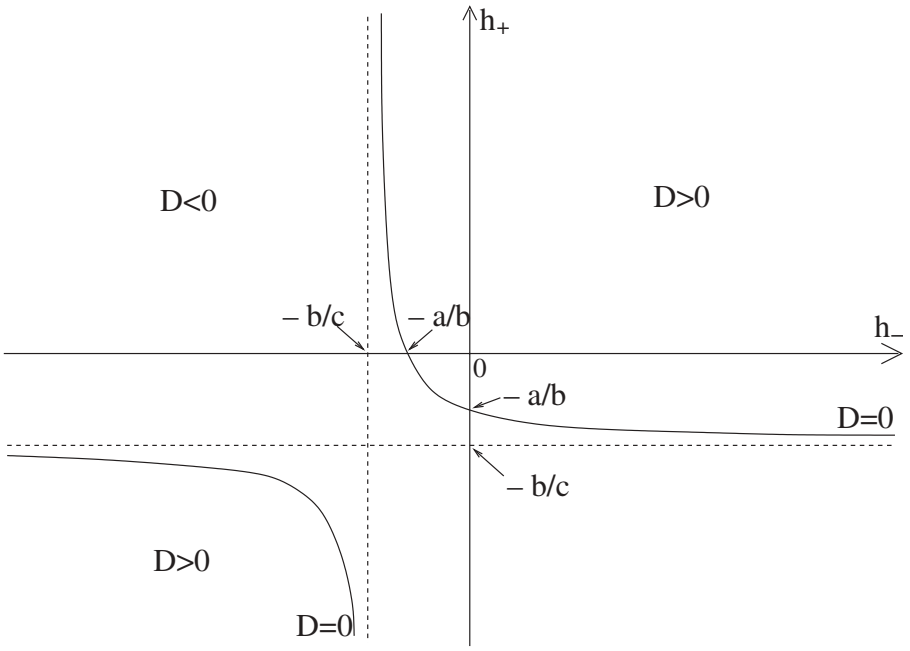

FIG. 1. $\kappa_* l < \pi/2, a < 0, b < 0, c < 0.$

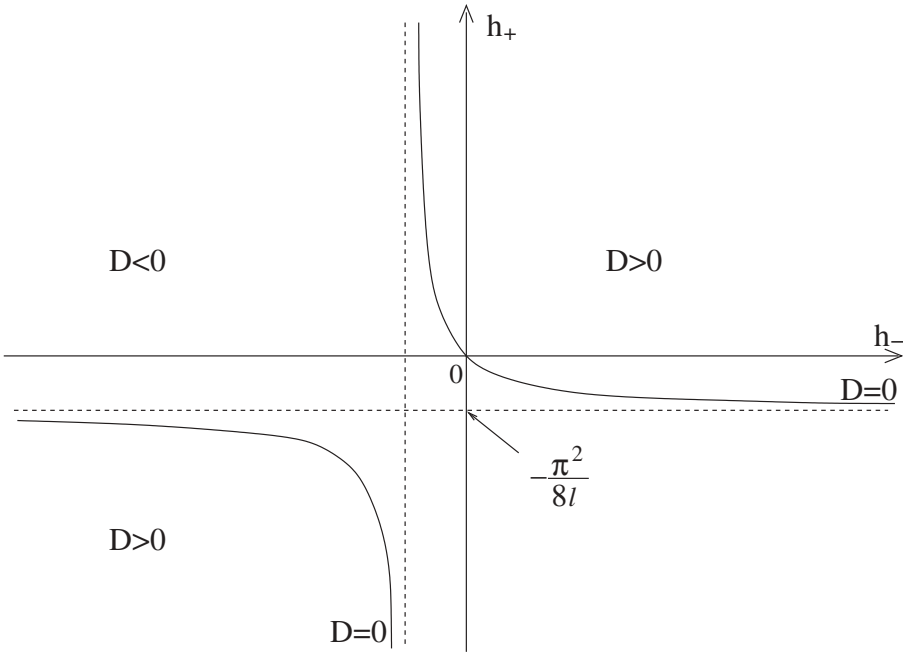

FIG. 2. $\kappa_* l = \pi/2, a = 0, b = -\kappa_* l, c = -2/\kappa_*.$

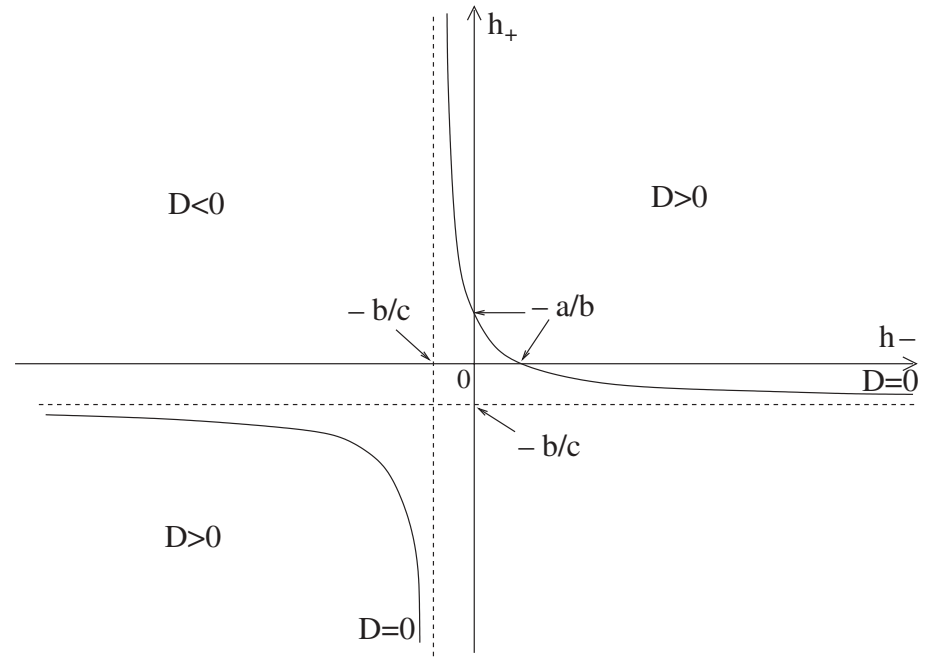

FIG. 3. $\kappa_* l > \pi/2, a > 0, b < 0, c < 0$.

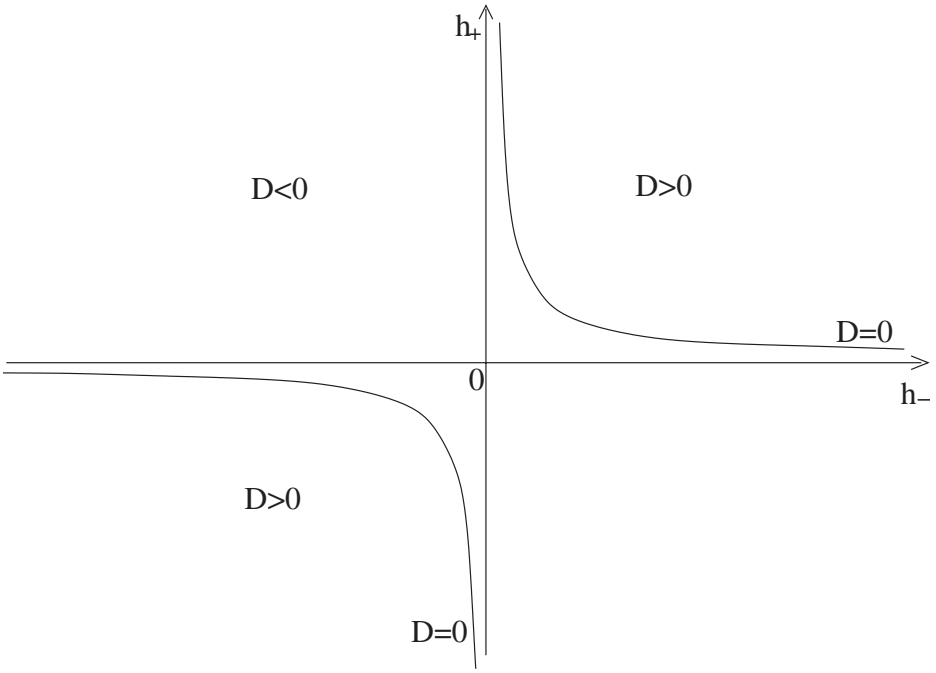

FIG. 4. $\kappa_* l > \pi/2, a > 0, b = 0, c < 0$.

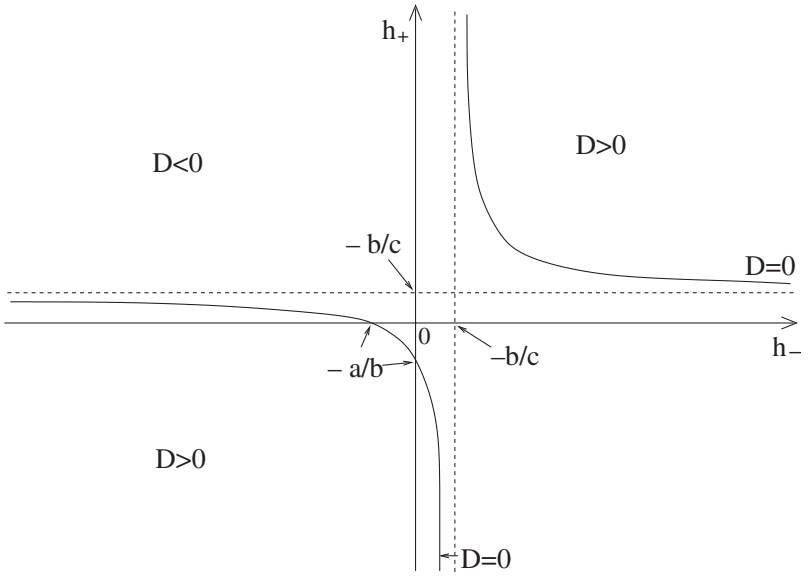

FIG. 5. $\kappa_* l > \pi/2, a > 0, b > 0, c < 0$.

Then the operator $\mathcal{A}$ has a zero eigenvalue if and only if the equation

$$M(h_+, h_-, \kappa_*)\, {}^t(\alpha_0, \alpha_c, \alpha_s) = {}^t(0, 0, 0) \tag{6.6}$$

has nonzero solutions ${}^t(\alpha_0, \alpha_c, \alpha_s)$. Nonzero solutions of (6.6) are derived when

$$\det M(h_+, h_-, \kappa_*) = 0,$$

which implies (6.3). Furthermore, by the definition of a, b, c, we have

$$b^2 - ac = \{\kappa_* l - \sin(\kappa_* l)\cos(\kappa_* l)\}^2 = \frac{1}{4}\{2\kappa_* l - \sin(2\kappa_* l)\}^2 \geq 0.$$

It follows from $\kappa_* l \neq 0$ that $2\kappa_* l - \sin(2\kappa_* l) \neq 0$. This implies (6.4).

(ii) Assume that $-\partial_\sigma^4 \rho = 0$. Then the function ρ can be denoted by

$$\rho(\sigma) = \alpha_3 \sigma^3 + \alpha_2 \sigma^2 + \alpha_1 \sigma + \alpha_0$$

for constants $(\alpha_3, \alpha_2, \alpha_1, \alpha_0)$. By the boundary conditions $\partial_\sigma^3 \rho = 0$ at $\sigma = \pm l$, we have $\alpha_3 = 0$. In addition, the conditions $(\partial_\sigma \pm h_\pm)\rho = 0$ at $\sigma = \pm l$ and $\int_{-l}^{l} \rho = 0$ give the equation

$$M_0(h_+, h_-)\, {}^t(\alpha_2, \alpha_1, \alpha_0) = {}^t(0, 0, 0), \tag{6.7}$$

where the 3×3-matrix $M_0(h_+, h_-)$ is defined as

$$M_0(h_+, h_-) := \begin{pmatrix} 2l + h_+ l^2 & 1 + h_+ l & h_+ \\ -2l - h_- l^2 & 1 + h_- l & -h_- \\ l^2/3 & 0 & 1 \end{pmatrix}.$$

Applying an argument similar to that of the proof of (i), we find that the operator $\mathcal{A}$ with $\kappa_* = 0$ has a zero eigenvalue if and only if $\det M_0(h_+, h_-) = 0$, which implies (6.5).

(iii) This follows readily from the expressions for a/c and b/c with the help of L'Hospital's rule.

(iv) In the case $\kappa_* = 0$, we needed to find nonzero solutions of (6.7) in order to derive a zero eigenvalue of $\mathcal{A}$. Each of the solutions to the linear systems (6.7) corresponds one eigenfunction to the eigenvalue zero. Assume that the multiplicity of an eigenvalue zero is larger than one. This implies that the matrix $M_0(h_+, h_-)$ has rank 1 (less is not possible). This implies

$$1 + h_+ l = 1 + h_- l = 0\,.$$

Hence

$$h_+ = h_- = -\frac{1}{l}\,.$$

But then the first and third columns are not linear dependent. This is a contradiction and shows the assertion for $\kappa_* = 0$. A similar argument works in the case $\kappa_* \neq 0$. ∎

We denote by N_U and N_N the number of unstable and zero eigenvalues of $\mathcal{A}$ (counting the multiplicity). Then we obtain the following theorem.

THEOREM 6.5. *Case* A: *If* $\mathcal{D}(h_-, h_+, \kappa_*) > 0$ *and if* $h_- > -b/c$, *then*

$$N_U = N_N = 0\,.$$

Case B: *If* $\mathcal{D}(h_-, h_+, \kappa_*) = 0$ *and if* $h_- > -b/c$, *then*

$$N_U = 0\,,\ N_N = 1\,.$$

Case C: *If* $\mathcal{D}(h_-, h_+, \kappa_*) < 0$, *then*

$$N_U = 1\,,\ N_N = 0\,.$$

Case D: *If* $\mathcal{D}(h_-, h_+, \kappa_*) = 0$ *and if* $h_- < -b/c$, *then*

$$N_U = 1\,,\ N_N = 1\,.$$

Case E: *If* $\mathcal{D}(h_-, h_+, \kappa_*) > 0$ *and if* $h_- < -b/c$, *then*

$$N_U = 2\,,\ N_N = 0\,.$$

Remark 6.6. (a) In Cases A, B, D, and E the condition $h_- > -b/c$ ($h_- < -b/c$, respectively) can be replaced by $h_+ > -b/c$ ($h_+ < -b/c$, respectively).

(b) Theorem 6.5 says that we have stability above the upper arc of the hyperbola (see Figures 1–5). Underneath it we have instability where the number of instable modes is one when we are above the lower arc of the hyperbola and two when we are underneath of it.

Proof of Theorem 6.5. The proof is a simple consequence of Lemmas 5.6, 6.1, and 6.3. For large h_+ and h_- we have stability. If we decrease h_+ or h_-, the stability behavior changes only on the curves defined by $\mathcal{D}(h_-, h_+, \kappa_*) = 0$. By virtue of (iv) in Lemma 6.3, only one eigenvalue can pass through zero when crossing the curves $\mathcal{D}(h_-, h_+, \kappa_*) = 0$. The monotonicity of the eigenvalues with respect to h_+ and h_- implies that the number of unstable modes can increase only if we further decrease h_+ or h_-. This proves the theorem. ∎

Let us discuss the signs of a, b, and c, which depend on $\kappa_* l$. It is easy to see

$$\begin{cases} a < 0 & \text{for} \quad \kappa_* l < \pi/2, \\ a = 0 & \text{for} \quad \kappa_* l = \pi/2, \\ a > 0 & \text{for} \quad \kappa_* l > \pi/2. \end{cases}$$

To derive the signs of b, we rewrite b as

$$\begin{aligned} b &= \frac{1}{2}\{2\kappa_* l \cos(2\kappa_* l) - \sin(2\kappa_* l)\} \\ &= \frac{1}{2}\cos(2\kappa_* l)\{2\kappa_* l - \tan(2\kappa_* l)\} \quad \text{if} \quad 2\kappa_* l \neq \pi/2, 3\pi/2. \end{aligned}$$

It follows from the relations between $2\kappa_* l$ and $\tan(2\kappa_* l)$ in $0 < 2\kappa_* l < 2\pi$ that

$$\begin{cases} b < 0 & \text{for} \quad \kappa_* l < \theta_0, \\ b = 0 & \text{for} \quad \kappa_* l = \theta_0, \\ b > 0 & \text{for} \quad \kappa_* l > \theta_0 \end{cases}$$

for some $\theta_0 \in (\pi/2, \pi)$. Finally, we investigate the sign of c. If $\kappa_* l \geq \pi/2$, we can easily derive $c < 0$. If $\kappa_* l < \pi/2$, we rewrite c as

$$c = \frac{2}{\kappa_*}\sin(\kappa_* l)\cos(\kappa_* l)\{\kappa_* l - \tan(\kappa_* l)\}.$$

Then $\kappa_* l < \pi/2$ implies that $\sin(\kappa_* l) > 0$, $\cos(\kappa_* l) > 0$, and $\kappa_* l - \tan(\kappa_* l) < 0$, so that $c < 0$. Thus we see $c < 0$ in all cases. Consequently the behavior illustrated in Figures 1–5 follows.

7. Examples. Finally we want to discuss how the linearized stability of equilibria depends on the parameters l, κ_*, h_+, and h_-. In the following the expressions "stable" and "unstable" are to be understood in the linearized sense.

If κ_* is zero and h_+ and h_- are negative, then the stability depends crucially on the length of Γ_*. For fixed h_+ and h_- equilibria with a small length are stable and equilibria with a large length are unstable. They are separated by a case which is neutral in the sense that the linearized evolution operator has, besides negative eigenvalues, one zero eigenvalue. This is, for example, the case when Ω is a ball and Γ_* is a segment intersection $\partial\Omega$ perpendicular (see Figure 6). In this case a nonlinear analysis has to decide on the stability.

If κ_* is nonzero, then the linearized stability behavior depends on the curvature of the outer boundary, roughly speaking, in the following sense. Cases with large positive outer curvatures h_+ and h_- are stable, and cases with large negative outer curvatures are unstable. In Figure 7 we demonstrate this stability behavior for a case where we fix κ_* and l. An equilibrium Γ_* is stable for $h > 0$ and unstable for $h < 0$. The case $h = 0$ is neutral and again a nonlinear analysis has to decide on stability. An interesting special case is when the outer boundary has constant curvature. This case is illustrated in Figure 8 and is always neutral. Indeed, let h be a constant curvature of the outer boundary, which implies $h_+ = h_- = h$. For the case $h = 0$, see the above explanation. If $h \neq 0$, then h is represented as

$$h = -\frac{\kappa_*}{\tan(\kappa_* l)}.$$

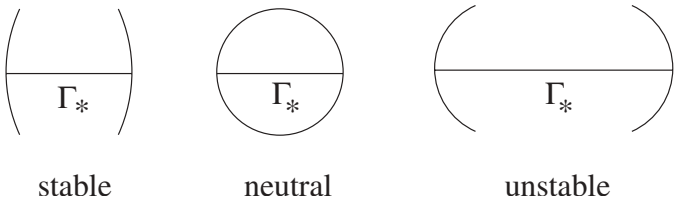

FIG. 6. *Three equilibria with $h_+ = h_-$ and $\kappa_* = 0$. (The stability depends on the length of Γ_*.)*

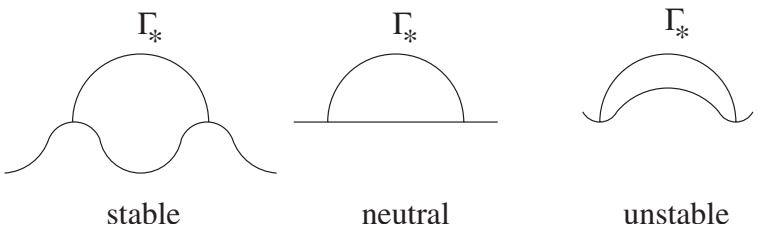

FIG. 7. *Three cases with κ_* and l fixed.*

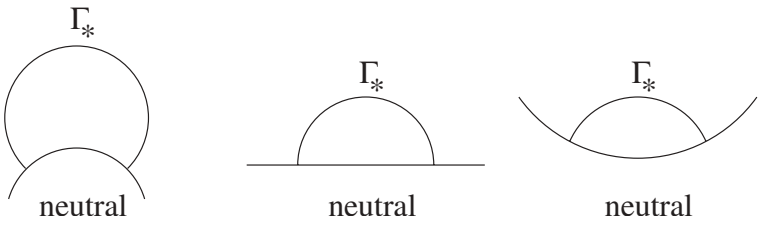

FIG. 8. *Three cases with the same κ_* and with constant curvature of $\partial\Omega$.*

By the definition of a, b, c, we derive

$$\frac{a}{c} = -\frac{\kappa_*^3 l}{\kappa_* l - \tan(\kappa_* l)}, \quad \frac{b}{c} = -\frac{h}{2}\left\{1 - \frac{\kappa_* l \tan^2(\kappa_* l)}{\kappa_* l - \tan(\kappa_* l)}\right\}.$$

This implies that

$$D(h, h, \kappa_*) = \frac{a}{c} + \frac{b}{c} \cdot 2h + h^2 = 0.$$

In addition, $h > -b/c$ for $0 < \kappa_* l < \pi/2$ follows from

$$1 - \frac{\kappa_* l \tan^2(\kappa_* l)}{\kappa_* l - \tan(\kappa_* l)} > 2 \quad \text{for} \quad 0 < \kappa_* l < \pi/2,$$

and we also find $h > 0$ for $\pi/2 < \kappa_* l < \pi$. This means that this case is included in the line $D = 0$ on the right-hand side of Figure 1 and Figures 3–5, so that this case is neutral.

Choosing, for example, $h_+ = h_- = 0$, we observe that $\kappa_* l$, is an important quantity (see Figure 9). As long as $\kappa_* l < \pi/2$ (i.e., Γ_* is less than a half circle) we

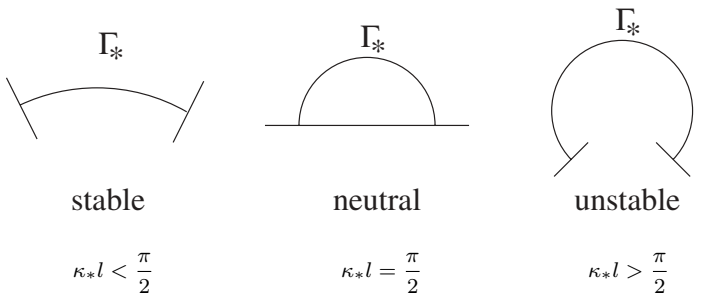

FIG. 9. *Three cases with* $h_+ = h_- = 0$.

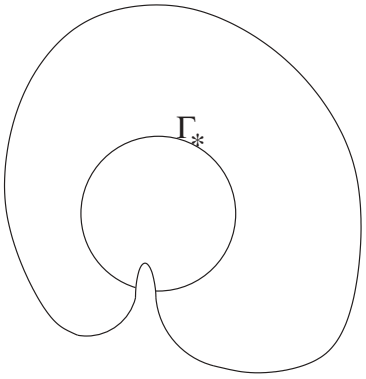

FIG. 10. *Instability for the case* $h_+, h_- > 0$.

have stability, the case $\kappa_* l = \pi/2$ is neutral (i.e., Γ_* is a half circle), and the case $\kappa_* l > \pi/2$ is unstable (i.e., Γ_* is more than a half circle).

Finally, we remark that instability for h_+, h_- positive and large is also possible. In this case $\kappa_* l$ has to be close to π, i.e., Γ_* has to be close to a full circle (see Figure 10).

Acknowledgments. The second author would like to express his special gratitude to Professor J. Escher for helpful discussions and warm hospitality during his stay at the Universität Hannover. Also the third author is grateful to the Universität Regensburg for the kind hospitality during his stay.

REFERENCES

[1] N. D. ALIKAKOS, P. W. BATES, X. CHEN, AND G. FUSCO, *Mullins-Sekerka motion of small droplets on a fixed boundary*, J. Geom. Anal., 10 (2000), pp. 575–596.

[2] J. W. BARRETT, J. F. BLOWEY, AND H. GARCKE, *On fully practical finite element approximations of degenerate Cahn-Hilliard systems*, Math. Model. Numer. Anal., 35 (2001), pp. 713–748.

[3] P. W. BATES AND P. C. FIFE, *Spectral comparison principles for the Cahn–Hilliard and phase-field equations, and time scales for coarsening*, Phys. D, 43 (1990), pp. 335–348.

[4] J. W. CAHN, C. M. ELLIOTT, AND A. NOVICK-COHEN, *The Cahn-Hilliard equation with a concentration dependent mobility: Motion by minus the Laplacian of the mean curvature*, European J. Appl. Math., 7 (1996), pp. 287–301.

[5] J. W. CAHN AND J. E. TAYLOR, *Surface motion by surface diffusion*, Acta Metallurgica, 42 (1994), pp. 1045–1063.

[6] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. I, Interscience, New York, 1953.
[7] F. Davi and M. Gurtin, *On the motion of a phase interface by surface diffusion*, Z. Angew. Math. Phys., 41 (1990), pp. 782–811.
[8] S.-I. Ei, M.-H. Sato, and E. Yanagida, *Stability of stationary interfaces with contact angle in a generalized mean curvature flow*, Amer. J. Math., 118 (1996), pp. 653–687.
[9] S.-I. Ei and E. Yanagida, *Stability of stationary interfaces in a generalized mean curvature flow*, J. Fac. Sci. Univ. Tokyo Sect. IA Math., 40 (1993), pp. 651–661.
[10] C. M. Elliott and H. Garcke, *Existence results for diffusive surface motion laws*, Adv. Math. Sci. Appl., 7 (1997), pp. 467–490.
[11] J. Escher, U. F. Mayer, and G. Simonett, *The surface diffusion flow for immersed hypersurfaces*, SIAM J. Math. Anal., 29 (1998), pp. 1419–1433.
[12] J. Escher, H. Garcke, and K. Ito, *Exponential stability for a mirror-symmetric three phase boundary motion by surface diffusion*, Math. Nachr., 257 (2003), pp. 3–15.
[13] P. C. Fife, *Models for phase separation and their mathematics*, in Nonlinear Partial Differential Equations and Applications, M. Mimura and T. Nishida, eds., KTK, Tokyo, 1993.
[14] H. Garcke and A. Novick-Cohen, *A singular limit for a system of degenerate Cahn-Hilliard equations*, Adv. Differential Equations, 5 (2000), pp. 401–434.
[15] M. E. Gurtin, *Thermodynamics of Evolving Phase Boundaries in the Plane*, Clarendon Press, Oxford, UK, 1993.
[16] R. Ikota and E. Yanagida, *A stability criterion for stationary curves to the curvature-driven motion with a triple junction*, Differential Integral Equations, 16 (2003), pp. 707–726.
[17] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.
[18] K. Ito and Y. Kohsaka, *Three phase boundary motion by surface diffusion: Stability of a mirror symmetric stationary solution*, Interfaces Free Bound., 3 (2001), pp. 45–80.
[19] K. Ito and Y. Kohsaka, *Three phase boundary motion by surface diffusion in triangular domain*, Adv. Math. Sci. Appl., 11 (2001), pp. 753–779.
[20] W. W. Mullins, *Theory of thermal grooving*, J. Appl. Phys., 28 (1957), pp. 333–339.
[21] J. Rubinstein, P. Sternberg, and J. B. Keller, *Fast reaction, slow diffusion, and curve shortening*, SIAM J. Appl. Math., 49 (1989), pp. 116–133.
[22] P. Sternberg and W. P. Ziemer, *Local minimizers of a three phase partition problem with triple junctions*, Proc. Roy. Soc. Edinburgh Sect. A, 124 (1994), pp. 1059–1073.
[23] J. E. Taylor and J. W. Cahn, *Linking anisotropic sharp and diffuse surface motion laws via gradient flows*, J. Statist. Phys., 77 (1994), pp. 183–197.
[24] E. Zeidler, *Applied Functional Analysis: Applications to Mathematical Physics*, Springer-Verlag, New York, 1995.

SIAM J. MATH. ANAL.
Vol. 36, No. 4, pp. 1057–1079

DENSITY ESTIMATES FOR A DEGENERATE/SINGULAR PHASE-TRANSITION MODEL*

ARSHAK PETROSYAN† AND ENRICO VALDINOCI‡

Abstract. We consider a Ginzburg–Landau type phase-transition model driven by a p-Laplacian type equation. We prove density estimates for absolute minimizers and we deduce the uniform convergence of level sets and the existence of plane-like minimizers in periodic media.

Key words. density estimates, p-Laplacian equation, phase-transition models, Ginzburg–Landau models, uniform convergence of level sets, plane-like minimizers

AMS subject classifications. 35J70, 35B45

DOI. 10.1137/S0036141003437678

1. Introduction.

For a bounded domain Ω in $\mathbb{R}^n$ and $\varepsilon > 0$ consider an energy functional

$$\mathcal{J}_\varepsilon(u;\Omega) = \int_\Omega [A(x,\varepsilon\nabla u) + F(x,u)]\,dx, \tag{1.1}$$

where $A(x,\eta) \simeq |\eta|^p$, $1 < p < \infty$, and $F(x,u) \simeq |1-u^2|^\alpha$, $0 < \alpha \le p$ (see below for precise assumptions). Functionals of this type appear in the context of minimal surfaces, and it has been shown by Γ-convergence methods that sequences of minimizers converge in L^1_{loc} to suitable step functions satisfying a minimal interface property, as $\varepsilon \to 0+$ (see [MM77] for $p = 2$ and [Bou90] for the general case). Functionals of type (1.1) also have a physical relevance, since they appear in the study of the equilibrium of elastic rods under tension (see [Ant73]), in the context of fluid jets (see [AC81] and [ACF84]), and in the van der Waals–Cahn–Hilliard and Ginzburg–Landau theories of phase transition (see, for instance, [Row79]). In the phase-transition setting, the term $A(x,\varepsilon\nabla u)$ in the energy functional (1.1) can be seen as an interfacial energy contribution to the total energy, which penalizes the formation of interfaces (see [Gur85] for details).

The main purpose of this paper is to obtain Caffarelli–Córdoba [CC95] type density estimates for the absolute minimizers of the normalized functional

$$\mathcal{J}(u;\Omega) = \int_\Omega [A(x,\nabla u) + F(x,u)]\,dx. \tag{1.2}$$

Roughly speaking, such density estimates state that, if u is an absolute minimizer of $\mathcal{J}$, then the set $\{|u| < 1/2\}$ behaves in measure as an $(n-1)$-dimensional set, while $\{u > 1/2\}$ and $\{u < -1/2\}$ behave in measure as n-dimensional sets (a precise statement will be given in Theorem 1.1 below). As a consequence, we obtain the uniform convergence of the level sets of minimizers of $\mathcal{J}_\varepsilon$ to a surface of minimal

*Received by the editors November 14, 2003; accepted for publication (in revised form) May 28, 2004; published electronically January 27, 2005.
http://www.siam.org/journals/sima/36-4/43767.html
†Department of Mathematics, Purdue University, West Lafayette, IN 47907 (arshak@math.purdue.edu).
‡Dipartimento di Matematica, Università di Roma Tor Vergata, Roma, I-00133, Italy (valdinoci@mat.uniroma2.it). The research of this author was partially supported by MURST Variational Methods and Nonlinear Differential Equations.

"area" as $\varepsilon \to 0+$; see Theorem 7.1. Another application of the density estimates is the existence of plane-like minimizers of $\mathcal{J}$ in periodic media; see Theorem 7.3.

We now state in detail the assumptions required in this paper. We assume that $A : \Omega \times \mathbb{R}^n \ni (x, \eta) \longrightarrow \mathbb{R}$ is in $C^1(\Omega \times \mathbb{R}^n)$ and that

$$a(x, \eta) := D_\eta A(x, \eta)$$

is in $C(\Omega \times \mathbb{R}^n) \cap C^1(\Omega \times \mathbb{R}^n - \{0\})$. We require that

$$A(x, 0) = 0, \quad a(x, 0) = 0, \tag{1.3}$$

for every $x \in \Omega$ and that there exists $\Lambda > 0$ such that

$$\zeta \cdot D_\eta a(x, \eta)\, \zeta \geq \Lambda^{-1} |\zeta|^2\, |\eta|^{p-2} \text{ for any } \zeta \in \mathbb{R}^n, \tag{1.4}$$

$$|D_\eta a(x, \eta)| \leq \Lambda\, |\eta|^{p-2} \qquad \text{and} \tag{1.5}$$

$$|D_x a(x, \eta)| \leq \Lambda\, |\eta|^{p-1}\,, \tag{1.6}$$

$$\eta \cdot a(x, \eta) \geq \Lambda^{-1}\, |\eta|^p \tag{1.7}$$

for every $x \in \Omega$ and $\eta \in \mathbb{R}^n$.

Next, we assume that $F : \Omega \times \mathbb{R} \ni (x, u) \longrightarrow \mathbb{R}$ is a Carathéodory function, i.e., continuous in u for a.e. $x \in \Omega$ and measurable in x for every $u \in \mathbb{R}$, and satisfies

$$0 \leq F \leq M, \qquad F(x, \pm 1) = 0, \qquad \inf_{|u| \leq \theta} F(x, u) \geq \gamma(\theta) \tag{1.8}$$

for every $0 \leq \theta < 1$, where $\gamma(\theta)$ and M are positive constants. Here and below all structural inequalities on F are assumed to be uniform for a.e. $x \in \Omega$. Further, we assume that the partial derivative $F_u(x, u)$ exists for every $u \in (-1, 1)$ for a.e. $x \in \Omega$ and that

$$\sup_{|u| \leq \theta} |F_u(x, u)| \leq M(\theta) \tag{1.9}$$

for every $0 \leq \theta < 1$. We also assume the following growth condition near $u = \pm 1$: there exists $s_0 > 0$ and $d \leq p$ such that

$$F_u(x, -1 + s) \geq C s^{d-1}, \qquad F_u(x, 1 - s) \leq -C s^{d-1} \tag{1.10}$$

for every $s \in (0, s_0)$. Without loss of generality, we may and do assume $1 \leq d \leq p$. In the case $d = p$ we additionally require that

$$F_u \text{ is monotone increasing in } u \text{ for } u \in (-1, -1 + s_0) \text{ and } u \in (1 - s_0, 1). \tag{1.11}$$

Finally, if $1 < p \leq 2n/(n+2)$, we require F to be uniformly Lipschitz in $u \in (-1, 1)$. More precisely, we assume that

$$\sup_{|u| < 1} |F_u(x, u)| \leq M \tag{1.12}$$

for a certain constant M.

We will refer to the constants that appear in (1.3)–(1.12), including n and p, as the *structural constants*. Quantities depending only on structural constants will be called *universal constants*.

A "typical" example of the functional $\mathcal{J}$, which satisfies the assumptions above, is given by

$$\int \Big(a_{i,j}(x)\,\partial_i u\,\partial_j u\Big)^{p/2} + Q(x)\,|1-u^2|^\alpha,$$

where $a_{i,j}$ is a C^1 symmetric positive definite matrix, $0 < Q_{\min} \le Q(x) \le Q_{\max}$ and $0 < \alpha \le p$. (The case $\alpha = 0$, which corresponds to $F(x,u) = Q(x)\chi_{(-1,1)}(u)$, has been treated recently in [PV03].)

We say that $u \in W^{1,p}(\Omega)$ is an *absolute minimizer* for $\mathcal{J}$ in Ω if $\mathcal{J}(u;\Omega) \le \mathcal{J}(u+\phi;\Omega)$ for any $\phi \in W_0^{1,p}(\Omega)$. In this paper, we will be concerned only with absolute minimizers u that satisfy $|u| \le 1$. Conditions (1.3)–(1.12) that we impose on $\mathcal{J}$ make it possible to apply the regularity results of Giaquinta and Giusti [GG82]. In particular, by Theorem 3.1 there, we will have that u is locally uniformly Hölder continuous in Ω. Moreover, in the region $\{|u| < 1\}$, the standard variational arguments show that u satisfies the Euler–Lagrange equation

$$\operatorname{div} a(x,\nabla u) = F_u(x,u)$$

in the weak sense. Then u is also $C^{1,\alpha}$ regular in $\{|u| < 1\}$ for some $0 < \alpha < 1$; see, e.g., [Tol84].

We will also denote by $\mathcal{L}^n$ the standard Lebesgue measure on $\mathbb{R}^n$.

The main result of this paper is as follows.

THEOREM 1.1. *For $1 < p < \infty$ assume that the hypotheses* (1.3)–(1.12) *hold. Fix $\theta \in (0,1)$ and let $|u| \le 1$ be an absolute minimizer for $\mathcal{J}$ in a bounded domain Ω, $x \in \{-\theta < u < \theta\}$ and $y \in \Omega$. Then, for every $\delta > 0$, there exist positive r_0, c, and C depending only on θ, on the structural constants, and on δ such that*

$$\mathcal{L}^n\big(B_r(x) \cap \{u > \theta\}\big) \ge c\,r^n \quad \text{and} \quad \mathcal{L}^n\big(B_r(x) \cap \{u < -\theta\}\big) \ge c\,r^n, \tag{1.13}$$

$$\mathcal{L}^n\big(B_r(x) \cap \{|u| < \theta\}\big) \ge c\,r^{n-1} \text{ and } \mathcal{L}^n\big(B_r(y) \cap \{|u| < \theta\}\big) \le C\,r^{n-1} \tag{1.14}$$

for any $r \ge r_0$, provided $B_{r+\delta}(x)$, $B_{r+\delta}(y) \subset\subset \Omega$.

The density estimates of this type have been obtained originally in [CC95] for $p = 2$ and $A(x,\nabla u) = |\nabla u|^2$ and generalized in [Val04] to $A(x,\nabla u) = a_{i,j}\partial_i u\,\partial_j u$. The case of a general $p \in (1,\infty)$ with $A(x,\nabla u)$ satisfying the hypotheses above and $F(x,u) = Q(x)\,\chi_{(-1,1)}(u)$ has been considered in [PV03] as a model for non-Newtonian power-law fluid jets. The case treated here can be seen as a degenerate/singular phase-transition model driven by a p-Laplacian type equation.

We explicitly point out here that there is a restriction in Theorem 1.1 on the decay rate of the "double-well" potential $F(x,u)$ near $u = \pm 1$. In particular, for $F(x,u) = |1-u^2|^\alpha$ for some $\alpha > 0$, we must have $\alpha \le p$ by (1.10). The density estimates as in Theorem 1.1 are not known for $\alpha > p$. Thus, the larger we take p, the wider is the class of admissible potentials $F(x,u)$ for which the density estimates are known. In that sense, the perturbations with $A(x,\eta) \simeq |\eta|^p$ behave better for larger values of p.

We also note that additional difficulties appear in the case $1 < p < 2$. We need to require uniform Lipschitz continuity of the double-well potential $F(x,u)$ in $u \in (-1,1)$ in order to obtain the desired density estimates. This excludes the potentials $F(x,u) = |1-u^2|^\alpha$ with $0 < \alpha < 1$. However, we show that *at least* for the range

of the exponents $2n/(n+2) < p < 2$, one can drop this uniform Lipschitz continuity assumption; see section 6.2.

The paper is organized as follows. In section 2 we collect some short-proof lemmas that will be of use in what follows. A Caccioppoli-type inequality is stated and proved in section 3. The proof of Theorem 1.1 is dealt with in section 4, and it makes use of an auxiliary result, namely, Lemma 4.1 below, which is interesting in itself and which roughly says that as soon as the density of sublevels of minimizers is positive in some ball, it must grow as r^n in balls of bigger radius r. We devote sections 5 and 6 to the proof of such an auxiliary result, considering the cases $p \geq 2$ and $1 < p \leq 2$ separately. In section 7 we point out some consequences that can be derived from Theorem 1.1, such as the uniform convergence of level sets of absolute minima to minimal interfaces and the existence of plane-like minimizers in periodic media.

2. Technical and elementary lemmas. We start this section with a recursive lemma.

LEMMA 2.1. *Let $v_k \geq 0$ and $a_k \geq 0$ be two nondecreasing sequences such that $v_1 + a_1 \geq c_0$,*

$$v_k^{(n-1)/n} \leq C_0 \left(v_{k+1} + a_{k+1} - v_k - a_k - c_1 a_k\right)^{1-\alpha} k^{\alpha(n-1)}$$

for any $k \in \mathbb{N}$ and some positive constants c_0, c_1, C_0, and $0 \leq \alpha < 1/n$. Then there exists $\gamma = \gamma(c_0, c_1, C_0, \alpha) > 0$ such that

$$v_k + a_k \geq \gamma\, k^n$$

for any $k \in \mathbb{N}$.

Proof. We start with an observation that it is enough to prove the estimate for $k \geq k_0$, since

$$v_k + a_k \geq v_1 + a_1 \geq c_0 \geq (c_0/k_0^n)k^n \quad \text{for } k \leq k_0.$$

The proof is by induction. Assume that $v_k + a_k \geq \gamma\, k^n$. Then either $v_k \geq (\gamma/2)k^n$ or $a_k \geq (\gamma/2)k^n$.

1. Assume first $v_k \geq (\gamma/2)k^n$. Then, using the recurrence relationship, we have

$$C_0 \left(v_{k+1} + a_{k+1} - v_k - a_k - c_1 a_k\right)^{1-\alpha} \geq (\gamma/2)^{(n-1)/n}\, k^{(1-\alpha)(n-1)}$$

and consequently

$$v_{k+1} + a_{k+1} \geq \gamma\, k^n + C\gamma^{\frac{1}{1-\alpha}\cdot\frac{n-1}{n}} k^{n-1}.$$

By our assumption, $\alpha < 1/n$, which implies that $\frac{1}{1-\alpha} \cdot \frac{n-1}{n} < 1$. Hence, if γ is sufficiently small,

$$v_{k+1} + a_{k+1} \geq \gamma(k^n + C_* k^{n-1})$$

for C_* as large as we wish. On the other hand, if we choose $C_* \geq 2^n$,

$$k^n + C_* k^{n-1} \geq (k+1)^n$$

and we obtain

$$v_{k+1} + a_{k+1} \geq \gamma(k+1)^n.$$

2. Assume now that $a_k \geq (\gamma/2)k^n$. Then

$$v_{k+1} + a_{k+1} \geq v_k + a_k + c_1 a_k \geq \gamma(k^n + (ck)k^{n-1}) \geq \gamma(k+1)^n$$

for sufficiently large k.

The proof is complete. □

The next lemma is similar in spirit. Its proof can be found on page 10 in [CC95] and is omitted here.

LEMMA 2.2. *Let $a_k \geq 0$ be a sequence such that $a_1 \geq c_0$, $a_k \leq C_0 L^n k^{n-1}$,*

$$\left(\sum_{1\leq j\leq k} a_j\right)^{(n-1)/n} \leq C_0 \left(a_{k+1} + \sum_{1\leq j\leq k} e^{-L(k+1-j)} a_j\right)$$

for any $k \in \mathbb{N}$ and some positive constants L, c_0, and C_0. Then, if $L = L(c_0, C_0)$ is suitably large, there exists $\gamma = \gamma(c_0, C_0) > 0$ such that

$$a_k \geq \gamma\, k^{n-1}$$

for any $k \in \mathbb{N}$. □

The next several lemmas are direct consequences of the structural hypotheses on $A(x,\eta)$ and $F(x,u)$.

LEMMA 2.3. *There exists a universal constant $\gamma > 0$ such that*

$$(a(x,\xi') - a(x,\xi)) \cdot (\xi' - \xi) \geq \gamma \cdot \begin{cases} (|\xi'| + |\xi|)^{p-2}|\xi' - \xi|^2 & \text{if } 1 < p \leq 2, \\ |\xi' - \xi|^p & \text{if } p \geq 2 \end{cases}$$

for every $\xi, \xi' \in \mathbb{R}^n$ and $x \in \Omega$.

Proof. For the reader's convenience we include a standard proof of this lemma. Set

$$\xi^s = s\,\xi' + (1-s)\,\xi, \quad 0 \leq s \leq 1. \tag{2.1}$$

Then $\xi^0 = \xi$ and $\xi^1 = \xi'$ and we have

$$a(x,\xi') - a(x,\xi) = \int_0^1 D_\eta a(x,\xi^s)(\xi' - \xi)ds.$$

By the hypothesis (1.4) we obtain

$$(a(x,\xi') - a(x,\xi)) \cdot (\xi' - \xi) \geq \Lambda^{-1}|\xi' - \xi|^2 \int_0^1 |\xi^s|^{p-2} ds.$$

Without loss of generality we may assume that $|\xi'| \leq |\xi|$. Then

$$(1/4)|\xi' - \xi| \leq |\xi^s| \leq |\xi'| + |\xi| \quad \text{for } 0 \leq s \leq 1/4.$$

Using the left-hand inequality for $p \geq 2$ and the right-hand inequality for $1 < p \leq 2$, we conclude the proof of the lemma. □

LEMMA 2.4. *For any $p \geq 2$ there exists a universal constant $c > 0$ such that*

$$c\,|\xi' - \xi|^p \leq A(x,\xi') - A(x,\xi) - a(x,\xi)\cdot(\xi' - \xi) \tag{2.2}$$

for every $\xi, \xi' \in \mathbb{R}^n$ and $x \in \Omega$.

Proof. Let ξ^s be as in (2.1). Then

$$\begin{aligned} A(x,\xi') - A(x,\xi) &= \int_0^1 a(x,\xi^s)\cdot(\xi' - \xi)ds \\ &= \int_0^1 (a(x,\xi^s) - a(x,\xi))\cdot(\xi' - \xi)ds + a(x,\xi)\cdot(\xi' - \xi) \\ &= \int_0^1 (a(x,\xi^s) - a(x,\xi))\cdot(\xi^s - \xi)\frac{ds}{s} + a(x,\xi)\cdot(\xi' - \xi). \end{aligned}$$

From Lemma 2.3 for $p \geq 2$ we have that

$$(a(x,\xi^s) - a(x,\xi))\cdot(\xi^s - \xi) \geq \gamma\,|\xi^s - \xi|^p.$$

Hence

$$\begin{aligned} A(x,\xi') - A(x,\xi) &\geq \gamma \int_0^1 |\xi^s - \xi|^p \frac{ds}{s} + a(x,\xi)\cdot(\xi' - \xi) \\ &= \gamma\,|\xi' - \xi|^p \int_0^1 s^{p-1}ds + a(x,\xi)\cdot(\xi' - \xi) \\ &= c\,|\xi' - \xi|^p + a(x,\xi)\cdot(\xi' - \xi). \qquad \square \end{aligned}$$

The analogue of the preceding Lemma 2.4 for $1 < p \leq 2$ is as follows.

LEMMA 2.5. *For any $1 < p \leq 2$ and $M \geq 0$ there exists a universal constant $c > 0$ such that*

$$c\,M^{p-2}\,|\xi' - \xi|^2 \leq A(x,\xi') - A(x,\xi) - a(x,\xi)\cdot(\xi' - \xi) \tag{2.3}$$

for every $\xi, \xi' \in \mathbb{R}^n$ with $|\xi| + |\xi'| \leq M$ and $x \in \Omega$.

Proof. The proof repeats the one for Lemma 2.4, except that we have to use the counterpart of Lemma 2.3 for $1 \leq p \leq 2$:

$$(a(x,\xi^s) - a(x,\xi))\cdot(\xi^s - \xi) \geq \gamma\,(|\xi^s| + |\xi|)^{p-2}|\xi^s - \xi|^2.$$

Then, also using $|\xi^s| + |\xi| \leq 2(|\xi'| + |\xi|)$, we will obtain

$$c\,(|\xi| + |\xi'|)^{p-2}|\xi' - \xi|^2 \leq A(x,\xi') - A(x,\xi) - a(x,\xi)\cdot(\xi' - \xi),$$

which implies (2.3) if $|\xi| + |\xi'| \leq M$. $\square$

The following result is elementary, and we omit the proof.

LEMMA 2.6. *Let $d \geq 1$. There exists $c_d > 0$ such that*

$$(u+1)^d - (u'+1)^d \geq c_d\,(u - u')^d$$

for any $u \geq u' \geq -1$.

Next, we deduce an estimate on the double-well potential.

LEMMA 2.7. *There exists $c > 0$ such that, for any $-1 \le u' \le u \le \theta$,*

$$F(x,u) - F(x,u') \ge c\,(u-u')^d\,,$$

provided $1+\theta > 0$ is small enough.

We omit the proof of Lemma 2.7, which easily follows from (1.10) and Lemma 2.6. The proof of the next two lemmas is also elementary.

LEMMA 2.8. *Let us assume (1.11). Then, there exists $c > 0$ so that, for any $-1 \le u' \le u \le \theta$, $F(x,u) - F(x,u') \ge c\,(u'+1)^{d-1}\,(u-u')$, provided $1+\theta > 0$ is small enough.*

LEMMA 2.9. *Let us assume that F is uniformly Lipschitz continuous in u. Then, there exists $c > 0$ so that, for any $u \in [-1,1]$, $F(x,u) \le c(1+u)$.*

We now construct a barrier that will be of use during the proof of the main result.

LEMMA 2.10. *Fix $T \ge 1$, $\Theta \in (0,1]$, and $k \in \mathbb{N}$. Then, there exists a function $h \in C^2(B_{(k+1)T})$ so that $-1 \le h \le 1$, $h = 1$ on $\partial B_{(k+1)T}$,*

$$(h+1) + |\nabla h| + |D^2 h| \le C(h+1) \le Ce^{-\Theta T(k+1-j)} \tag{2.4}$$

in $B_{jT} - B_{(j-1)T}$ for $j = 1,\dots,k$, and

$$|\nabla h| + |D^2 h| \le C\Theta(h+1) \tag{2.5}$$

in $B_{(k+1)T}$.

Proof. Define the following functions $\Phi : [0,1] \longrightarrow \mathbb{R}$, $\Psi : [1,(k+1)T] \longrightarrow \mathbb{R}$:

$$\Phi(t) = 2e^{\Theta[\frac{3}{8}t^6 - \frac{10}{8}t^4 + \frac{15}{8}t^2 - (k+1)T]} - 1 \quad \text{and}$$
$$\Psi(t) = 2e^{\Theta[t-(k+1)T]} - 1\,.$$

By explicit computations,

$$\Phi(1) = \Psi(1)\,, \quad \Phi'(1) = \Psi'(1), \quad \text{and} \quad \Phi''(1) = \Psi''(1)\,.$$

Thus, the function $\bar h$ agreeing with Φ in $[0,1]$ and with Ψ in $[1,(k+1)T]$ belongs to $C^2([0,(k+1)T])$. Define $h(x) = \bar h(|x|)$. Notice that $h \in C^2(B_{(k+1)T})$, since $\bar h'(0) = \Phi'(0) = 0$. Furthermore,

$$|\Phi'(t)| \le C\Theta t(\Phi+1)\,, \qquad |\Phi''(t)| \le C\Theta(\Phi+1) \tag{2.6}$$

in $[0,1]$ and

$$|\Psi'(t)| + |\Psi''(t)| \le C\Theta(\Psi+1) \tag{2.7}$$

in $[1,(k+1)T]$. Moreover,

$$(h+1) + |\nabla h| + |D^2 h| \le (\bar h + 1) + \left(1 + \frac{2}{|x|}\right)|\bar h'| + |\bar h''|\,. \tag{2.8}$$

By means of (2.6), we bound the right-hand side of (2.8) in B_1 by

$$C(\Phi+1) \le Ce^{\Theta(C-(k+1)T)} \le Ce^{-\Theta Tk}\,.$$

Similarly, using (2.7), we bound (2.8) by

$$C(\Psi+1) \le Ce^{-\Theta[(k+1)T-j]} \le Ce^{-\Theta T(k+1-j)}$$

in $B_{jT} - B_{(j-1)T}$ for $j = 2,\dots,k$. This proves (2.4). In a similar way, one can prove (2.5). ∎

3. A Caccioppoli-type inequality. We now state and prove a weaker version of the Caccioppoli inequality.

LEMMA 3.1. *Fix $\delta > 0$. Let $|u| \le 1$ be an absolute minimizer for $\mathcal{J}$ in a domain Ω. Then, there exists $C > 0$, depending only on δ and on the structural constants, such that*

$$\int_{B_r(x_0)} |\nabla u|^p \le C\,(r+\delta)^n$$

for any $r > 0$ and any $x_0 \in \Omega$, provided $B_{r+\delta}(x_0) \subset \Omega$.

Proof. We start with a claim that

$$\int_\Omega a(x,\nabla u)\cdot\nabla\phi + \int_{\Omega\cap\{|u|\neq 1\}} F_u(x,u)\,\phi\,dx \ge 0 \tag{3.1}$$

for any nonnegative $\phi \in C_0^\infty(\Omega\cap\{u > -1\})$ and

$$\int_\Omega a(x,\nabla u)\cdot\nabla\psi + \int_{\Omega\cap\{|u|\neq 1\}} F_u(x,u)\,\psi\,dx \le 0 \tag{3.2}$$

for any nonnegative $\psi \in C_0^\infty(\Omega\cap\{u < 1\})$. Let us show (3.2), the proof of (3.1) being analogous. For ψ as above and a small $\varepsilon > 0$, let

$$\psi_\varepsilon(x) = \psi(x)\chi_\varepsilon(u(x)),$$

where

$$\chi_\varepsilon(u) = \begin{cases} 0 & \text{if } u \le -1+\varepsilon,\\ (u+1)/\varepsilon - 1 & \text{if } -1+\varepsilon < u < -1+2\varepsilon,\\ 1 & \text{if } 1+u \ge 2\varepsilon. \end{cases}$$

Then $\psi_\varepsilon \in W^{1,p}(\Omega)$, $\operatorname{supp}\psi_\varepsilon \subset \Omega\cap\{|u|<1\}$, and therefore

$$\int_\Omega a(x,\nabla u)\cdot\nabla\psi_\varepsilon + F_u(x,u)\,\psi_\varepsilon = 0.$$

On the other hand,

$$\begin{aligned}&\int_\Omega a(x,\nabla u)\cdot\nabla\psi_\varepsilon\\ &= \int_\Omega [a(x,\nabla u)\cdot\nabla\psi]\chi_\varepsilon(u) + \frac{1}{\varepsilon}\int_{\Omega\cap\{\varepsilon<u+1<2\varepsilon\}} [a(x,\nabla u)\cdot\nabla u]\psi\\ &\ge \int_\Omega [a(x,\nabla u)\cdot\nabla\psi]\chi_\varepsilon(u) \to \int_\Omega a(x,\nabla u)\cdot\nabla\psi\end{aligned}$$

as $\varepsilon \to 0+$ and

$$\int_\Omega F_u(x,u)\psi_\varepsilon \to \int_{\Omega\cap\{u>-1\}} F_u(x,u)\psi.$$

The passage to the limit is legitimate, since

$$\int_\Omega |a(x,\nabla u)\cdot\nabla\psi| < \infty,$$

$\psi_\varepsilon \nearrow \psi\chi_{\{u>-1\}}$, and $F_u(x,u) \geq 0$ by (1.10) for u close to -1. Collecting the estimates above, we obtain (3.2).

Now fix $0 < \theta < 1$. If θ is sufficiently close to 1, the assumptions (1.9)–(1.10) and (3.1)–(3.2) above imply that

$$\int_\Omega a(x,\nabla u)\cdot\nabla\phi + K\,\phi \geq 0 \quad \text{and} \quad \int_\Omega a(x,\nabla u)\cdot\nabla\psi - K\,\psi \leq 0 \tag{3.3}$$

for any nonnegative $\phi \in C_0^\infty(\Omega\cap\{u > -\theta\})$ and $\psi \in C_0^\infty(\Omega\cap\{u<\theta\})$ with $K = M(\theta)$ as in (1.9). By standard density arguments, (3.3) also holds for nonnegative $\phi \in W_0^{1,p}(\Omega\cap\{u>-\theta\})$ and $\psi \in W_0^{1,p}(\Omega\cap\{u<\theta\})$.

Next, we observe that in light of Theorem 3.1 in [GG82], the distance between the level sets $\{u=-\theta\}$ and $\{u=\theta\}$ in $B_{r+\delta/2}(x_0)$ is bounded from below by some universal constant (depending only on θ, δ, and the structural constants). Therefore, by partition of unity, there exist two smooth functions η_- and η_+, supported in $B_{r+\delta/2}(x_0)$, so that $0 \leq \eta_-(x), \eta_+(x) \leq 1$ for any $x \in \Omega$, whose gradients are uniformly bounded by a universal constant and which satisfy

$$\begin{aligned} \eta_-(x)+\eta_+(x) \;&=\; 1 \quad \forall x \in B_r(x_0)\,,\\ \operatorname{supp}\eta_- &\subseteq \{-1 \leq u < \theta\}\,,\\ \operatorname{supp}\eta_+ &\subseteq \{-\theta < u \leq 1\}\,,\\ \eta_-(x)+\eta_+(x) &\leq 1 \quad \forall x\in\Omega\,. \end{aligned}$$

We set $\phi := (1-u)\,\eta_+^p$ and $\psi := (1+u)\,\eta_-^p$. By repeating the standard arguments in the proof of the Caccioppoli inequality (e.g., see Lemma 3.27 in [HKM93]), one infers that

$$\int_\Omega |\nabla u|^p\eta_-^p \;\leq\; C\,(r+\delta)^n \qquad \text{and} \qquad \int_\Omega |\nabla u|^p\eta_+^p \;\leq\; C\,(r+\delta)^n. \tag{3.4}$$

For the reader's convenience, we sketch the details of the proof of the second inequality in (3.4), the first being analogous. From (3.3),

$$0 \leq \int_\Omega -a(x,\nabla u)\cdot\nabla u\eta_+^p + p\,a(x,\nabla u)\,(1-u)\,\eta_+^{p-1}\nabla\eta_+ + K(1-u)\eta_+^p\,.$$

Therefore, introducing a parameter $\varepsilon \in (0,1)$, to be chosen suitably small in what follows, and using Young's inequality, we have

$$\begin{aligned} \int_\Omega |\nabla u|^p\eta_+^p &\leq C\left(\int_\Omega (|\nabla u|\eta_+)^{p-1}|\nabla\eta_+| + \eta_+^p\right)\\ &= C\left(\int_\Omega (\varepsilon|\nabla u|\eta_+)^{p-1}\frac{|\nabla\eta_+|}{\varepsilon^{p-1}} + \eta_+^p\right)\\ &\leq C\left(\int_\Omega (\varepsilon|\nabla u|\eta_+)^p + \frac{|\nabla\eta_+|^p}{\varepsilon^{p(p-1)}} + \eta_+^p\right)\\ &\leq C\varepsilon^p\int_\Omega (|\nabla u|\eta_+)^p + \frac{C}{\varepsilon^{p(p-1)}}\int_{B_{r+\delta}}(|\nabla\eta_+|^p + \eta_+^p)\\ &\leq C\varepsilon^p\int_\Omega |\nabla u|^p\eta_+^p + \frac{C}{\varepsilon^{p(p-1)}}(r+\delta)^n\,. \end{aligned}$$

Thus, the second inequality in (3.4) follows by choosing ε suitably small.

Using (3.4), we easily conclude the proof of the lemma:

$$\int_{B_r(x_0)} |\nabla u|^p \,=\, \int_{B_r(x_0)} |\nabla u|^p (\eta_- + \eta_+)^p$$
$$\le C \int_{B_r(x_0)} |\nabla u|^p \eta_-^p + C \int_{B_r(x_0)} |\nabla u|^p \eta_+^p \,\le\, C\,(r+\delta)^n\,. \qquad \square$$

4. Proof of Theorem 1.1. First, we point out that, since u is an absolute minimizer,

$$\mathcal{J}(u; B_r(y)) \le C\, r^{n-1} \tag{4.1}$$

for any $r \ge r_0$, for a suitable universal r_0, provided $B_{r+\delta}(y) \subset \Omega$. To prove (4.1), with no loss of generality assume $y = 0$ and proceed as follows. Let h be a smooth function such that $h\big|_{B_{r-1}} = -1$ and $h\big|_{\partial B_r} = 1$. Let $u^* = \min\{u, h\}$. Then,

$$\begin{aligned}\mathcal{J}(u; B_r) &\le \mathcal{J}(u^*; B_r)\\ &\le C \int_{B_r - B_{r-1}} (|\nabla u|^p + |\nabla h|^p) + r^{n-1}\\ &\le C \int_{B_r - B_{r-1}} |\nabla u|^p + r^{n-1}.\end{aligned}$$

Covering $B_r - B_{r-1}$ with balls of radius $\delta/2$, $\mathcal{B}_1, \dots, \mathcal{B}_N$, with $N \le C\, r^{n-1}$ and applying Lemma 3.1, we obtain

$$\int_{B_r - B_{r-1}} |\nabla u|^p \le \sum_{i=1}^{N} \int_{\mathcal{B}_i} |\nabla u|^p \le C\, r^{n-1}\,.$$

This completes the proof of (4.1).

We now focus our attention on the proof of (1.13). We will deal only with the first inequality in (1.13), the proof of the second one being analogous. To this end, we state the following result, the proof of which is deferred to sections 5 and 6.

LEMMA 4.1. *Let us assume the same hypotheses on A and F as in Theorem* 1.1. *Fix $\theta \in (-1, 1)$ and let u be an absolute minimizer for $\mathcal{J}$ in a domain Ω. Assume that there exist $\mu_1, \mu_2 > 0$ so that $B_{\mu_1}(x) \subset \Omega$ and $\mathcal{L}^n\big(B_{\mu_1}(x) \cap \{u > \theta\}\big) \ge \mu_2$. Then, for fixed $\delta > 0$, there exist positive r_0 and c depending only on θ, μ_1, μ_2, δ, and on the structural constants, such that $\mathcal{L}^n\big(B_r(x) \cap \{u > \theta\}\big) \ge c\, r^n$, for any $r \ge r_0$, provided $B_{r+\delta}(x) \subset\subset \Omega$.*

Analogously, if $\mathcal{L}^n\big(B_{\mu_1}(x) \cap \{u < \theta\}\big) \ge \mu_2$, then $\mathcal{L}^n\big(B_r(x) \cap \{u < \theta\}\big) \ge c\, r^n$ for any $r \ge r_0$, provided $B_{r+\delta}(x) \subset\subset \Omega$.

We now use the above result to prove (1.13). Let $\theta^\star = (1+\theta)/2 \in (\theta, 1)$. Since u is uniformly continuous (with a modulus of continuity depending only on the structural constants; see Theorem 3.1 in [GG82]) and $|u(x)| < \theta$, we have that $|u(x')| < \theta^\star$ for any $x' \in B_{\mu^\star}(x)$, for a suitable universal $\mu^\star > 0$. Hence, in view of Lemma 4.1, $\mathcal{L}^n\big(B_r(x) \cap \{u > -\theta^\star\}\big) \ge c\, r^n$ and $\mathcal{L}^n\big(B_r(x) \cap \{u < \theta^\star\}\big) \ge c\, r^n$. Therefore, by (1.8) and (4.1),

$$\begin{aligned}&c\, r^n - \mathcal{L}^n\big(B_r(x) \cap \{u > \theta\}\big)\\ &\le \mathcal{L}^n\big(B_r(x) \cap \{u > -\theta^\star\}\big) - \mathcal{L}^n\big(B_r(x) \cap \{u > \theta\}\big)\\ &\le \mathcal{L}^n\big(B_r(x) \cap \{-\theta^\star < u \le \theta\}\big)\\ &\le C \int_{\{-\theta^\star < u \le \theta\} \cap B_r(x)} F\\ &\le C\, \mathcal{J}(u; B_r(x)) \,\le\, C\, r^{n-1}\,.\end{aligned}$$

Hence, if r is suitably large, $\mathcal{L}^n\big(B_r(x)\cap\{u>\theta\}\big)\geq c\,r^n$, thus proving (1.13).

We now deal with the proof of (1.14). The second inequality follows from (4.1); hence we focus on the proof of the first one. Let

$$u^\star(x)=\begin{cases} u(x) & \text{if} \quad |u(x)|<\theta\,,\\ \theta & \text{if} \quad u(x)\geq\theta\,,\\ -\theta & \text{if} \quad u(x)\leq-\theta. \end{cases}$$

Using a standard notation in geometric measure theory, we denote by $\mathcal{P}(E;U)$ the perimeter of the Borel set E in an open set U. Then, using the coarea formula, the isoperimetric inequality, and (1.13), we have

$$\begin{aligned}
&\int_{B_r(x)}|\nabla u^\star|\\
&\geq\int_{-\theta}^{\theta}\mathcal{P}\Big(\{u^\star<s\};\,B_r(x)\Big)\,ds\\
&\geq c\int_{-\theta}^{\theta}\min\Big\{\mathcal{L}^n\left(B_r(x)\cap\{u^\star<s\}\right),\,\mathcal{L}^n\left(B_r(x)\cap\{u^\star\geq s\}\right)\Big\}^{\frac{n-1}{n}}\,ds\\
&=c\int_{-\theta}^{\theta}\min\Big\{\mathcal{L}^n\left(B_r(x)\cap\{u<s\}\right),\,\mathcal{L}^n\left(B_r(x)\cap\{u\geq s\}\right)\Big\}^{\frac{n-1}{n}}\,ds\\
&\geq c\int_{-\theta}^{\theta}\min\Big\{\mathcal{L}^n\left(B_r(x)\cap\{u<-\theta\}\right),\,\mathcal{L}^n\left(B_r(x)\cap\{u\geq\theta\}\right)\Big\}^{\frac{n-1}{n}}\,ds\\
&\geq c\,r^{n-1}\,.
\end{aligned}$$

Let us now fix a suitably large parameter $K>0$. In view of the above estimate, denoting by p' the conjugated exponent of p, using Young's inequality and (4.1), we deduce that

$$\begin{aligned}
c\,r^{n-1}&\leq\frac{1}{K^p}\int_{B_r(x)}|\nabla u|^p+K^{p'}\,\mathcal{L}^n\left(B_r(x)\cap\{|u|<\theta\}\right)\\
&\leq\frac{C}{K^p}\mathcal{J}(u;B_r(x))\,+\,K^{p'}\,\mathcal{L}^n\left(B_r(x)\cap\{|u|<\theta\}\right)\\
&\leq\frac{C}{K^p}\,r^{n-1}\,+\,K^{p'}\,\mathcal{L}^n\left(B_r(x)\cap\{|u|<\theta\}\right)\,.
\end{aligned}$$

Then, (1.14) follows by choosing K large enough here above. ∎

5. Proof of Lemma 4.1. The case $p\geq 2$. We will prove the first claim in Lemma 4.1, the second claim being analogous. We point out that it is enough to show the validity of the first claim of Lemma 4.1 for θ as close to -1 as we wish. Indeed, let us assume that the claim holds true for $\theta_\star$ and $-1<\theta_\star<\theta<1$. Then, if $\mathcal{L}^n\left(B_{\mu_1}(x)\cap\{u>\theta\}\right)\geq\mu_2$, then of course $\mathcal{L}^n\left(B_{\mu_1}(x)\cap\{u>\theta_\star\}\right)\geq\mu_2$, and so, since the claim holds for $\theta_\star$, $\mathcal{L}^n\left(B_r(x)\cap\{u>\theta_\star\}\right)\geq c\,r^n$. Thus, using the second part of (1.14) (which has already been proved via (4.1)),

$$\begin{aligned}
&\mathcal{L}^n\left(B_r(x)\cap\{u>\theta\}\right)\\
&\geq\mathcal{L}^n\left(B_r(x)\cap\{u>\theta_\star\}\right)\,-\,\mathcal{L}^n\left(B_r(x)\cap\{\theta_\star<u\leq\theta\}\right)\\
&\geq c\,r^n\,-\,C\,r^{n-1}\,\geq\,\tilde{c}\,r^n
\end{aligned}$$

if r is sufficiently big. This shows that we need only to prove the first claim of Lemma 4.1 for θ close to -1. Thus, we may assume that (1.10) is satisfied for $0<s\leq\theta+1$.

In this section we will assume $p \geq 2$. We will distinguish the cases $d < p$ and $d = p$, where d is the exponent that appears in (1.10).

5.1. The case $d < p$. For θ close to -1 let

$$\mathcal{V}_r = \mathcal{L}^n(\{u \geq \theta\} \cap B_r), \qquad \mathcal{A}_r = \int_{B_r} F(x,u)dx\,, \tag{5.1}$$

where B_r is short for $B_r(x)$. Then, we claim that

$$\mathcal{V}_r^{(n-1)/n} + \mathcal{A}_r \leq C_0\,(\mathcal{V}_{r+1} + \mathcal{A}_{r+1} - \mathcal{V}_r - \mathcal{A}_r)\,. \tag{5.2}$$

With no loss of generality, we can take $r_0 \geq \mu_1$ and $\delta \geq 2$. Thus, by assumption, $\mathcal{V}_{r_0} \geq \mu_2 > 0$. Therefore, by means of Lemma 2.1, the above inequality implies that

$$\mathcal{V}_r \geq c\, r^n$$

for $r \geq 1$.

We now prove (5.2). We use a barrier function $h \in C^2(B_{r+1})$ such that

$$h\big|_{\partial B_{r+1}} = 1, \qquad h\big|_{B_r} = -1.$$

Let $\varepsilon = 1 + \theta$ and define $u^* = \min(u, h)$ and $\beta = \min(u - u^*, \varepsilon)$. Using the Sobolev inequality applied to β^p and then Young's inequality, we have

$$\begin{aligned}
&\left(\int_{B_{r+1}} |\beta|^{pn/(n-1)}\right)^{(n-1)/n} \\
&\leq C \int_{B_{r+1}\cap\{u-u^*<\varepsilon\}} |\beta|^{p-1}|\nabla\beta| \\
&\leq C\,K^p \int_{B_{r+1}\cap\{u-u^*<\varepsilon\}} |\nabla(u-u^*)|^p \\
&\quad + \frac{C}{K^{p'}} \int_{B_{r+1}\cap\{u-u^*<\varepsilon\}} (u-u^*)^p\,.
\end{aligned} \tag{5.3}$$

Here, $K > 0$ is a free parameter that will be conveniently chosen in what follows. As customary, we also denoted the conjugated exponent of p by p'. Since $u^* = -1$ in B_r, $u - u^* \geq \varepsilon$ in $B_r \cap \{u \geq \theta\}$, the left-hand side of the inequality above is bounded from below by

$$c\,\mathcal{L}^n(\{u \geq \theta\} \cap B_r)^{(n-1)/n} = c\,\mathcal{V}_r^{(n-1)/n}.$$

Next, we apply (2.2) with $\xi = \nabla u^*$ and $\xi' = \nabla u$ to estimate $|\nabla(u - u^*)|^p$ in the right-hand side of (5.3). Thus, we obtain

$$\begin{aligned}
\mathcal{V}_r^{(n-1)/n} &\leq C\,K^p \int_{B_{r+1}} A(x,\nabla u) - A(x,\nabla u^*) \\
&\quad - C\,K^p \int_{B_{r+1}} a(x,\nabla u^*)\cdot\nabla(u-u^*) \\
&\quad + \frac{C}{K^{p'}} \int_{B_{r+1}\cap\{u-u^*<\varepsilon\}} (u-u^*)^p.
\end{aligned} \tag{5.4}$$

Since $\operatorname{supp}(u-u^*) \subset B_{r+1} \subset\subset \Omega$, the minimality of u implies that $\mathcal{J}(u;\Omega) \leq \mathcal{J}(u^*;\Omega)$ or, equivalently,

$$\int_{B_{r+1}} A(x,\nabla u) - A(x,\nabla u^*) \leq \int_{B_{r+1}} F(x,u^*) - F(x,u).$$

Using this, and integrating by parts the term $a(x,\nabla u^*)\cdot\nabla(u-u^*)$ in the right-hand side of (5.4), we obtain

$$\begin{aligned} \mathcal{V}_r^{(n-1)/n} &\leq C\,K^p \int_{B_{r+1}} F(x,u^*) - F(x,u) \\ &\quad + C\,K^p \int_{B_{r+1}} \operatorname{div} a(x,\nabla u^*)(u-u^*) \\ &\quad + \frac{C}{K^{p'}} \int_{B_{r+1}\cap\{u-u^*<\varepsilon\}} (u-u^*)^p. \end{aligned} \tag{5.5}$$

Notice also that, by definition of u^*,

$$\int_{B_{r+1}} \operatorname{div} a(x,\nabla u^*)(u-u^*) = \int_{B_{r+1}} \operatorname{div} a(x,\nabla h)(u-u^*).$$

Thus, to proceed, we recall that by (1.5) and (1.6) we have

$$\operatorname{div} a(x,\nabla h) \leq C|\nabla h|^{p-2}(|\nabla h| + |D^2 h|). \tag{5.6}$$

Now let h be a C^2 radial function, defined by

$$h(x) = -1 + 2(|x| - r)_+^\alpha,$$

with some fixed

$$\alpha > \max\left\{\frac{p}{p-d}, 2\right\}.$$

Then

$$|\nabla h| \leq C(h+1)^{(\alpha-1)/\alpha}, \qquad |D^2 h| \leq C(h+1)^{(\alpha-2)/\alpha}.$$

Hence,

$$\operatorname{div} a(x,\nabla h) \leq C(h+1)^{(p-2)(\alpha-1)/\alpha + (\alpha-2)/\alpha} \leq C(h+1)^{d-1} \tag{5.7}$$

and consequently

$$\begin{aligned} \mathcal{V}_r^{(n-1)/n} &\leq C\,K^p \int_{B_{r+1}} [F(x,u^*) - F(x,u)] \\ &\quad + C\,K^p \int_{B_{r+1}} (u^*+1)^{d-1}\,(u-u^*) \\ &\quad + \frac{C}{K^{p'}} \int_{B_{r+1}\cap\{u-u^*<\varepsilon\}} (u-u^*)^p. \end{aligned} \tag{5.8}$$

We now split the right-hand side of the above inequality into three parts, namely, the contribution in B_r, the one in $\{u<\theta\}\cap(B_{r+1}-B_r)$, and the one in $\{u\geq\theta\}\cap(B_{r+1}-B_r)$.

1. The contribution in B_r. Here the second integrand in the right-hand side of (5.8) vanishes, as well as the term $F(x,u^*)$ of the first integrand. Besides, the third integral is taken over the region, where $u-u^*<\varepsilon$. In B_r, the latter condition is equivalent to $u<\theta$, since $u^*=-1$. Furthermore, if K is sufficiently large, for $u<\theta$ we have

$$-C\,K^p F(x,u)+C\,K^{-p'}(u+1)^p\le -c\,F(x,u),$$

since by our assumption $F(x,u)\ge c(u+1)^d\ge c(u+1)^p$ for $-1\le u<\theta$. Hence, the contribution of the right-hand side of (5.8) in B_r is bounded from above by $-c\mathcal{A}_r$.

2. The contribution in $\{u<\theta\}\cap(B_{r+1}-B_r)$. Since $-1\le u^*\le u<\theta$, from Lemma 2.7 we have that

$$F(x,u^*)-F(x,u)\le -c\,(u-u^*)^d.$$

Since both $u^*+1\le u+1$ and $u-u^*\le u+1$, we also have

$$(u^*+1)^{d-1}(u-u^*)\le (u+1)^d\le C\,F(x,u).$$

Thus,

$$K^p[F(x,u^*)-F(x,u)+(u^*+1)^{d-1}(u-u^*)]+K^{-p'}(u-u^*)^p\le C\,F(x,u),$$

and the total contribution of the right-hand side of (5.8) in $\{u<\theta\}\cap(B_{r+1}-B_r)$ is bounded from above by $C(\mathcal{A}_{r+1}-\mathcal{A}_r)$.

3. Finally, the contribution in $\{u\ge\theta\}\cap(B_{r+1}-B_r)$ is easily estimated by

$$C\,\mathcal{L}^n(\{u\ge\theta\}\cap(B_{r+1}-B_r))=C(\mathcal{V}_{r+1}-\mathcal{V}_r),$$

since the terms inside the integrals are bounded.

Collecting the estimates from 1–3, we obtain (5.2), which completes the proof of Lemma 4.1 in the case $p\ge 2$, $d<p$. □

5.2. The case $d=p$. The proof is a refinement of the one in the case $d<p$. Here we use suitable positive parameters Θ and T: we will fix Θ small enough and then choose T suitably large (and in fact ΘT suitably large).

We define $\mathcal{V}_r$ as in (5.1) and set $a_k=\mathcal{V}_{kT}-\mathcal{V}_{(k-1)T}$. Then, we claim that

$$\left(\sum_{1\le j\le k}a_j\right)^{(n-1)/n}\le C\left(a_{k+1}+\sum_{1\le j\le k}e^{-L(k+1-j)}a_j\right),\tag{5.9}$$

with $L=\Theta T$ as large as we wish. With no loss of generality, we can take $r_0\ge\mu_1$. Thus, by assumption, $\mathcal{V}_{r_0}\ge\mu_2>0$. Therefore, by means of Lemma 2.2, the above inequality implies that

$$\mathcal{V}_r\ge c\,r^n$$

for $r\ge 1$.

We now prove (5.9). We use the barrier function $h=h_k\in C^2(B_{(k+1)T})$ introduced in Lemma 2.10. Then, $-1\le h\le 1$, $h=1$, on $\partial B_{(k+1)T}$,

$$(h+1)+|\nabla h|+|D^2h|\le C(h+1)\le Ce^{-\Theta T(k+1-j)}\tag{5.10}$$

in $B_{jT} - B_{(j-1)T}$, and

$$|\nabla h| + |D^2 h| \le C\Theta(h+1) \tag{5.11}$$

in $B_{(k+1)T}$.

Fix $\varepsilon \in (0, 1+\theta)$. Define $u^* = \min(u, h)$ and $\beta = \min(u - u^*, \varepsilon)$. Using the Sobolev inequality applied to β^p and then Young's inequality, we have

$$\begin{aligned}\left(\int_{B_{(k+1)T}} |\beta|^{pn/(n-1)}\right)^{(n-1)/n} \\ \le C\,K^p \int_{B_{(k+1)T}\cap\{u-u^*<\varepsilon\}} |\nabla(u-u^*)|^p \\ + \frac{C}{K^{p'}} \int_{B_{(k+1)T}\cap\{u-u^*<\varepsilon\}} (u-u^*)^p\end{aligned} \tag{5.12}$$

with the parameter $K > 0$ to be chosen later. From (5.10),

$$u - u^* \ge \theta + 1 - Ce^{-\Theta T} > \varepsilon \tag{5.13}$$

in $B_{kT} \cap \{u \ge \theta\}$, provided ΘT is conveniently large; hence the left-hand side of the inequality above is bounded from below by

$$c\,\mathcal{L}^n(\{u \ge \theta\} \cap B_{kT})^{(n-1)/n} = c\,\mathcal{V}_{kT}^{(n-1)/n}.$$

Now, combining (5.12) and (2.2), using the minimality of u, and integrating by parts the term $a(x, \nabla u^*)\cdot\nabla(u-u^*)$ (for more details, see the respective part in section 5.1), we obtain

$$\begin{aligned}\mathcal{V}_{kT}^{(n-1)/n} &\le C\,K^p \int_{B_{(k+1)T}} F(x,u^*) - F(x,u) \\ &\quad + C\,K^p \int_{B_{(k+1)T}} \operatorname{div} a(x,\nabla u^*)(u-u^*) \\ &\quad + \frac{C}{K^{p'}} \int_{B_{(k+1)T}\cap\{u-u^*<\varepsilon\}} (u-u^*)^p.\end{aligned}$$

Recalling (1.5) and (1.6), we have that

$$\operatorname{div} a(x,\nabla h) \le C|\nabla h|^{p-2}(|\nabla h| + |D^2 h|). \tag{5.14}$$

Hence,

$$\begin{aligned}\mathcal{V}_{kT}^{(n-1)/n} &\le C\,K^p \int_{B_{(k+1)T}} F(x,u^*) - F(x,u) \\ &\quad + C\,K^p \int_{B_{(k+1)T}} |\nabla u^*|^{p-2}\,(|\nabla u^*| + |D^2 u^*|)\,(u-u^*) \\ &\quad + \frac{C}{K^{p'}} \int_{B_{(k+1)T}\cap\{u-u^*<\varepsilon\}} (u-u^*)^p\,.\end{aligned}$$

Thanks to the definition of u^*, we may replace u^* with h in the second integral in the previous inequality in order to gather

$$\begin{aligned}\mathcal{V}_{kT}^{(n-1)/n} &\le C\,K^p \int_{B_{(k+1)T}} F(x,u^*) - F(x,u)\\ &\quad + C\,K^p \int_{B_{(k+1)T}} |\nabla h|^{p-2}\,(|\nabla h| + |D^2 h|)\,(u-u^*)\\ &\quad + \frac{C}{K^{p'}} \int_{B_{(k+1)T}\cap\{u-u^*<\varepsilon\}} (u-u^*)^p.\end{aligned} \tag{5.15}$$

We now split the right-hand side of the above inequality into three parts, namely, the contribution in $\{u<\theta\}$, the one in $\{u\ge\theta\}\cap(B_{(k+1)T}-B_{kT})$, and the one in $\{u\ge\theta\}\cap B_{kT}$.

1. The contribution in $\{u<\theta\}$ is estimated using (5.11) and Lemmas 2.7 and 2.8. We actually show that such contribution is negative. Indeed, using the above mentioned results and taking K suitably big (so as to kill the last term with the first one) and Θ suitably small (so as to kill the constant c in Lemma 2.8), the contribution in $\{u<\theta\}$ is bounded by

$$C\int_{B_{(k+1)T}\cap\{u<\theta\}} F(x,u^*) - F(x,u) + C\Theta(u^*+1)^{d-1}(u-u^*) \le 0\,.$$

2. The contribution in $\{u\ge\theta\}\cap(B_{(k+1)T}-B_{kT})$ of the right-hand side of (5.15) can be easily bounded by $C\,a_{k+1}$, since the terms inside the integrals are bounded.

3. We now estimate the contribution of the right-hand side of (5.15) in $\{u\ge\theta\}\cap B_{kT}$. First, notice that, by (5.13),

$$\int_{B_{kT}\cap\{u-u^*<\varepsilon\}\cap\{u\ge\theta\}} (u-u^*)^p = \int_{\emptyset}(u-u^*)^p = 0\,.$$

Also, from (1.9) and (1.11), it follows that F is uniformly Lipschitz continuous in u; thus, from Lemma 2.9, $F(x,h)\le c\,(1+h)$. Therefore, by (5.10), we bound the contribution in $\{u\ge\theta\}\cap B_{kT}$ by

$$\begin{aligned}&C\left(\sum_{j=1}^{k}\int_{(B_{jT}-B_{(j-1)T})\cap\{u\ge\theta\}} F(x,h) + |\nabla h| + |D^2 h|\right)\\ &\le C\left(\sum_{j=1}^{k} e^{-\Theta T(k+1-j)}\,\mathcal{L}^n((B_{jT}-B_{(j-1)T})\cap\{u\ge\theta\})\right).\end{aligned}$$

In light of 1–3, we bound the right-hand side of (5.15) by

$$C\left(a_{k+1} + \sum_{1\le j\le k} e^{-L(k+1-j)}a_j\right).$$

This proves (5.9) and completes the proof of Theorem 1.1 in the case $p\ge 2$.

6. Proof of Lemma 4.1. The case $1 < p < 2$.

6.1. The case of uniformly Lipschitz F. Under the assumption (1.12) of the uniform Lipschitz continuity of the double-well potential F, every absolute minimizer u of $\mathcal{J}$ with $|u| \le 1$ will satisfy an equation

$$\operatorname{div} a(x, \nabla u) = g(x), \tag{6.1}$$

weakly in Ω for some $g \in L^\infty(\Omega)$. Indeed, if M is as in (1.12) and $\psi \in C_0^\infty(\{u < 1\})$ is nonnegative, using that $\mathcal{J}(u + \varepsilon\psi; \Omega) \ge \mathcal{J}(u; \Omega)$, we will easily obtain

$$\int_\Omega a(x, \nabla u) \cdot \nabla\psi + M\,\psi \ge 0.$$

On the other hand, by (3.2), we also have

$$\int_\Omega a(x, \nabla u) \cdot \nabla\psi - M\,\psi \le 0.$$

Hence (6.1) is satisfied with $|g| \le M$ in $\{u < 1\}$. Similarly, we prove (6.1) in $\{u > -1\}$ and consequently in Ω.

Note that $g(x) = F_u(x, u)$ a.e. in $\{|u| < 1\}$ and $g(x) = 0$ in $\Omega \setminus \overline{\{|u| < 1\}}$, but we have no information on $g(x)$ on the "free boundary" $\partial\{|u| < 1\} \cap \Omega$, except that it is bounded. However, that is sufficient for our purposes.

The equation (6.1) implies that u is locally uniformly $C^{1,\alpha}$ regular in Ω; see [Tol84]. Then the proof of Lemma 4.1 in the case $1 < p \le 2$ is a slight variation of the one for $p \ge 2$. The main difference is that we use Lemma 2.5 instead of Lemma 2.4. Technically, we should separately consider the cases $d < p$ and $d = p$. However, since the changes from the case $p \ge 2$ are similar in both cases, we sketch only the proof for the more subtle case $d = p$.

We consider suitable positive parameters Θ, T, and K (playing the same role as in section 5.2) and we define h, u^*, and β as we did in section 5.2 above. In analogy with (5.12), using the Sobolev inequality applied to β^2 and then Young's inequality, we have

$$\begin{aligned} &\left(\int_{B_{(k+1)T}} |\beta|^{2n/(n-1)} \right)^{(n-1)/n} \\ &\le C\,K^2 \int_{B_{(k+1)T} \cap \{u - u^* < \varepsilon\}} |\nabla(u - u^*)|^2 \\ &\quad + \frac{C}{K^2} \int_{B_{(k+1)T} \cap \{u - u^* < \varepsilon\}} (u - u^*)^2 \,. \end{aligned} \tag{6.2}$$

Arguing as in (5.13), we get that the left-hand side of the inequality above is estimated from below by

$$c\,\mathcal{L}^n(\{u \ge \theta\} \cap B_{kT})^{(n-1)/n} = c\,\mathcal{V}_{kT}^{(n-1)/n}.$$

Notice that $|\nabla u|$ is uniformly bounded by means of Theorem 1 in [Tol84] (and, indeed, u is $C^{1,\alpha}$ with uniform estimates in the interior of Ω). Thus, using (2.3), the

minimality of u, and an integration by parts, we infer from (6.2) the following inequality:

$$\begin{aligned}\mathcal{V}_{kT}^{(n-1)/n} &\le C\,K^2 \int_{B_{(k+1)T}} F(x,u^*) - F(x,u)\\ &\quad + C\,K^2 \int_{B_{(k+1)T}} \operatorname{div} a(x,\nabla u^*)(u-u^*)\\ &\quad + \frac{C}{K^2}\int_{B_{(k+1)T}\cap\{u-u^*<\varepsilon\}} (u-u^*)^2.\end{aligned}$$

In light of (5.14), we deduce that

$$\begin{aligned}\mathcal{V}_{kT}^{(n-1)/n} &\le C\,K^2 \int_{B_{(k+1)T}} F(x,u^*) - F(x,u)\\ &\quad + C\,K^2 \int_{B_{(k+1)T}} |\nabla u^*|^{p-2}\,(|\nabla u^*| + |D^2 u^*|)\,(u-u^*)\\ &\quad + \frac{C}{K^2}\int_{B_{(k+1)T}\cap\{u-u^*<\varepsilon\}} (u-u^*)^2.\end{aligned}$$

By the definition of u^*, we may replace u^* with h in the second integral in the previous inequality, obtaining

$$\begin{aligned}\mathcal{V}_{kT}^{(n-1)/n} &\le C\,K^2 \int_{B_{(k+1)T}} F(x,u^*) - F(x,u)\\ &\quad + C\,K^2 \int_{B_{(k+1)T}} |\nabla h|^{p-2}\,(|\nabla h| + |D^2 h|)\,(u-u^*)\\ &\quad + \frac{C}{K^2}\int_{B_{(k+1)T}\cap\{u-u^*<\varepsilon\}} (u-u^*)^2.\end{aligned} \tag{6.3}$$

As done in section 5.2, one splits the right-hand side of the above inequality into three parts, namely, the contribution in $\{u<\theta\}$, the one in $\{u\ge\theta\}\cap(B_{(k+1)T}-B_{kT})$, and the one in $\{u\ge\theta\}\cap B_{kT}$. Such estimates follow the lines of section 5.2. Namely, the contribution in $\{u\le\theta\}$ is estimated by using (5.11) and Lemmas 2.7 and 2.8, obtaining, for big K and small $\Theta>0$, the bound

$$C\int_{B_{(k+1)T}\cap\{u<\theta\}} F(x,u^*) - F(x,u) + C\Theta^{p-1}(u^*+1)^{d-1}(u-u^*) \le 0\,,$$

which is negative. The contribution in $\{u\ge\theta\}\cap(B_{(k+1)T}-B_{kT})$ of the right-hand side of (5.15) can be easily bounded by $C\,a_{k+1}$. As above, the contribution in $\{u\ge\theta\}\cap B_{kT}$ is bounded by using Lemma 2.9 and (5.10), obtaining

$$C\left(\sum_{j=1}^{k} e^{-(p-1)\Theta T(k+1-j)}\,\mathcal{L}^n((B_{jT}-B_{(j-1)T})\cap\{u\ge\theta\})\right).$$

This proves (5.9) and hence completes the proof of Theorem 1.1 in the case $1<p\le 2$ for potentials F satisfying (1.12). □

6.2. The case $2n/(n+2) < p < 2$. We now show that the density estimate in Lemma 4.1 can be obtained at least for $2n/(n+2) < p < 2$ without the technical assumption (1.12) of uniform Lipschitz continuity of the double-well potential F in $u \in (-1,1)$. This will be achieved with a more effective use of the inequality

$$c(|\xi'|+|\xi|)^{p-2}|\xi'-\xi|^2 \le A(x,\xi') - A(x,\xi) - a(x,\xi)\cdot(\xi'-\xi) \tag{6.4}$$

for $1 < p < 2$; see the proof of Lemma 2.5.

Without loss of generality we may assume that $d < p$. Indeed, the additional hypothesis (1.11) in the case $d = p$ implies (1.12), contrary to our assumption.

We revisit the proof of Lemma 4.1 in section 5.1, now with $1 < p < 2$, and let h, u^*, and β be the same as there. We also introduce the weight

$$\omega = (|\nabla u| + |\nabla u^*|)^{1-p/2}.$$

Integrating (6.4) over B_r with $\xi' = \nabla u$ and $\xi = \nabla u^*$, we obtain that

$$c\int_{B_r} \frac{|\nabla\beta|^2}{\omega^2} \le \int_{B_r} A(x,\nabla u) - A(x,\nabla u^*) - a(x,\nabla u^*)\cdot\nabla(u-u^*). \tag{6.5}$$

To avoid complications, related to the vanishing of ω, we also introduce its "regularization"

$$\omega_\varepsilon = (|\nabla u| + |\nabla u^*| + \varepsilon)^{1-p/2}, \quad \varepsilon > 0.$$

Observe that we always have $\omega_\varepsilon > \omega$ and $\omega_\varepsilon \searrow \omega$ as $\varepsilon \searrow 0$.

Analyzing the proof in section 5.1, we realize that one can improve the step when the Sobolev and Young inequalities are applied to the function β^p; see (5.3). Indeed, let κ and $\lambda(x)$ be a certain positive number and a function, to be chosen later. Then

$$c\left(\int_{B_r} \beta^{\kappa\frac{n}{n-1}}\right)^{\frac{n-1}{n}} \le \int_{B_r} \beta^{\kappa-1}|\nabla\beta| = \int_{B_r} \left(\frac{|\nabla\beta|}{\omega_\varepsilon}\lambda\right)\left(\beta^{\kappa-1}\frac{\omega_\varepsilon}{\lambda}\right),$$

where in the first step we have applied the Sobolev inequality to the function β^κ. Now we use Young's inequality, with a certain parameter q, $1 < q < 2$, and its conjugate $q' = q/(q-1)$. We obtain

$$c\int_{B_r} \left(\frac{|\nabla\beta|}{\omega_\varepsilon}\lambda\right)\left(\beta^{\kappa-1}\frac{\omega_\varepsilon}{\lambda}\right) \le K^q \int_{B_r} \frac{|\nabla\beta|^q}{\omega_\varepsilon^q}\lambda^q + K^{-q'}\int_{B_r} \beta^{(\kappa-1)q'}\frac{\omega_\varepsilon^{q'}}{\lambda^{q'}}.$$

Applying the Hölder inequality with exponents $2/q$ and $2/(2-q)$ in both integrals on the right-hand side of the inequality above, we estimate it by

$$\begin{aligned} &K^q\left(\int_{B_r}\frac{|\nabla\beta|^2}{\omega_\varepsilon^2}\right)^{q/2}\left(\int_{B_r}\lambda^{2q/(2-q)}\right)^{1-q/2} \\ &+K^{-q'}\left(\int_{B_r}\beta^{(\kappa-1)q'(2/q)}\right)^{q/2}\left(\int_{B_r}\left(\frac{\omega_\varepsilon}{\lambda}\right)^{2q'/(2-q)}\right)^{1-q/2}. \end{aligned}$$

Let us now choose λ so that

$$\lambda^{2q/(2-q)} = \left(\frac{\omega_\varepsilon}{\lambda}\right)^{2q'/(2-q)}.$$

A simple computation shows that

$$\lambda = \omega_\varepsilon^{q'/(q+q')} = \omega_\varepsilon^{1/q}.$$

Then the expression above transforms to

$$\left[K^q \left(\int_{B_r} \frac{|\nabla\beta|^2}{\omega_\varepsilon^2}\right)^{q/2} + K^{-q'} \left(\int_{B_r} \beta^{(\kappa-1)q'(2/q)}\right)^{q/2}\right] \left(\int_{B_r} \omega_\varepsilon^{1/(1-q/2)}\right)^{1-q/2},$$

which is bounded from above by

$$C\left[K^2 \int_{B_r} \frac{|\nabla\beta|^2}{\omega_\varepsilon^2} + K^{-2q'/q} \int_{B_r} \beta^{(\kappa-1)q'(2/q)}\right]^{q/2} \left(\int_{B_r} \omega_\varepsilon^{1/(1-q/2)}\right)^{1-q/2}.$$

Collecting the estimates above and then letting $\varepsilon \to 0$, we will arrive at the inequality

$$\left(\int_{B_r} \beta^{\kappa\frac{n}{n-1}}\right)^{\frac{n-1}{n}} \leq$$
$$C\left[K^2 \int_{B_r} \frac{|\nabla\beta|^2}{\omega^2} + K^{-2q'/q} \int_{B_r} \beta^{(\kappa-1)q'(2/q)}\right]^{q/2} \left(\int_{B_r} \omega^{1/(1-q/2)}\right)^{1-q/2}.$$

Now observe that the term inside square brackets can be estimated similarly as in section 5.1, recalling also (6.5). We now have arrived at a point when we have to choose q. For that purpose we turn our attention to the term $\omega^{1/(1-q/2)}$. If we knew that ω is bounded, we could let $q \nearrow 2$. This is so, for instance, when F is uniformly Lipschitz in u, and we recover the proof in section 6.1 above. However, for non-Lipschitz F, we a priori know only the L^p integrability of $|\nabla u|$ and $|\nabla u^*|$. Moreover, we have

$$\int_{B_r} |\nabla u|^p + |\nabla u^*|^p \leq \int_{B_r} 2|\nabla u|^p + |\nabla h|^p \leq C\, r^{n-1},$$

by (4.1), for sufficiently large r. Thus, in order to obtain the desired density estimate, we choose q so that $\omega^{1/(1-q/2)} \simeq (|\nabla u|^p + |\nabla u^*|^p)$. Since $\omega = (|\nabla u| + |\nabla u^*|)^{1-p/2}$, we require

$$(1-p/2)/(1-q/2) = p \quad \Longleftrightarrow \quad q = 3 - \frac{2}{p}.$$

Observe that the condition $1 < q < 2$ is satisfied for $1 < p < 2$. As for the value of κ, we choose it to have

$$(\kappa-1)q'(2/q) = p \quad \Longleftrightarrow \quad \kappa = p.$$

Thus, with this choice of constants we obtain

$$\left(\int_{B_r} \beta^{p\frac{n}{n-1}}\right)^{\frac{n-1}{n}} \leq C\left[K^2 \int_{B_r} \frac{|\nabla\beta|^2}{\omega^2} + K^{-p'} \int_{B_r} \beta^p\right]^{q/2} \left(r^{n-1}\right)^{1-q/2}.$$

Now, using (6.5) and repeating the arguments as in section 5.1, we can deduce the following recursive inequality:

$$c\,\mathcal{V}_r^{\frac{n-1}{n}} \leq [\mathcal{V}_{r+1} - \mathcal{V}_r + \mathcal{A}_{r+1} - \mathcal{A}_r - c\mathcal{A}_r]^{q/2}\,(r^{n-1})^{1-q/2}. \tag{6.6}$$

Now the question is whether we can infer from (6.6) that

$$\mathcal{V}_r \geq c\, r^n$$

for $r \geq 1$ if $\mathcal{V}_1 \geq \mu > 0$. The answer is affirmative when

$$\frac{n-1}{n} \cdot \frac{2}{q} < 1 \quad \Longleftrightarrow \quad p > \frac{2n}{n+2}.$$

This follows from Lemma 2.1 with $\alpha = 1 - q/2$. (Unfortunately, (6.6) alone does not imply the density estimate for $p \leq 2n/(n+2)$, since we do need $\alpha < 1/n$ in Lemma 2.1.)

Summarizing, we obtain that for the range of the exponents $2n/(n+2) < p < 2$, one can drop the assumption (1.12) to prove Lemma 4.1. This completes the proof of Theorem 1.1.

7. Consequences of the density estimates. We briefly show in this section two consequences that can be easily derived from Theorem 1.1, thanks to the techniques developed in the last years.

The first consequence is that level sets of absolute minimizers converge, up to subsequence, to minimal interfaces in $L^\infty_{\rm loc}$. More precisely, it has been proved in [Bou90] that minimizers u_ε of

$$\mathcal{J}_\varepsilon(u;\Omega) = \int_\Omega A(x, \varepsilon\,\nabla u) + F(x,u) \tag{7.1}$$

converge, up to subsequence, in $L^1_{\rm loc}$ to a step function u_0 which has a minimal interface with respect to a suitably weighted area. Indeed, from the above density estimates we have that level sets converge in $L^\infty_{\rm loc}$.

THEOREM 7.1. *Fix $\theta \in (0,1)$. Let $|u_\varepsilon| \leq 1$ be an absolute minimizer of (7.1) in a bounded domain Ω. Assume that, as ε tends to zero, u_ε converges in $L^1_{\rm loc}$ to*

$$u_0 := \chi_E - \chi_{\Omega - E}$$

for a suitable $E \subset \Omega$. Then, $\{|u_\varepsilon| \leq \theta\}$ converges locally uniformly to ∂E.

The latter convergence is understood in the sense that $\operatorname{dist}(x, \partial E) \to 0$ uniformly for $x \in \{|u_\varepsilon| \leq \theta\} \cap K$ for any $K \subset\subset \Omega$.

Proof. The proof repeats the one of Theorem 2 in [CC95]. Assume that the claim of the theorem is not correct. Then there is $\delta > 0$, $K \subset\subset \Omega$, and $\varepsilon_n \to 0$, such that there exist $x_n \in \{|u_{\varepsilon_n}| < \theta\}$ with, say, $B_\delta(x_n) \subset E \cap K$. Since the rescalings $\tilde{u}_\varepsilon(x) := u_\varepsilon(\varepsilon x)$ are absolute minimizers of the normalized functional $\mathcal{J}$, applying the density estimates in Theorem 1.1 to $\tilde{u}_\varepsilon$ and then scaling back to u_ε, we will obtain that

$$\mathcal{L}^n(B_{\delta/2}(x_n) \cap \{u_{\varepsilon_n} < \theta\}) \geq c\,\delta^n$$

for some $c > 0$. But then,

$$\int_{B_{\delta/2}(x_n)} |u_{\varepsilon_n} - u_0| \geq c(1-\theta)\,\delta^n,$$

in contradiction with the hypothesis. This proves Theorem 7.1. ∎

Remark 7.2. We point out the following particular case of the above theorem. Let $F(x,u) = |1-u^2|^\alpha$ with $\alpha > 2$. Then it is unknown whether there is a uniform convergence of the level sets of minimizers in the singular perturbation problem

$$\int \varepsilon^2 |\nabla u|^2 + |1-u^2|^\alpha, \qquad \varepsilon \to 0+ .$$

However, if one perturbs with $\varepsilon^p |\nabla u|^p$ with $p \geq \alpha$, the uniform convergence follows from Theorem 7.1.

The second consequence of the density estimates is the existence of plane-like minimizers in the periodic setting. We say that u is a class A minimizer for $\mathcal{J}$ if it is an absolute minimizer for $\mathcal{J}$ in any ball B. With this setting, we can prove the following theorem.

THEOREM 7.3. *Assume that $A(x+e,\eta) = A(x,\eta)$ and $F(x+e,\eta) = F(x,\eta)$ for any $e \in \mathbb{Z}^n$. Fix $\theta \in (0,1)$. Then, there exists a positive constant M_0, depending only on θ and on the structural constants, such that, given any $\omega \in \mathbb{R}^n - \{0\}$, there exists a class A minimizer $u = u_\omega$ for the functional $\mathcal{J}$ for which the set $\{|u| \leq \theta\}$ is constrained in the strip $\{x \cdot \omega \in [0, M_0|\omega|]\}$.*

Furthermore, such u enjoys the following property of "quasi periodicity": if $\omega \in \mathbb{Q}^n - \{0\}$, then u is periodic (with respect to the identification induced by ω, i.e., $u(x+k) = u(x)$ for any $k \in \mathbb{Z}^n \cap \omega^\perp$); if $\omega \in \mathbb{R}^n - \mathbb{Q}^n$, then u can be approximated uniformly on compact sets by periodic class A minimizers.

Notice that M_0 above is *independent* of the frequency ω. These kinds of plane-like structures have been considered in [CdlL01] in the minimal surfaces case, and generalized to fluid jets and Ginzburg–Landau models in [Val04] and [PV03]. See also [Tor04] for a case with a degenerate metric. The proof of Theorem 7.3 is analogous to the one presented in section 8 of [PV03], with minor obvious changes, and we therefore omit the details.

REFERENCES

[AC81] H. W. ALT AND L. A. CAFFARELLI, *Existence and regularity for a minimum problem with free boundary*, J. Reine Angew. Math., 325 (1981), pp. 105–144.

[ACF84] H. W. ALT, L. A. CAFFARELLI, AND A. FRIEDMAN, *Jets with two fluids. I. One free boundary*, Indiana Univ. Math. J., 33 (1984), pp. 213–247.

[Ant73] S. S. ANTMAN, *Nonuniqueness of equilibrium states for bars in tension*, J. Math. Anal. Appl., 44 (1973), pp. 333–349.

[Bou90] G. BOUCHITTÉ, *Singular perturbations of variational problems arising from a two-phase transition model*, Appl. Math. Optim., 21 (1990), pp. 289–314.

[CC95] L. A. CAFFARELLI AND A. CÓRDOBA, *Uniform convergence of a singular perturbation problem*, Comm. Pure Appl. Math., 48 (1995), pp. 1–12.

[CdlL01] L. A. CAFFARELLI AND R. DE LA LLAVE, *Planelike minimizers in periodic media*, Comm. Pure Appl. Math., 54 (2001), pp. 1403–1441.

[GG82] M. GIAQUINTA AND E. GIUSTI, *On the regularity of the minima of variational integrals*, Acta Math., 148 (1982), pp. 31–46.

[Gur85] M. E. GURTIN, *On a theory of phase transitions with interfacial energy*, Arch. Rational Mech. Anal., 87 (1985), pp. 187–212.

[HKM93] J. HEINONEN, T. KILPELÄINEN, AND O. MARTIO, *Nonlinear Potential Theory of Degenerate Elliptic Equations*, Oxford Math. Monogr., The Clarendon Press, Oxford University Press, New York, 1993.

[MM77] L. MODICA AND S. MORTOLA, *Un esempio di Γ-convergenza*, Boll. Un. Mat. Ital. B (5), 14 (1977), pp. 285–299.

[PV03] A. PETROSYAN AND E. VALDINOCI, *Geometric properties of Bernoulli-type minimizers*, Interfaces Free Bound., to appear.

[Row79] J. S. Rowlinson, *Translation of J. D. van der Waals' "The thermodynamic theory of capillarity under the hypothesis of a continuous variation of density,"* J. Statist. Phys., 20 (1979), pp. 197–244.

[Tol84] P. Tolksdorf, *Regularity for a more general class of quasilinear elliptic equations*, J. Differential Equations, 51 (1984), pp. 126–150.

[Tor04] M. Torres, *Plane-like minimal surfaces in periodic media with exclusions*, SIAM J. Math. Anal. 36 (2004), pp. 523–551.

[Val04] E. Valdinoci, *Plane-like minimizers in periodic media: Jet flows and Ginzburg–Landau-type functionals*, J. Reine Angew. Math., 574 (2004), pp. 147–185.

SIAM J. MATH. ANAL.
Vol. 36, No. 4, pp. 1080–1094

MINIMIZATION METHODS FOR QUASI-LINEAR PROBLEMS WITH AN APPLICATION TO PERIODIC WATER WAVES*

B. BUFFONI†, É. SÉRÉ‡, AND J. F. TOLAND§

Abstract. Penalization and minimization methods are used to give an abstract "semiglobal" result on the existence of nontrivial solutions of parameter-dependent quasi-linear differential equations in variational form. A consequence is a proof of existence, by infinite-dimensional variational means, of bifurcation points for quasi-linear equations which have a line of trivial solutions.

The approach is to penalize the functional twice. Minimization gives the existence of critical points of the resulting problem, and a priori estimates show that the critical points lie in a region unaffected by the leading penalization. The other penalization contributes to the value of the parameter.

As applications we prove the existence of periodic water waves, with and without surface tension.

Key words. variational method, critical-point theory, minimization, quasi-linear elliptic problems, periodic water waves, free boundaries

AMS subject classifications. 76B15, 35B38, 58E50

DOI. 10.1137/S0036141003432766

1. Introduction.

Using local finite-dimensional reduction followed by a constrained minimization argument, Stuart [20], following Krasnosel'skii [15], gave a bifurcation theory for variational problems which was applied in [6] to Babenko's [2] quasi-linear equation for Stokes waves. But how to deal with such problems directly, using variational methods in infinite dimensions, remained unclear. In [8] we took a step in this direction by adapting some ideas of Turner [22] and the mountain-pass lemma in infinite dimensions to obtain an existence theory which was "semiglobal," in the sense that parameter values were quantifiably not infinitesimally small, and finite-dimensional reduction was not involved.

Now, in section 2, we present an abstract result that covers a general class of quasi-linear problems. In section 3 we show that the existence problem for two-dimensional capillary-gravity waves on a flow of infinite depth, with its curvature term that represents surface tension effects, is a special case. Our method should give explicit (and hopefully good) lower bounds on the size of the periodic capillary-gravity waves so obtained, but here the aim is merely to illustrate the generality of the abstract result. Moreover, our abstract method, based on a direct minimization, could probably be enriched to encompass the various kinds of two-dimensional periodic capillary-gravity waves found in [10, 14, 17] by extending it to more involved minimax principles. In section 4, we apply the present method to the Stokes-wave problem (steady periodic water waves without surface tension). As in [8], we obtain the existence of a nonzero symmetric Stokes wave which is not a consequence of local existence theories [1, 6, 20].

*Received by the editors July 30, 2003; accepted for publication (in revised form) April 16, 2004; published electronically January 27, 2005.

http://www.siam.org/journals/sima/36-4/43276.html

†Institut de Mathématiques, École Polytechnique Fédérale, Lausanne, CH 1015, Switzerland (boris.buffoni@epfl.ch). Supported by the Swiss National Science Foundation.

‡CEREMADE, Université de Paris Dauphine, Pl. du Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16, France (sere@ceremade.dauphine.fr). Partially supported by the "Institut Universitaire de France."

§Department of Mathematical Sciences, University of Bath, Claverton Down, Bath, BA2 7AY, UK (jft@maths.bath.ac.uk).

In contrast to [8], we no longer appeal to the mountain-pass principle or to Morse theory, and therefore the present proof is simpler.

The vast literature on steady two-dimensional water waves, with or without surface tension, which has been developed since the work of Stokes [19] in the middle of the nineteenth century, is surveyed from various viewpoints in [12, 13, 18, 23]. A great deal is now known about existence theory using global continuation methods [5, 6, 7], numerical investigations [9, 16], and computer assisted proofs [3, 4]. But the need for a global or "semiglobal" variational theory of these and similarly degenerate variational problems, including Morse indices of solutions, remains our focus. An extension of such "semiglobal" methods to cover the three-dimensional waves considered in [10] would be most interesting but is beyond the scope of the present work.

2. Abstract setting. Consider a real Hilbert space X_0 with inner-product $\langle\cdot,\cdot\rangle_0$ and norm $\|\cdot\|_0$, and suppose that A is a (possibly unbounded) positive-definite self-adjoint operator on X_0 such that $A^{-1} : X_0 \to X_0$ exists and is continuous. For $k \geq 1$ let X_k denote the domain of $A^{k/2}$, which is dense in X_0. Then X_k is a Hilbert space with inner-product and norm defined by $\langle u, v\rangle_k = \langle A^{k/2}u, A^{k/2}v\rangle_0$ and $\|u\|_k = \|A^{k/2}u\|_0$ for u, $v \in X_k$, and

$$\|u\|_k \leq \|A^{-1/2}\|\,\|u\|_{k+1} \quad \text{for all} \quad u \in X_{k+1}.$$

For $R_2 > 0$, let $U \subset X_2$ be the open ball $\{u \in X_2 : \|u\|_2 < R_2\}$, and suppose that $\mathcal{K}$, $\mathcal{L} \in C^1(U;\mathbb{R})$ are functionals with Fréchet derivatives at u denoted by $\partial\mathcal{K}(u)$ and $\partial\mathcal{L}(u)$. We are interested in the equation

$$(\star) \qquad \gamma\partial\mathcal{K}(w) + \partial\mathcal{L}(w) = 0, \quad w \in U \setminus \{0\}, \quad \gamma \geq \gamma_0 \geq 0,$$

when the following inequalities hold for constants $C_1\,, C_2 > 0$ and for a continuous function $\psi : [0,\infty)^2 \to \mathbb{R}$, the precise form of which depends on the problem:

$$\text{for} \quad u \in U : \quad \mathcal{K}(u) \geq C_1\|u\|_1^2 \quad \text{and } \mathcal{K}(0) = 0, \tag{2.1a}$$

$$\text{for} \quad u \in U : \quad \mathcal{L}(u) \geq -C_2\|u\|_1^2 \text{ and } \mathcal{L}(0) = 0, \tag{2.1b}$$

$$\text{for} \quad u \in U \cap X_4 : \partial\mathcal{K}(u)Au \geq 0 \quad \text{and} \tag{2.1c}$$

$$\gamma_0\partial\mathcal{K}(u)Au + \partial\mathcal{L}(u)Au \geq \psi(\|u\|_1\,, \|u\|_2). \tag{2.1d}$$

Observe that $\partial\mathcal{K}(0) = \partial\mathcal{L}(0) = 0$ so that $w = 0$ is a trivial solution of $(\star)$.

Roughly speaking, the function ψ takes positive and negative values and will be such that $\psi(s,t) > 0$ when s is "not too large" and t is "not small" (see assumption (2.2) below). In the existence proof, we construct a functional $\mathcal{J}$ on U, whose minimizer w satisfies $\gamma_0\partial\mathcal{K}(w)Aw + \partial\mathcal{L}(w)Aw \leq 0$ with $s = \|w\|_1$ "not too large." This yields an upper bound, better than R_2, on $t = \|w\|_2$ which ensures that w is solution of our problem. To verify the assumptions in practice, it will often be necessary to choose U small enough, whence the term "semiglobal" in the abstract.

The next hypothesis is about weak solutions of a regularized problem: for all $\gamma \geq \gamma_0$, $\epsilon > 0$, and $w \in U$,

$$\text{if} \quad \gamma\partial\mathcal{K}(w) + \partial\mathcal{L}(w) + \epsilon A^2 w = 0 \quad \text{in } X_2^*\,, \text{ then } w \in X_4. \tag{2.1e}$$

Finally, we make the following assumptions:

$$\mathcal{K} \text{ and } \mathcal{L} \quad \text{are weakly lower semicontinuous on } U \subset X_2. \tag{2.1f}$$

$$\text{There exists } u_* \in U \text{ with } \gamma_0\mathcal{K}(u_*) + \mathcal{L}(u_*) < 0 = \gamma_0\mathcal{K}(0) + \mathcal{L}(0). \tag{2.1g}$$

THEOREM 1. *Suppose that hypotheses* (2.1) *hold and that*

$$\psi(s, R_2) > 0 \quad \text{for all} \quad s \in \left[0,\, \sqrt{\mathcal{K}(u_*)/C_1}\right]. \tag{2.2}$$

Then there exist $w \in U\backslash\{0\}$ *and* $\gamma \geq \gamma_0$ *such that*

$$\begin{gathered} \gamma \partial \mathcal{K}(w) + \partial \mathcal{L}(w) = 0, \\ \gamma_0 \mathcal{K}(w) + \mathcal{L}(w) \leq \gamma_0 \mathcal{K}(u_*) + \mathcal{L}(u_*) < 0, \end{gathered} \tag{$\star$}$$

and, as a consequence of (2.1a) *and* (2.1b),

$$\|w\|_1^2 \geq \frac{\gamma_0 \mathcal{K}(u_*) + \mathcal{L}(u_*)}{\gamma_0 C_1 - C_2} > 0.$$

Remarks.

1. Note that in (2.1a) and (2.1b), $\gamma_0 C_1 - C_2 < 0$; otherwise u_* would not exist.
2. A typical form of the function ψ is $\alpha t^2 - \varphi(s,t)$, where $\alpha > 0$, φ is continuous from $[0,\infty)^2$ to $\mathbb{R}$, and $\varphi(0, R_2) = 0$. Then condition (2.2) becomes

$$\varphi(s, R_2) \,<\, \alpha R_2^2 \quad \text{for all} \quad s \in \left[0,\, \sqrt{\mathcal{K}(u_*)/C_1}\right]. \tag{2.3a}$$

This condition is satisfied for $\mathcal{K}(u_*)$ small enough, by continuity of φ at $(0, R_2)$. Note that α may depend on R_2.

3. In the examples of sections 3 and 4 the function ψ is constructed in two steps. First we establish an inequality

$$\partial \mathcal{K}(u) Au \geq C_3 \|u\|_2^2 \quad \text{for all} \quad u \in U \cap X_4, \tag{2.3b}$$

for some constant $C_3(R_2) > 0$. Then we find a function $\varphi(s,t)$ and a constant $C_4 \in \mathbb{R}$ such that

$$\partial \mathcal{L}(u) Au \geq -\varphi(\|u\|_1, \|u\|_2) - C_4 \|u\|_2^2 \quad \text{for all} \quad u \in U \cap X_4. \tag{2.3c}$$

If $\gamma_0 > C_4/C_3$, the estimate (2.1d) will follow for $\psi(s,t) := \alpha t^2 - \varphi(s,t)$, with $\alpha := \gamma_0 C_3 - C_4 > 0$.

Proof of Theorem 1. Let $R_1, R_{\min} > 0$ be finite numbers such that

$$\mathcal{K}(u_*) < R_1^2, \quad \|u_*\|_2 \leq R_{\min} < R_2$$

and

$$\psi(s,t) > 0 \quad \text{for all} \quad s \in \left[0,\, R_1/\sqrt{C_1}\right] \quad \text{and} \quad t \in \left[R_{\min}, R_2\right]. \tag{2.4}$$

These numbers exist because of assumption (2.2) and the uniform continuity of ψ on $\left[0\,,\, 2\sqrt{\mathcal{K}(u_*)/C_1}\right] \times \left[0\,,\, R_2\right]$. We define two smooth, nondecreasing *penalization* functions $\rho_i : [0, R_i^2) \to \mathbb{R}$ such that

$$\begin{gathered} \rho_i(s) \to \infty \text{ as } s \nearrow R_i^2, \quad i = 1,\, 2, \\ 0 \leq s \leq \mathcal{K}(u_*) \Rightarrow \rho_1(s) = 0, \quad 0 \leq s \leq R_{\min}^2 \Rightarrow \rho_2(s) = 0, \end{gathered}$$

and consider the real-valued functional defined for $u \in U$ with $\mathcal{K}(u) < R_1^2$ by

$$\mathcal{J}(u) = \gamma_0 \mathcal{K}(u) + \mathcal{L}(u) + \rho_2(\|u\|_2^2) + \rho_1(\mathcal{K}(u)).$$

The indices denote the facts that ρ_2 and ρ_1 control, respectively, the norms in X_2 and X_1 (through the constant C_1). We refer to ρ_2 as the *leading penalization*. Together they allow us to work in the domain $V := \{u \in U : \mathcal{K}(u) < R_1^2\}$. The inequalities (2.1a), (2.1d), (2.2) will ensure that critical points w of $\mathcal{J}$ in this domain satisfy $\|w\|_2 \le R_{\min}$ and are thus unaffected by the leading penalization. At critical points the other penalization may be nonzero, which leads to the loss of control of the value of the parameter γ in the statement of the theorem.

We must now find a critical point of $\mathcal{J}$, and a natural idea is to look for a minimizer. Note that $\mathcal{J}$ is bounded from below on V, with $\mathcal{J}(u) \to \infty$ as $\|u\|_2 \nearrow R_2$ and $\mathcal{J}(u) \to \infty$ as $\mathcal{K}(u) \nearrow R_1^2$, by the existence of the constants C_1 and C_2. From (2.1f) it follows that $\mathcal{J}$ has a minimizer $w \in V$. We have $\mathcal{K}(w) < R_1^2$, $\|w\|_2 < R_2$, and w is a weak solution of the Euler equation

$$\{\gamma_0 + \rho_1'(\mathcal{K}(w))\}\partial\mathcal{K}(w) + \partial\mathcal{L}(w) + 2\rho_2'(\|w\|_2^2)A^2w = 0. \tag{2.5}$$

Seeking a contradiction, assume that $\epsilon := 2\rho_2'(\|w\|_2^2) > 0$. Then, by (2.1e), $w \in X_4$. Hence $Aw \in X_2$, and we have $\partial\mathcal{J}(w)Aw = 0$. From (2.1c) and (2.1d) it follows that

$$\psi(\|w\|_1, \|w\|_2) \le -\rho_1'(\mathcal{K}(w))\partial\mathcal{K}(w)Aw - 2\rho_2'(\|w\|_2^2)\|w\|_3^2 < 0.$$

Since $\mathcal{K}(w) < R_1^2$, (2.1a) gives the estimate $\|w\|_1 \le R_1/\sqrt{C_1}$. Therefore, by (2.4), $\|w\|_2 < R_{\min}$. This shows that $\rho_2'(\|w\|_2^2) = 0$, which is the required contradiction.

Hence, we have proved that $w \in U$ satisfies (2.5), with $\rho_2'(\|w\|_2^2) = 0$. Therefore

$$\gamma\partial\mathcal{K}(w) + \partial\mathcal{L}(w) = 0 \text{ with } \gamma := \gamma_0 + \rho_1'(\mathcal{K}(w)).$$

Since w is a minimizer of $\mathcal{J}$ and $\rho_1(\mathcal{K}(u_*)) = \rho_2(\|u_*\|_2^2) = 0$, we have

$$\gamma_0\mathcal{K}(w) + \mathcal{L}(w) \le \mathcal{J}(w) \le \mathcal{J}(u_*) = \gamma_0\mathcal{K}(u_*) + \mathcal{L}(u_*) < 0.$$

The critical point w is thus nonzero, and we have the estimate

$$(\gamma_0 C_1 - C_2)\|w\|_1^2 \le \gamma_0\mathcal{K}(u_*) + \mathcal{L}(u_*) < 0. \qquad \square$$

Additional hypotheses yield more information on the critical point w.

THEOREM 2. *Suppose that* (2.1) *holds and*

$$\partial\mathcal{K}(u) + \mu A^2 u \ne 0 \quad \textit{for all} \quad u \in U \setminus \{0\} \quad \textit{and} \quad \mu \ge 0. \tag{2.6}$$

Let the two constants $\underline{R}, \overline{R}$ *satisfy* $\|u_*\|_2 \le \overline{R} < R_2$, $\underline{R} \ge \sqrt{\mathcal{K}(u_*)}$, *and (instead of* (2.2)*) suppose that*

$$\psi(s, \overline{R}) \ge 0 \quad \textit{for all} \quad s \in \big[0, \underline{R}/\sqrt{C_1}\big]. \tag{2.7}$$

Then there exists $w \in U\setminus\{0\}$ *such that* $\|w\|_2 \le \overline{R}$, $\mathcal{K}(w) \le \underline{R}^2$,

$$\gamma_0\mathcal{K}(w) + \mathcal{L}(w) = \min\{\gamma_0\mathcal{K}(u) + \mathcal{L}(u) : \ u \in U, \ \|u\|_2 \le \overline{R}, \ \mathcal{K}(u) \le \underline{R}^2\},$$

and the following hold:

(i) *if* $\mathcal{K}(w) < \underline{R}^2$, *then* $\gamma_0\partial\mathcal{K}(w) + \partial\mathcal{L}(w) = 0$;

(ii) *if* $\mathcal{K}(w) = \underline{R}^2$, *then* $\gamma\partial\mathcal{K}(w) + \partial\mathcal{L}(w) = 0$ *for some* $\gamma \ge \gamma_0$.

Moreover, as a consequence of properties (2.1a), (2.1b),

$$\|w\|_1^2 \geq \frac{\gamma_0 \mathcal{K}(u_*) + \mathcal{L}(u_*)}{\gamma_0 C_1 - C_2} > 0.$$

Proof. By assumptions (2.1a) and (2.1b), the functional $\mathcal{I}(u) := \gamma_0 \mathcal{K}(u) + \mathcal{L}(u)$ is bounded from below on the set $C := \{u \in X_2 : \mathcal{K}(u) \leq \underline{R}^2, \|u\|_2 \leq \overline{R}\}$. By assumption (2.1f), the set C is weakly closed in X_2, and $\mathcal{I}$ is weakly lower semicontinuous. So there exists a minimizer w of $\mathcal{I}$ on C. Since $u_* \in C$,

$$\mathcal{I}(w) \leq \mathcal{I}(u_*) < 0.$$

Hence $(\gamma_0 C_1 - C_2)\|w\|_1^2 \leq \mathcal{I}(u_*) < 0$. By assumption (2.6) and the general theorem on Lagrange multipliers, w is a weak solution of

$$\epsilon A^2 w + \gamma \partial \mathcal{K}(w) + \partial \mathcal{L}(w) = 0$$

for some $\gamma \geq \gamma_0$ and $\epsilon \geq 0$ and

$$(\gamma - \gamma_0)(\underline{R}^2 - \mathcal{K}(w)) = 0 = \epsilon\,(\overline{R} - \|w\|_2).$$

It follows by contradiction, as in the proof of Theorem 1, that $\epsilon = 0$ (this follows from (2.1a), (2.1d), (2.7)). The alternative (i)–(ii) is thus satisfied. ∎

3. Gravity-capillary water waves. Let $L^2_{2\pi}$ denote the usual real Banach space of 2π-periodic, real-valued, square-integrable measurable "functions" on $\mathbb{R}$, and let $L^\infty_{2\pi}$ denote the analogous space of essentially bounded functions. We denote by $C^n_{2\pi}$ (resp., $C^\infty_{2\pi}$) the space of 2π-periodic functions u which are n times continuously differentiable (resp., infinitely differentiable).

With respect to the orthonormal basis $\{(2\pi)^{-\frac{1}{2}} e^{ikt} : k \in \mathbb{Z}\}$, let the Fourier coefficients of $u \in L^2_{2\pi}$ be denoted by $\hat{u}_k$, $k \in \mathbb{Z}$. Then $\hat{u}_{-k} = \overline{\hat{u}_k}$, since u is real, and $L^2_{2\pi}$ is a real Hilbert space with inner-product

$$\langle u, v \rangle = \sum_{k \in \mathbb{Z}} \hat{u}_k \overline{\hat{v}_k}.$$

For $u \in L^2_{2\pi}$ let

$$[u] = \frac{1}{2\pi} \int_{-\pi}^{\pi} u(t)\, dt = \frac{\hat{u}_0}{\sqrt{2\pi}}.$$

The fractional order Sobolev space $H^s_{2\pi}$, $s \geq 0$, is the Hilbert space of functions $u \in L^2_{2\pi}$ with norm given by

$$\|u\|_s^2 = \hat{u}_0^2 + \sum_{k \in \mathbb{Z}} |k|^{2s} |\hat{u}_k|^2 < \infty.$$

Note that if $u \in C^n_{2\pi}$, $k \in \mathbb{N} \cup 0$, then

$$\|u\|_n^2 = 2\pi [u]^2 + \|u^{(n)}\|^2_{L^2_{2\pi}},$$

where $u^{(n)}$ denotes the nth derivative of u. The conjugation operation [24] on $L^2_{2\pi}$ is defined by

$$(\widehat{\mathcal{C}u})_0 = 0 \text{ and } (\widehat{\mathcal{C}u})_k = -i \operatorname{sgn}(k) \hat{u}_k \text{ for } k \in \mathbb{Z} \setminus \{0\}, \text{ when } u \in L^2_{2\pi};$$

equivalently, $\mathcal{C}(\cos kt) = \sin kt, k \geq 0$, and $\mathcal{C}(\sin kt) = -\cos kt, k \geq 1$. Clearly $\mathcal{C} : L^2_{2\pi} \to L^2_{2\pi}$ is a bounded linear operator and $u \mapsto \mathcal{C}u'$ is nonnegative and symmetric in the sense that

$$0 \leq \langle u, \mathcal{C}u' \rangle \text{ and } \langle u, \mathcal{C}v' \rangle = \langle \mathcal{C}u', v \rangle \text{ for all } u,\, v \in C^\infty_{2\pi}.$$

For any function $w \in H^1_{2\pi}$ with $[w] = 0$, writing $w + i\mathcal{C}w = \sqrt{\frac{2}{\pi}} \sum_{k>0} \hat{w}_k e^{ikt}$, one gets

$$\begin{aligned} \|w + i\mathcal{C}w\|_\infty &\leq \frac{1}{\sqrt{2\pi}} \sum_{k \neq 0} |\hat{w}_k| \leq \frac{1}{\sqrt{2\pi}} \left(\sum_{k \neq 0} k^2 |\hat{w}_k|^2 \right)^{1/2} \left(\sum_{k \neq 0} \frac{1}{k^2} \right)^{1/2} \\ &= \sqrt{\frac{\pi}{6}} \left(\sum_{k \neq 0} k^2 |\hat{w}_k|^2 \right)^{1/2} = \sqrt{\frac{\pi}{6}} \|w\|_1 = \sqrt{\frac{\pi}{6}} \|w'\|_{L^2_{2\pi}}. \end{aligned} \tag{3.1}$$

When surface-tension effects are included, the steady water-wave problem can be formulated as follows [6]: find w such that

$$\frac{\nu^2}{2} \{w'^2 + (1 + \mathcal{C}w')^2\}^{-1} + \lambda w - \beta \frac{(1 + \mathcal{C}w')w'' - w'(1 + \mathcal{C}w')'}{\{w'^2 + (1 + \mathcal{C}w')^2\}^{3/2}} = \frac{1}{2}\nu^2, \tag{3.2a}$$

$$w'^2 + (1 + \mathcal{C}w')^2 > 0, \quad w \in H^2_{2\pi} \setminus \{0\}, \quad \lambda \geq 0, \quad \beta, \nu > 0. \tag{3.2b}$$

Here β is the coefficient of surface tension and the parameters λ and ν^2 are defined in terms of the wavelength 2Λ, the wave speed c, the gravitational acceleration g, and the density d by

$$\lambda = \frac{g\Lambda^2 d}{\pi^2}, \quad \nu^2 = \frac{\Lambda c^2 d}{\pi}.$$

Note that (3.2) is not a variational problem as it stands. However, it is known [6] that (3.2) is satisfied by any $w \in H^2_{2\pi}$ such that $w'^2 + (1 + \mathcal{C}w')^2 > 0$ and such that, almost everywhere,

$$\begin{aligned} 0 = {} & -\nu^2 \mathcal{C}w' + \lambda \{w + w\mathcal{C}w' + \mathcal{C}(ww')\} \\ & - \beta \left\{ \frac{w'}{\sqrt{w'^2 + (1 + \mathcal{C}w')^2}} \right\}' + \beta \mathcal{C} \left\{ \frac{1 + \mathcal{C}w'}{\sqrt{w'^2 + (1 + \mathcal{C}w')^2}} \right\}'. \end{aligned} \tag{3.3}$$

Equation (3.3) is the Euler equation of the functional

$$\begin{aligned} J(w) = \int_{-\pi}^{\pi} \Big\{ & -\frac{1}{2}\nu^2 w \mathcal{C}w' + \frac{1}{2}\lambda w^2 (1 + \mathcal{C}w') \\ & + \beta \sqrt{w'^2 + (1 + \mathcal{C}w')^2} - \beta(1 + \mathcal{C}w') \Big\} dt. \end{aligned}$$

For all w, the integral of the last term is $-2\beta\pi$, and it does not contribute to the variational principle (it is a null Lagrangian). It is included here only to ensure that the constant and linear parts of the integrand vanish when $w = 0$.

Observe that, when $\lambda = 0$, every constant function w is a solution of (3.3) and any translate of a solution is also a solution. These superfluous solutions complicate

the problem unnecessarily, and, to eliminate them, we work in the subspace X_2 of $H^2_{2\pi}$ consisting of even functions of zero mean with norm given by

$$\|w\|^2_{X_2} = \int_{-\pi}^{\pi} |w''(t)|^2\, dt = \|w\|_2^2.$$

The critical points of J under the constraint $[w] = 0$ satisfy (3.3) almost everywhere, but with 0 on the left-hand side replaced by the constant $\lambda[w\mathcal{C}w']$. So instead we consider the functional $\widetilde{J}$ defined on X_2 by

$$\widetilde{J}(w) := J(w) - \frac{\lambda}{4\pi}\left\{\int_{-\pi}^{\pi} w\mathcal{C}w'\,dt\right\}^2.$$

Critical points $w \in X_2$ of $\widetilde{J}$ satisfy

$$\frac{\lambda}{2\pi}\int_{-\pi}^{\pi} w\mathcal{C}w'\,dt = -\left(\nu^2 + \frac{\lambda}{\pi}\int_{-\pi}^{\pi} w\mathcal{C}w'\,dt\right)\mathcal{C}w' + \lambda(w + w\mathcal{C}w' + \mathcal{C}(ww')) + \beta\left\{\frac{-w'}{\sqrt{w'^2 + (1+\mathcal{C}w')^2}} + \mathcal{C}\frac{1+\mathcal{C}w'}{\sqrt{w'^2 + (1+\mathcal{C}w')^2}}\right\}', \tag{3.4}$$

and $\widetilde{w} := w - [w\mathcal{C}w']$ satisfies (3.3), from which (3.2a) follows.

Since we can divide (3.2) by any one of the parameters ν^2, λ, and β, there are effectively only two dimensionless parameters in the problem. Now divide J, $\widetilde{J}$, and (3.3) by ν^2 so that, in the remainder of section 3,

$$\lambda = \frac{g\Lambda}{\pi c^2}, \quad \nu^2 = 1, \tag{3.5a}$$

and β has been replaced by the dimensionless parameter

$$\gamma = \frac{\beta\pi}{\Lambda c^2 d}. \tag{3.5b}$$

We now apply the abstract result of section 1 to $\widetilde{J}$. To put the functional $\widetilde{J}$ in the context of section 2, let

$$\begin{aligned} &X_0 = \{w \in L^2_{2\pi} : [w] = 0,\ w \text{ is even}\},\\ &Aw = -w'',\\ &X_k = \{w \in H^k_{2\pi} : [w] = 0,\ w \text{ is even}\} \quad (k \geq 1). \end{aligned}$$

If $R_2 < \sqrt{6/\pi}$, then (3.1) implies that

$$\|w' + i\mathcal{C}w'\|_\infty \leq \sqrt{\pi/6}R_2 < 1 \text{ when } \|w\|_{X_2} < R_2\ . \tag{3.6}$$

For $w \in U$, the ball of radius R_2 centered at the origin in X_2, let

$$\begin{aligned} \mathcal{K}(w) &= \int_{-\pi}^{\pi} \sqrt{w'^2 + (1+\mathcal{C}w')^2} - (1+\mathcal{C}w')\ dt\\ &= \int_{-\pi}^{\pi} \frac{w'^2 dt}{|1+\mathcal{C}w' - iw'| + (1+\mathcal{C}w')}\ ,\\ \mathcal{L}(w) &= -\frac{1}{2}\int_{-\pi}^{\pi} w\mathcal{C}w'dt - \frac{\lambda}{4\pi}\left\{\int_{-\pi}^{\pi} w\mathcal{C}w'dt\right\}^2 + \frac{\lambda}{2}\int_{-\pi}^{\pi} w^2(1+\mathcal{C}w')dt. \end{aligned}$$

With β in (3.4) replaced by γ and ν^2 by 1, we check the hypotheses of Theorem 1 for $0 < R_2 < \sqrt{6/\pi}$ small enough. Obviously, $\mathcal{L}$ is of class C^1 on X_2. Since (3.6) holds, we have

$$0 < 2(1 - \sqrt{\pi/6}R_2) \le |1 + \mathcal{C}w' - iw'| + (1 + \mathcal{C}w') \le 2(1 + \sqrt{\pi/6}R_2). \tag{3.7}$$

So $\mathcal{K}$ is of class C^1 on U. Moreover, if we define

$$C_1(R_2) := \frac{1}{2(1 + \sqrt{\pi/6}R_2)}, \tag{3.8}$$

then

$$C_1\|w'\|^2_{L^2_{2\pi}} \le \int_{-\pi}^{\pi} \frac{w'^2 dt}{|1 + \mathcal{C}w' - iw'| + (1 + \mathcal{C}w')} = \mathcal{K}(w).$$

So (2.1a) is satisfied.

Now to check (2.1b) for all $w \in U$ and $\lambda > 0$, note that

$$\begin{aligned}
\mathcal{L}(w) &= \int_{-\pi}^{\pi} \left\{ -\frac{1}{2} w\mathcal{C}w' + \frac{\lambda}{2} w^2(1 + \mathcal{C}w') \right\} dt - \frac{\lambda}{4\pi} \left\{ \int_{-\pi}^{\pi} w\mathcal{C}w' dt \right\}^2 \\
&\ge -\frac{1}{2}\|w\|_1^2 + \frac{\lambda}{2}(1 - \sqrt{\pi/6}\|w\|_2)\|w\|_0^2 - \frac{\lambda}{4\pi}\|w\|_2^2\|w\|_0^2 \\
&\ge \frac{\lambda}{2}\left(1 - \sqrt{\pi/6}R_2 - \frac{1}{2\pi}R_2^2\right)\|w\|_0^2 - \frac{1}{2}\|w\|_1^2 \\
&\ge -\frac{1}{2}\|w\|_1^2
\end{aligned}$$

under the condition $\sqrt{\pi/6}R_2 + \frac{1}{2\pi}R_2^2 \le 1$, which is satisfied, for instance, when $R_2 < 1$. So, under this restriction on R_2, (2.1b) holds for $C_2 := 1/2$.

Next we show the existence of $C_3(R_2) > 0$ satisfying (2.3b) for R_2 small enough.

LEMMA 3. *If $w \in X_4$ is such that $\|w\|_{X_2} < R_2$ with $0 < R_2 < \sqrt{3/4\pi}$, then*

$$\partial\mathcal{K}(w)\, Aw \ge C_3(R_2)\|w\|^2_{X_2},$$

where

$$C_3(R_2) := \frac{1 - 2\sqrt{\pi/3}R_2}{(1 + \sqrt{\pi/6}R_2)^3}. \tag{3.9}$$

Proof. The product $\partial\mathcal{K}(w)Aw$ is the directional derivative of the functional $\mathcal{K}$ at w in the direction $-w''$. It is also the derivative of the length of the parametrized curve $\{c(t) = (t + \mathcal{C}w(t), w(t)), 0 \le t \le 2\pi\}$ in the direction $\{\delta(t) = (-\mathcal{C}w''(t), -w''(t)),$

$0 \le t \le 2\pi\}$. Using this interpretation, one easily gets the formula

$$\begin{aligned}
\partial\mathcal{K}(w)Aw &= \int_{-\pi}^{\pi} \frac{\{(1+\mathcal{C}w')w'' - w'\mathcal{C}w''\}^2}{\{(1+\mathcal{C}w')^2 + w'^2\}^{3/2}}dt \\
&= \int_{-\pi}^{\pi} \frac{|w'' + \mathcal{R}e\{(\mathcal{C}w' + iw')(w'' + i\mathcal{C}w'')\}|^2}{|1+\mathcal{C}w' + iw'|^3}dt \\
&\ge \int_{-\pi}^{\pi} \frac{|w''|^2 - 2|\mathcal{C}w' + iw'|\,|w'' + i\mathcal{C}w''|\,|w''|}{(1+\sqrt{\pi/6}R_2)^3}dt \\
&\ge \int_{-\pi}^{\pi} \frac{|w''|^2 - 2\sqrt{\pi/6}R_2\,|w'' + i\mathcal{C}w''|\,|w''|}{(1+\sqrt{\pi/6}R_2)^3}dt \\
&\ge \frac{1 - 2\sqrt{\pi/3}R_2}{(1+\sqrt{\pi/6}R_2)^3}\int_{-\pi}^{\pi} |w''|^2\,dt
\end{aligned}$$

since (3.6) holds. □

We now construct a function φ such that (2.3c) is true when $C_4 = 0$. For $\lambda \ge 0$ and $w \in U \cap X_4$, we have

$$\begin{aligned}
\partial\mathcal{L}(w)(-w'') &= -\int_{-\pi}^{\pi} w'\mathcal{C}w''dt - \frac{\lambda}{\pi}\left(\int_{-\pi}^{\pi} w\mathcal{C}w'dt\right)\left(\int_{-\pi}^{\pi} w'\mathcal{C}w''dt\right) \\
&\quad - \lambda\int_{-\pi}^{\pi} \big(w + w\mathcal{C}w' + \mathcal{C}(ww')\big)w''dt \\
&= -\int_{-\pi}^{\pi} w'\mathcal{C}w''dt + \lambda\int_{-\pi}^{\pi} w'^2dt - \lambda\int_{-\pi}^{\pi} w\big(w''\mathcal{C}w' - w'\mathcal{C}w''\big)dt \\
&\quad - \frac{\lambda}{\pi}\left(\int_{-\pi}^{\pi} w\mathcal{C}w'dt\right)\left(\int_{-\pi}^{\pi} w'\mathcal{C}w''dt\right) \\
&\ge -\|w\|_2\|w\|_1 + \lambda\|w\|_1^2 - 2\lambda\|w\|_\infty\|w\|_2\|w\|_1 - \frac{\lambda}{\pi}\|w\|_1^2\|w\|_2^2 \\
&\ge -\varphi(\|w\|_1, \|w\|_2),
\end{aligned}$$

where

$$\varphi(s,t) := ts - \lambda\big(1 - \sqrt{2\pi/3}\,t - t^2/\pi\big)s^2.$$

Let $\gamma_0 > 0$ be given. Then, in (2.1d), we can choose

$$\psi(s,t) := \gamma_0 C_3(R_2)t^2 - \varphi(s,t).$$

If $w \in U$ is a weak solution of $\partial\widetilde{J}(w) + \epsilon A^2 w = 0$, with $\nu^2 = 1$ and $\beta = \gamma$,

$$\begin{aligned}
\frac{\lambda}{2\pi}\int_{-\pi}^{\pi} w\mathcal{C}w'dt = &-\left(1 + \frac{\lambda}{\pi}\int_{-\pi}^{\pi} w\mathcal{C}w'dt\right)\mathcal{C}w' + \lambda(w + w\mathcal{C}w' + \mathcal{C}(ww')) \\
&+ \gamma\left\{\frac{-w'}{\sqrt{w'^2 + (1+\mathcal{C}w')^2}} + \mathcal{C}\frac{1+\mathcal{C}w'}{\sqrt{w'^2 + (1+\mathcal{C}w')^2}}\right\}' + \epsilon w^{iv}
\end{aligned}$$

with $\epsilon > 0$, a standard regularity argument shows that $w \in H^4_{2\pi}$ and thus hypothesis (2.1e) is verified. Hypothesis (2.1f) is a consequence of the compact Sobolev embedding $X_2 \subset C^1_{2\pi}$.

THEOREM 4. *Let $\lambda \geq 0, \gamma_0 > 0$, $0 < R_2 < \sqrt{3/4\pi}$, and suppose that there exists $u_* \in U$ such that*

$$\widetilde{J}(u_*) < 0, \tag{3.10}$$

$$\varphi(s, R_2) < \gamma_0 C_3 R_2^2 \ \text{ for all } \ s \in \big[0\,,\, \sqrt{\mathcal{K}(u_*)/C_1}\,\big]. \tag{3.11}$$

Then there exists $w \in U\backslash\{0\}$ such that $\widetilde{J}(w) \leq \widetilde{J}(u_*)$ and (3.4) holds with $\nu^2 = 1$ and $\beta = \gamma \geq \gamma_0$. Hence (3.2) and (3.3) are satisfied for the same parameters if w is replaced by $\widetilde{w} := w - [w\mathcal{C}w']$.

Proof. This is a consequence of Theorem 1, since its hypotheses have been verified for this example. □

Now we must confirm the existence of u_* satisfying (3.10) and (3.11). The choice of u_* is motivated by a power-series expansion of solutions of the nonlinear problem which bifurcate from the trivial solution. From now on, $\lambda \in (0,1)$ is fixed, and we impose $\gamma_0 > 1 - \lambda$. For u_*, we try $u_*(t) = a\big(\cos t + k\cos 2t\big)$, where $a > 0$ and $k \in \mathbb{R}$ are small. Then $\|u_*\|^2_{L^2_{2\pi}} = \pi a^2(1 + 4k^2)$, $\|u_*''\|^2_{L^2_{2\pi}} = \pi a^2(1 + 16k^2)$, and

$$\int_{-\pi}^{\pi} u_* \mathcal{C}u_*' dt = \pi a^2(1 + 2k^2),$$

which yields

$$\widetilde{J}(u_*) = -\frac{\pi}{2}a^2\big\{1 + 2k^2 - \lambda(1 + k^2 + 2ak)\big\} - \frac{\lambda}{4\pi}\{\pi a^2(1 + 2k^2)\}^2 + \gamma_0\mathcal{K}(u_*).$$

To evaluate $\widetilde{J}(u_*)$, we choose $k = pa$, $\gamma_0 - 1 + \lambda = Ba^2$, where p, $B \in \mathbb{R}$ are yet to be determined, and consider only the terms of order at most 4 in a. Recall that, for $|s| < 1$,

$$\sqrt{1+s} = 1 + \frac{1}{2}s - \frac{1}{8}s^2 + \frac{1}{16}s^3 - \frac{5}{128}s^4 + \cdots.$$

Hence

$$\begin{aligned}
&\sqrt{{u_*'}^2 + (1 + \mathcal{C}u_*')^2} - (1 + \mathcal{C}u_*') \\
&= \sqrt{1 + 2\mathcal{C}u_*' + (\mathcal{C}u_*')^2 + (u_*')^2} - (1 + \mathcal{C}u_*') \\
&= \frac{1}{2}(u_*')^2 - \frac{1}{2}(u_*')^2\mathcal{C}u_*' - \frac{1}{8}(u_*')^4 + \frac{1}{2}(u_*')^2(\mathcal{C}u_*')^2 + \cdots \\
&= \frac{1}{2}a^2\left(\frac{1 - \cos 2t}{2} + 2k^2(1 - \cos 4t) + 2k(\cos t - \cos 3t)\right) \\
&\quad -\frac{1}{2}a^3\left(\frac{\cos t}{2} - \frac{\cos t + \cos 3t}{4} + 2k\sin^2 2t + k\sin 4t\right) \\
&\quad -\frac{1}{8}a^4\sin^2 t + \frac{5}{32}a^4\sin^2 2t + \cdots
\end{aligned}$$

and

$$\begin{aligned}\widetilde{J}(u_*) &= -\frac{\pi}{2}a^2\big\{1+2k^2-\lambda(1+k^2+2ak)\big\} - \frac{\lambda}{4\pi}\{\pi a^2(1+2k^2)\}^2 \\ &\quad +\pi\gamma_0\left\{\frac{a^2}{2}+2a^2k^2-a^3k+\frac{a^4}{32}\right\}+\cdots \\ &= \pi a^4\left\{\frac{B}{2}+\big(1-(3/2)\lambda\big)p^2+\big(2\lambda-1\big)p+(1-9\lambda)/32\big)\right\}+\cdots .\end{aligned}$$

Our aim is to get $\widetilde{\mathcal{J}}(u_*)<0$ for small enough $a>0$. If $\lambda\in[2/3,1)$, we can choose

$$p=-\frac{1-9\lambda}{32(2\lambda-1)}-1 \text{ and } B=2\lambda-1>0.$$

On the other hand, if $\lambda\in(0,2/3)$, then we can choose for p the value at which the following minimum is attained:

$$\min_{p\in\mathbb{R}}\Big\{\big(1-(3/2)\lambda\big)p^2+\big(2\lambda-1\big)p+(1-9\lambda)/32\Big\}=\frac{-74\lambda^2+86\lambda-28}{64(2-3\lambda)}<0.$$

Hence we can choose in this case

$$p=-\frac{2\lambda-1}{2-3\lambda} \text{ and } B=-\frac{-74\lambda^2+86\lambda-28}{64(2-3\lambda)}>0.$$

We can now check the other hypotheses of Theorem 4 for $\lambda>0$ fixed. If R_2 is fixed small enough, then $\varphi(s,R_2)\le R_2\,s$ and $C_1\ge 1/4$, $C_2=1/2$, $C_3\ge 1/2$. If $a>0$ is small, then $\|u_*\|_2^2=\pi a^2(1+16k^2)<R_2^2$. Since $\mathcal{K}(u_*)=\frac{\pi a^2}{2}+O(a^4)$ and $\gamma_0>1-\lambda$, it is clear that $R_2\sqrt{\mathcal{K}(u_*)/C_1}<\gamma_0 C_3R_2^2$ for a small, and (3.11) is satisfied.

To sum up, we have proved the following result that is a particular case of those in [10, 14, 17].

THEOREM 5. *Let λ and γ be given by* (3.5). *Then for all $\lambda\in(0,1)$ and all $\delta>0$, there exist $\gamma>1-\lambda$ and $w\in U\backslash\{0\}$ such that $0<\|w''\|_{L^2_{2\pi}}<\delta$ and* (3.4) *holds. Hence* (3.3) *and* (3.2) *with $\nu=1$ and $\beta=\gamma$ are satisfied if w is replaced by $\widetilde{w}:=w-[w\mathcal{C}w']$.*

By refining the method, it should be possible to obtain explicit lower bounds on the size of w for not too small δ. But the choice of our test function u_*, based on local arguments, is probably not optimal for this purpose. Note that in the case of pure capillary waves ($\lambda=0$), explicit large amplitude solutions have been obtained by Crapper [11]. In the general capillary-gravity case, this could help in the search for better test functions u_* more adapted to the global nature of the problem.

4. Stokes waves. We turn to the case of pure gravity waves. Divide by λ so that, in the abstract theory,

$$\gamma=\frac{\nu^2}{\lambda}=\frac{\pi c^2}{g\Lambda}. \tag{4.1}$$

In (3.2), (3.3), (3.4), J, and $\widetilde{J}$, let $\beta=0$ and $\lambda=1$. Having done so, $w\in H^1_{2\pi}$ in (3.2), (3.3) and we apply the abstract result of section 1 with $\gamma:=\nu^2$,

$$\begin{aligned}&X_0=\{w\in L^2_{2\pi}:[w]=0,\ w \text{ is even}\},\\ &Aw=\mathcal{C}w',\\ &X_k=\{w\in H^{k/2}_{2\pi}:[w]=0,\ w \text{ is even}\} \ \text{ for } \ k\in\{1,2,4\}.\end{aligned}$$

The radius $R_2 > 0$ will be specified later. For $w \in U$ (the ball in X_2 of radius R_2 centered at the origin) let

$$\mathcal{K}(w) = \int_{-\pi}^{\pi} w\mathcal{C}w'dt,$$
$$\mathcal{L}(w) = \frac{1}{2\pi}\left\{\int_{-\pi}^{\pi} w\mathcal{C}w'dt\right\}^2 - \int_{-\pi}^{\pi} w^2(1+\mathcal{C}w')dt.$$

Assumption (2.1a) is satisfied for $C_1 := 1$, and in (2.3b) we can take $C_3 := 2$, since

$$\partial\mathcal{K}(w)\mathcal{C}w' = 2\int_{-\pi}^{\pi} \mathcal{C}w'\mathcal{C}w'dt = 2\|w\|_2^2.$$

The following lemma from [21] is useful for finding the constant C_2 of (2.1b) and the function φ of (2.3c).

LEMMA 6. *If $w \in H^1_{2\pi}$ and if $h \in C^\infty(\mathbb{R})$ is convex on the range of w, then*

$$h'\big(w(t)\big)\mathcal{C}w'(t) - \mathcal{C}\big(h'(w)w'\big)(t) \geq 0$$

almost everywhere, and therefore

$$\int_{-\pi}^{\pi} h'\big(w(t)\big)\mathcal{C}w'(t)dt \geq 0.$$

Now

$$\int_{-\pi}^{\pi} w^2\mathcal{C}w'dt = \int_{-\pi}^{\pi} w\big(w\mathcal{C}w' - \mathcal{C}(ww')\big)dt + \frac{1}{2}\int_{-\pi}^{\pi} w^2\mathcal{C}w'dt,$$

which, with Lemma 6, gives

$$\left|\int_{-\pi}^{\pi} w^2\mathcal{C}w'dt\right| \leq 2\{\sup|w|\}\int_{-\pi}^{\pi}\{w\mathcal{C}w' - \mathcal{C}(ww')\}dt$$
$$= 2\{\sup|w|\}\int_{-\pi}^{\pi} w\mathcal{C}w'dt.$$

Thus

$$\mathcal{L}(w) \geq -(1+2\{\sup|w|\})\|w\|_1^2$$

and we can choose $C_2 := 1 + \sqrt{2\pi/3}R_2$ in (2.1b). We also have

$$\partial\mathcal{L}(w)\mathcal{C}w' = \frac{2}{\pi}\mathcal{K}(w)\int_{-\pi}^{\pi}\mathcal{C}w'\mathcal{C}w'dt - 2\int_{-\pi}^{\pi} w\mathcal{C}w'dt$$
$$- 2\int_{-\pi}^{\pi} w\mathcal{C}w'\mathcal{C}w'dt - \int_{-\pi}^{\pi}\mathcal{C}(w^2)'\mathcal{C}w'dt$$
$$\geq \frac{2}{\pi}\mathcal{K}(w)\int_{-\pi}^{\pi}\mathcal{C}w'\mathcal{C}w'dt - 2\mathcal{K}(w)$$
$$- 2\{\sup|w|\}\int_{-\pi}^{\pi}(\mathcal{C}w')^2 + (w')^2dt$$
$$\geq -\varphi(\|w\|_1, \|w\|_2) - C_4\|w\|_2^2$$

with $\varphi(s,t) := \frac{2}{\pi}s^2t^2 - 2s$ and $C_4 := 4\sqrt{\pi/6}R_2$. This gives (2.3c).

Now $A^2w = -w''$ and so, if $w \in H^1_{2\pi}$ is a weak solution of

$$-\frac{1}{2\pi}\int_{-\pi}^{\pi} w\mathcal{C}w'dt = \left(\nu^2 + \int_{-\pi}^{\pi} w\mathcal{C}w'dt/\pi\right)\mathcal{C}w' - w - w\mathcal{C}w' - \mathcal{C}(ww') + \epsilon A^2 w$$

with $\epsilon > 0$, a standard regularity argument shows that $w \in W^{2,2}_{2\pi}$, and thus hypothesis (2.1e) is verified.

To check assumption (2.1f), note that if $w_n \rightharpoonup w$ weakly in $L^2_{2\pi}$, then $\mathcal{C}w_n' \rightharpoonup \mathcal{C}w'$ weakly in $H^1_{2\pi}$ and $w_n \to w$ uniformly on $[-\pi,\pi]$. It follows that $\mathcal{K}(w_n) \to \mathcal{K}(w)$, $\mathcal{L}(w_n) \to \mathcal{L}(w)$.

THEOREM 7. *Let $\gamma > 0$ and $R_2 < \gamma\sqrt{3/(2\pi)}$. Assume that there exists $u_* \in U$ such that*

$$\gamma\mathcal{K}(u_*) + \mathcal{L}(u_*) < 0$$

and

$$\frac{\mathcal{K}(u_*)}{\gamma - \sqrt{2\pi/3}R_2 + \pi^{-1}\mathcal{K}(u_*)} < R_2^2. \tag{4.2}$$

Then there exist $w \in U\backslash\{0\}$ and $\widetilde{\gamma} \geq \gamma$ such that

$$\widetilde{\gamma}\partial\mathcal{K}(w) + \partial\mathcal{L}(w) = 0 \text{ and } \gamma\mathcal{K}(w) + \mathcal{L}(w) \leq \gamma\mathcal{K}(u_*) + \mathcal{L}(u_*).$$

Proof. This theorem follows from Theorem 1 and the particular form of ψ. Indeed, the inequality (4.2) is equivalent to (2.3a). □

Let $u_*(t) = a\big(\cos t + k\cos 2t\big)$, where $a,\ k > 0$. Then $\|u\|^2_{L^2_{2\pi}} = \pi a^2(1+4k^2)$,

$$\int_{-\pi}^{\pi} u_*\mathcal{C}u_*'dt = \pi a^2(1+2k^2), \quad \text{and} \quad \int_{-\pi}^{\pi} u_*^2\mathcal{C}u_*'dt = 2\pi a^3 k.$$

Moreover, setting $r = \sqrt{\pi a^2(1+2k^2)}$, we get

$$\nu^2\mathcal{K}(u_*) + \mathcal{L}(u_*) = \pi a^2\left\{\nu^2 - 1 + (2\nu^2-1)k^2 - \frac{2rk}{\sqrt{\pi(1+2k^2)}} + \frac{r^2(1+2k^2)}{2\pi}\right\}.$$

All hypotheses of Theorem 7 are verified if

$$r\frac{\sqrt{1+4k^2}}{\sqrt{1+2k^2}} < R_2 < \frac{\sqrt{3}\,\gamma}{\sqrt{2\pi}}, \tag{4.3}$$

$$R_2^2 > \frac{r^2}{\gamma - \sqrt{2\pi/3}R_2 + r^2/\pi},$$

$$\gamma - 1 + (2\gamma-1)k^2 - \frac{2rk}{\sqrt{\pi(1+2k^2)}} + \frac{r^2(1+2k^2)}{2\pi} \;<\; 0. \tag{4.4}$$

Then Theorem 4 provides us with a nontrivial solution w. Take $R_2 = 0.477$, $r = 0.28$, $k = 0.142$, $\gamma = 1/0.99$. Conditions (4.3) to (4.4) above are fulfilled, and we get a solution of (3.2) with $\beta = 0$ and $\lambda = 1$ in which w and ν are replaced by

$$w_* = w - \frac{1}{2\pi} \int_{-\pi}^{\pi} w\mathcal{C}w' dt$$

and $\nu_* \geq 0.99^{-1/2}$. Alternatively, we can fix $\nu = 1$ and let λ be the parameter, in which case the corresponding λ_* is in $(0, 0.99]$. The same result has been obtained in [8] via the mountain-pass theorem, but the present proof is simpler.

Acknowledgment. The authors wish to thank the referees for valuable comments and suggestions.

REFERENCES

[1] K. I. Babenko, *On a local existence theorem in the theory of surface waves of finite amplitude*, Soviet Math. Dokl., 35 (1987), pp. 647–650.
[2] K. I. Babenko, *Some remarks on the theory of surface waves of finite amplitude*, Soviet Math. Dokl., 35 (1987), pp. 599–603.
[3] K. I. Babenko, V. Yu. Petrovich, and A. I. Rakhmanov, *A computational experiment in the theory of surface waves of finite amplitude*, Soviet Math. Dokl., 38 (1989), pp. 327–331.
[4] K. I. Babenko, V. Yu. Petrovich, and A. I. Rakhmanov, *On a demonstrative experiment in the theory of surface waves of finite amplitude*, Soviet Math. Dokl., 38 (1989), pp. 626–630.
[5] C. Baesens and R. S. MacKay, *Uniformly travelling water waves from a dynamical systems viewpoint: Some insight into bifurcations from Stokes' family*, J. Fluid Mech., 241 (1992), pp. 333–347.
[6] B. Buffoni, E. N. Dancer, and J. F. Toland, *The regularity and local bifurcation of steady periodic water waves*, Arch. Ration. Mech. Anal., 152 (2000), pp. 207–240.
[7] B. Buffoni, E. N. Dancer, and J. F. Toland, *The sub-harmonic bifurcation of Stokes waves*, Arch. Ration. Mech. Anal., 152 (2000), pp. 241–271.
[8] B. Buffoni, É. Séré, and J. F. Toland, *Surface water waves as saddle points of the energy*, Calc. Var. Partial Differential Equations, 17 (2003), pp. 199–220.
[9] B. Chen and P. G. Saffman, *Numerical evidence for the existence of new types of gravity waves of permanent form on deep water*, Stud. Appl. Math., 62 (1980), pp. 1–21.
[10] W. Craig and D. P. Nicholls, *Travelling two and three dimensional capillary gravity water waves*, SIAM J. Math. Anal., 32 (2000), pp. 323–359.
[11] G. D. Crapper, *An exact solution for progressive capillary waves of arbitrary amplitude*, J. Fluid Mech., 2 (1957), pp. 532–540.
[12] F. Dias and C. Kharif, *Nonlinear gravity and capillary-gravity waves*, in Annual Review of Fluid Mechanics, Annu. Rev. Fluid Mech. 31, Annual Reviews, Palo Alto, CA, 1999, pp. 301–346.
[13] J. L. Hammack and D. M. Henderson, *Resonant interactions among surface water waves*, in Annual Review of Fluid Mechanics, Annu. Rev. Fluid Mech. 25, Annual Reviews, Palo Alto, CA, 1993, pp. 55–97.
[14] M. C. W. Jones and J. F. Toland, *Symmetry and the bifurcation of capillary-gravity waves*, Arch. Ration. Mech. Anal., 96 (1986), pp. 29–53.
[15] M. A. Krasnosel'skii, *Topological Methods for Nonlinear Eigenvalue Problems*, Pergamon Press, Oxford, UK, 1964.
[16] M. S. Longuet-Higgins, *Bifurcation in gravity waves*, J. Fluid Mech., 151 (1985), pp. 457–475.
[17] J. Reeder and M. Shinbrot, *On Wilton ripples,* I: *Formal derivation of the phenomenon*, Wave Motion, 3 (1981), pp. 115–135.
[18] L. W. Schwartz and J. D. Fenton, *Strongly nonlinear waves*, in Annual Review of Fluid Mechanics, Annu. Rev. Fluid Mech. 14, Annual Reviews, Palo Alto, CA, 1982, pp. 39–60.
[19] G. G. Stokes, *On the theory of oscillatory waves*, Trans. Camb. Phil. Soc., 8 (1847), pp. 441–455.
[20] C. A. Stuart, *An introduction to bifurcation theory based on differential calculus*, Nonlinear Mathematics and Mechanics: Heriot-Watt Symposium, Vol. IV, R. J. Knops, ed., Res. Notes in Math. 39, Pitman, Boston, 1979, pp. 76–132.

[21] J. F. Toland, *Continuity and differentiability of Nemytskii operators on the Hardy space* $H^{1,1}(T^1)$, Ark. Mat., 39 (2001), pp. 383–394.
[22] R. E. L. Turner, *A variational approach to surface solitary waves*, J. Differential Equations, 55 (1984), pp. 401–438.
[23] R. W. Yeung, *Numerical methods in free-surface flows*, in Annual Review of Fluid Mechanics, Annu. Rev. Fluid Mech. 14, Annual Reviews, Palo Alto, CA, 1982, pp. 395–442.
[24] A. Zygmund, *Trigonometric Series* I & II, Cambridge University Press, Cambridge, UK, 1959.

SIAM J. MATH. ANAL.
Vol. 36, No. 4, pp. 1095–1129

EXPLICIT SOLUTIONS OF THE EIGENVALUE PROBLEM $-\mathbf{div}\left(\frac{D}{|D|}\right) = u$ IN R^2 *

GIOVANNI BELLETTINI†, VICENT CASELLES‡, AND MATTEO NOVAGA§

Abstract. In this paper we compute explicit solutions of the eigenvalue problem $-\mathrm{div}(Du/|Du|) = u$ in R^2, in particular explicit solutions whose truncatures are in $W^{1,1}_{\mathrm{loc}}(R^2)$, and piecewise constant ones which are sums of characteristic functions of convex sets. The solutions of the above eigenvalue problem describe the asymptotic behavior of solutions of the minimizing total variation flow. As an application, we also construct explicit solutions of the denoising problem in image processing.

Key words. eigenvalue problem, total variation flow, finite perimeter sets, denoising problem

AMS subject classifications. 35J70, 35P30, 35K65

DOI. 10.1137/S0036141003430007

1. Introduction. The main aim of this paper is to compute explicit solutions of the following eigenvalue problem:

$$-\mathrm{div}\left(\frac{Du}{|Du|}\right) = u, \quad u \in L^1_{\mathrm{loc}}(R^2). \tag{1.1}$$

Solutions to (1.1) describe the asymptotic behavior, as $t \to +\infty$, of solutions of the minimizing total variation flow in R^2 given by the equation

$$\frac{\partial u}{\partial t} = \mathrm{div}\left(\frac{Du}{|Du|}\right) \quad \text{in } Q_T := \,]0,T[\times R^2, \tag{1.2}$$

coupled with the initial condition

$$u(0) = u_0 \in L^2(R^2). \tag{1.3}$$

Indeed, as was proved in [9], if $u_0 \in L^2(R^2)$, then the solution $u(t)$ vanishes in finite time $T(u_0)$ and the rescaled function $\frac{u(t)}{T(u_0)-t}$ converges along subsequences to a solution of (1.1) as $t \to T(u_0)$; see Theorem 2.8 below. Thus, solutions of (1.1) describe the profiles of extinction of solutions of (1.2). We also notice that a solution u of (1.1) allows us to construct a solution of (1.2) of the form $v(t,x) = (1-t)^+ u(x)$.

One of the main motivations of our study comes from the total variation approach to the problems of image denoising and restoration. Indeed, as was shown in [12], solutions of (1.1) allow us to construct explicit solutions of the total variation formulation of the denoising problem [33]. Assuming that our observed image (or data) f

*Received by the editors June 13, 2003; accepted for publication (in revised form) May 14, 2004; published electronically January 27, 2005.
http://www.siam.org/journals/sima/36-4/43000.html

†Dipartimento di Matematica, Università di Roma "Tor Vergata," via della Ricerca Scientifica, 00133 Roma, Italy (belletti@mat.uniroma2.it).

‡Dept. de Tecnologia, University of Pompeu-Fabra, Passeig de Circumvalacio, 8, 08003 Barcelona, Spain (vicent.caselles@upf.edu). This author was supported by the Departament d'Universitats, Recerca i Societat de la Informació de la Generalitat de Catalunya, and by PNPGC project BFM2000-0962-C02-01.

§Dipartimento di Matematica, Università di Pisa, via Buonarroti 2, 56127 Pisa, Italy (novaga@dm.unipi.it).

comes from noisy observations of an ideal undistorted image u, the image model can be written as

$$f = u + n, \tag{1.4}$$

where n represents the noise, typically assumed to be Gaussian. In [33], Rudin, Osher, and Fatemi proposed obtaining the denoised image u by solving the constrained minimization problem

$$\text{Minimize} \int_D |Du| \quad \text{with} \int_D (u-f)^2\,dx = \sigma^2|D|, \tag{1.5}$$

where D is the image domain, typically a rectangle in R^2, and the constraint incorporates the image acquisition model given by (1.4) in terms of the variance of the noise σ^2. Let us stress here that even if three-dimensional images occur, for instance, a medical image (or video data), the case of R^2, being the case of photographs and satellite or medical images, plays an important role in image processing. In practice, problem (1.5) is solved via the unconstrained minimization problem

$$\min\left\{\int_D |Du| + \frac{1}{2\lambda}\int_D (u-f)^2\,dx : u \in BV(D)\right\} \tag{1.6}$$

for some Lagrange multiplier $\lambda > 0$ [17]. The constraint has been introduced as a penalization term. The regularization parameter λ controls the trade-off between the goodness of fit of the constraint and the smoothness term given by the total variation. This formulation of the denoising problem pioneered the use of total variation as a regularization term and the use of bounded variation functions in image processing. The first regularization methods used the Sobolev (semi)norm $\int_D |Du|^2$ and proposed denoising the data f by solving

$$\min\left\{\int_D |Du|^2 + \frac{1}{2\lambda}\int_D (u-f)^2\,dx : u \in W^{1,2}(D)\right\}. \tag{1.7}$$

In case $D = R^2$, the solution of (1.7) in the Fourier domain is given by

$$\hat{u}(\xi) = \frac{\hat{f}(\xi)}{1 + 4\gamma\pi^2|\xi|^2}, \quad \xi \in R^2$$

(the constants appearing in the denominator being dependent on the form of the Fourier transform). From the above formula we see that high frequencies of f (hence, the noise) are attenuated by the smoothness constraint. This was an important step, but the results were not satisfactory, mainly due to the inability of the previous functional to resolve discontinuities (edges) and oscillatory textured patterns. The smoothness constraint is too restrictive. Indeed, functions in $W^{1,2}(D)$ cannot have discontinuities along rectifiable curves. These observations motivated the introduction of total variation in image restoration models by Rudin, Osher, and Fatemi in their seminal work [33]. The a priori hypothesis is that functions of bounded variation (the BV model) [6, 24, 36] are a reasonable functional model for many problems in image processing, in particular, for restoration problems [33]. Typically, functions of bounded variation have discontinuities along rectifiable curves, being continuous in the measure theoretic sense away from discontinuities. The discontinuities could be identified with edges. The ability of this functional to describe textures is less clear;

some textures can be recovered, but up to a certain scale of oscillation. An interesting experimental discussion of the adequacy of the BV model to describe real images can be seen in [3, 29].

The analysis of problem (1.5) has been the subject of much work in the last ten years, both numerical and theoretical. It will not be our purpose to review it here, and we refer the interested reader to [10] for an account of it. Let us mention only that the existence of solutions of (1.5) for any $f \in L^2(R^N)$ follows easily from the convexity of the functional and the properties of bounded variation functions; that the equivalence between (1.5) and (1.6) was proved in [17]; and that the characterization of the Euler–Lagrange equation in distributional terms was done in [8, 12] (see [10]). To describe the behavior of solutions of (1.6), the authors started in [12] the search for explicit solutions for some particular kind of functions $f \in L^2(R^2)$. Since, when $\lambda = \Delta t$, (1.6) corresponds to the implicit in time discretization of (1.2) (also called the Crandall–Liggett scheme in semigroup theory [19]), the behavior of solutions of one of them is analogous to those of the other. This has been exploited in the papers [8, 12] (see also [10] for a full account).

In particular, in [12] we showed how the explicit solutions of (1.1) could be used to construct data $f \in L^2(R^2)$ for which we could compute the explicit solution of (1.6) in R^2. In the most simple case, if $\overline{u} \in BV(R^2)$ is a solution of (1.1) and $b \in R$, then the function $a\overline{u}$ with $a = \operatorname{sign}(b)(|b| - \lambda)^+$ is the solution of the variational problem (1.6) when $f = b\overline{u}$. In other words, the solution of (1.6) is given by the soft-thresholding rule applied to b. Other more general results were also exhibited. In particular, this established a connection with the wavelet approach to denoising given by the soft-thresholding rule applied to the wavelet coefficients of a noisy function (the uncorrupted function being in some Besov space) [20, 21, 22, 23]. In this direction, let us recall the result of Meyer [31], which proves that by applying a soft-thresholding to the coefficients of the wavelet expansion of f with respect to some orthonormal wavelet basis, one obtains a quasi-optimal solution of (1.6) in the sense that its energy is bounded by a universal constant times the actual minimum energy. Further work exploring the connection between both approaches, variational and wavelet-based, to the denoising problem can be found in [35].

Our purpose in this paper will be to make progress in the study of the solutions of the eigenvalue problem (1.1) and to derive, as a consequence, other explicit solutions of the denoising problem

$$\min\left\{\int_{R^2} |Du| + \frac{1}{2\lambda}\int_{R^2} (u-f)^2\,dx : u \in BV(R^2)\right\} \tag{1.8}$$

for some data $f \in L^2(R^2)$, $\lambda > 0$. For that, in section 2 we shall begin by recalling some preliminary facts about functions of bounded variation, a generalized Green's formula [11], and the notion of solution for the evolution equation (1.2) and for the eigenvalue problem (1.1).

In section 3 we describe the regularity properties of the level lines of the solutions of (1.1). In section 4 we study the solutions of (1.1) which are in $W^{1,1}(R^2)$, and hence do not possess discontinuities along rectifiable curves. Indeed, we compute the explicit solutions u of (1.1) whose truncatures $T_k(u) := (-k) \vee u \wedge k$ are in $W^{1,1}_{\mathrm{loc}}(R^2)$ for any $k > 0$, and we prove that the level sets $\{u > t\}$, $t > 0$ (resp., $\{u < t\}$, $t < 0$), of the nonzero solutions are balls of radius $\frac{1}{t}$ (resp., $-\frac{1}{t}$).

Then we turn our attention to the consideration of piecewise constant solutions of (1.1) which can be described as sums of characteristic functions of convex sets forming

towers (or oscillating towers). As we shall prove, there are geometric restrictions on the curvature of the convex sets, as well as restrictions on their relative position to be able to combine them in towers which are solutions of (1.1). In some particular cases, this kind of geometric condition already appeared in the study of capillarity problems in domains of R^2 [18, 25, 26, 27, 28], and its analogues have also appeared in the case of crystalline variational problems [13, 14, 15]. Let us mention that consideration of convex sets is justified by the results in [12]. The analysis of piecewise constant solutions of (1.1) leads to the study of solutions of $\operatorname{div} z =$ constant in bounded and unbounded domains delimited by convex sets. Section 5 is devoted to solving the equation $\operatorname{div} z =$ constant in a bounded domain F of R^2 determined by an exterior Jordan curve ∂C_0 of class $\mathcal{C}^{1,1}$ and a finite number m of interior Jordan curves, also of class $\mathcal{C}^{1,1}$, where the unknown is a vector field $z \in L^\infty(F; R^2)$, $\|z\|_\infty \leq 1$, whose trace at the boundary is the inner or outer unit normal, depending on the Jordan curve. This is one of the basic building blocks in constructing piecewise constant explicit solutions of (1.1), the other being the solution of the equation $\operatorname{div} z = 0$ in the complement of a bounded domain made by a finite number of connected components whose boundary is a convex curve of class $\mathcal{C}^{1,1}$. This will be the purpose of section 6. By pasting together these solutions one can construct explicit piecewise constant solutions of (1.1). We shall call these solutions oscillating tower solutions of (1.1). We shall use them to construct some data $f \in L^2(R^2)$ for which the explicit solutions of (1.8) can be computed (with a soft-thresholding rule). This will be the purpose of section 8.

The solutions constructed here illustrate the behavior of solutions of (1.8), but do not exhibit all its features. The behavior of (1.8) for characteristic functions of general convex sets in R^2 (together with explicit solutions of (1.2)) was described in [1], where it was shown that the sets are eroded at high curvature points of its boundary. By the way, the extension of the above results to characteristic functions of convex sets in R^N has been started in [2]. The explicit behavior of (1.8) and (1.2) when the initial condition is the characteristic function of a general set in R^2 with smooth or piecewise smooth boundary is still to be described. We believe that with these elements on hand, one would be able to add them and produce a description of a more general class of piecewise constant solutions of (1.2). There is still a long way to go, but our explicit solutions are a first step in this direction and illustrate the behavior of soft-thresholding in some geometrically simple cases, exhibiting the role of the parameter λ in the elimination of small localized perturbations of the image (which could be assimilated to a multiple of a characteristic function of some small ball).

2. Some notation.

2.1. Functions of bounded variation and sets of finite perimeter. Let Q be an open subset of R^N. By $C_0^\infty(Q)$ (resp., $C_0^\infty(Q; R^N)$) we denote the space of functions (resp., vector fields with values in R^N) which are C^∞ and have compact support in Q.

A function $u \in L^1(Q)$ whose gradient Du in the sense of distributions is a (vector-valued) Radon measure with finite total variation in Q is called a function of bounded variation. The class of such functions will be denoted by $BV(Q)$. The total variation of Du on Q turns out to be

$$\sup\left\{\int_Q u \operatorname{div} z \, dx : z \in C_0^\infty(Q; R^N), \|z\|_{L^\infty(Q)} := \operatorname*{ess\,sup}_{x \in Q} |z(x)| \leq 1\right\} \tag{2.1}$$

(where for a vector $v = (v_1, \dots, v_N) \in R^N$ we set $|v|^2 := \sum_{i=1}^N v_i^2$) and will be denoted by $|Du|(Q)$ or by $\int_Q |Du|$. It turns out that the map $u \to |Du|(Q)$ is $L^1_{\rm loc}(Q)$-lower semicontinuous. $BV(Q)$ is a Banach space when endowed with the norm $\int_Q |u|\, dx + |Du|(Q)$. We recall that $BV(R^N) \subseteq L^{N/(N-1)}(R^N)$. The total variation of u on a Borel set $B \subseteq Q$ is defined as $\inf\{|Du|(A)\colon A \text{ open}, B \subseteq A \subseteq Q\}$. We denote by $BV_{\rm loc}(Q)$ the space of functions $w \in L^1_{\rm loc}(Q)$ such that $w\varphi \in BV(Q)$ for all $\varphi \in C_0^\infty(Q)$. For results and information on functions of bounded variation, we refer to [6, 24].

A measurable set $E \subseteq R^N$ is said to be of finite perimeter in Q if (2.1) is finite when u is substituted with the characteristic function χ_E of E. The perimeter of E in Q is defined as $P(E,Q) := |D\chi_E|(Q)$, and $P(E,Q) = P(R^N \setminus E, Q)$. We shall use the notation $P(E) := P(E, R^N)$. For sets of finite perimeter E one can define the essential boundary $\partial^* E$, which is countably $(N-1)$ rectifiable with finite $\mathcal{H}^{N-1}$ measure, and compute the outer unit normal $\nu^E(x)$ at $\mathcal{H}^{N-1}$ almost all points x of $\partial^* E$, where $\mathcal{H}^{N-1}$ is the $(N-1)$-dimensional Hausdorff measure. Moreover, $|D\chi_E|$ coincides with the restriction of $\mathcal{H}^{N-1}$ to $\partial^* E$.

For a Lebesgue measurable subset $E \subseteq R^N$ and a point $x \in R^N$, the upper and lower densities of E at x are, respectively, defined by

$$\overline{D}(x,E) := \limsup_{r \to 0^+} \frac{|E \cap B_r(x)|}{|B_r(x)|}, \quad \underline{D}(x,E) := \liminf_{r \to 0^+} \frac{|E \cap B_r(x)|}{|B_r(x)|}.$$

Here $B_r(x)$ denotes the open ball of radius r centered at x and $|\cdot|$ stands for the Lebesgue measure. If the upper and lower densities are equal, their common value will be called the density of E at x, and it will be denoted by $D(x,E)$. Each set E of finite perimeter will be identified with the representative (in its Lebesgue class) given by the set of all points $x \in R^N$ such that $D(x,E) = 1$. It is clear that if ∂E is Lipschitz continuous, then the precise representative we are choosing is an open set.

If μ is a (possibly vector-valued) Radon measure and f is a Borel function, the integration of f with respect to μ will be denoted by $\int f d\mu$. When μ is the Lebesgue measure, the symbol dx will be often omitted.

By $L^1_w(]0,T[; BV(R^N))$ we denote the space of functions $v :]0,T[\to BV(R^N)$ such that $v \in L^1\left(]0,T[\times R^N\right)$, the maps $t \in]0,T[\to \int_{R^N} \phi\, dDv(t)$ are measurable for every $\phi \in C_0^1(R^N; R^N)$, and $\int_0^T |Dv(t)|(R^N)\, dt < \infty$. By $L^1_w(]0,T[; BV_{\rm loc}(R^N))$ we denote the space of functions $v :]0,T[\to BV_{\rm loc}(R^N)$ such that $v\varphi \in L^1_w(]0,T[; BV(R^N))$ for all $\varphi \in C_0^\infty(R^N)$.

If E is a subset of R^N of class $\mathcal{C}^{1,1}$, we denote by $\kappa_{\partial E}$ the ($\mathcal{H}^{N-1}$-almost everywhere defined) curvature of ∂E, nonnegative for convex sets. The following result can be proved as in [32].

PROPOSITION 2.1. *Let $\mu \in R$ and E be a set of class $\mathcal{C}^{1,1}$. Assume that there exists an open set A such that $A \cap \partial E$ is the graph of a $\mathcal{C}^{1,1}$ function, and*

$$P(E,A) - \mu|E \cap A| \le P(E \cup B, A) - \mu|(E \cup B) \cap A| \tag{2.2}$$

for any bounded measurable set B with $\overline{B} \subset A$. Then $\kappa_{\partial E}(x) \ge \mu$ for $\mathcal{H}^{N-1}$-almost every $x \in A \cap \partial E$. Similarly, if in place of (2.2) *there holds the inequality*

$$P(E,A) - \mu|E \cap A| \le P(E \setminus B, A) - \mu|(E \setminus B) \cap A|,$$

then $\kappa_{\partial E}(x) \le \mu$ for $\mathcal{H}^{N-1}$-almost every $x \in A \cap \partial E$.

The following lemma will be used in several places. Let us include its proof for the sake of completeness.

LEMMA 2.2. *Let $A, B \subseteq R^N$ be two sets of finite perimeter such that $|A \cap B| = 0$. Then, up to a set of $\mathcal{H}^{N-1}$-measure zero, we have*

$$\partial^*(A \cup B) = (\partial^* A \setminus \partial^* B) \cup (\partial^* B \setminus \partial^* A).$$

In particular, we have

$$P(A \cup B) = P(A) + P(B) - 2\mathcal{H}^{N-1}(\partial^* A \cap \partial^* B).$$

Proof. Recall that if $E \subseteq R^N$ has finite perimeter, the essential boundary $\partial^* E$ is contained in the measure theoretic boundary $\partial^M E$ (i.e., the set of points $x \in R^N$ such that $\overline{D}(x, E) > 0$ and $\overline{D}(x, R^N \setminus E) > 0$) of E, and $\mathcal{H}^{N-1}(\partial^M E \setminus \partial^* E) = 0$ [6, 24, 36]. Let $p \in \partial^*(A \cup B)$. Then $\overline{D}(p, A \cup B) > 0$ and $\overline{D}(p, R^N \setminus (A \cup B)) > 0$. Since $R^N \setminus (A \cup B) \subseteq R^N \setminus A$ we have $\overline{D}(p, R^N \setminus A) > 0$. Similarly $\overline{D}(p, R^N \setminus B) > 0$. From $\overline{D}(p, A \cup B) > 0$, we have either $\overline{D}(p, A) > 0$ or $\overline{D}(p, B) > 0$. If $\overline{D}(p, A) > 0$ (resp., $\overline{D}(p, B) > 0$), we have $p \in \partial^* A$ (resp., $p \in \partial^* B$). Now, if $p \in \partial^* A \cap \partial^* B$, $\mathcal{H}^{N-1}$-almost everywhere, we have $D(p, A) = D(p, R^N \setminus A) = D(p, B) = D(p, R^N \setminus B) = \frac{1}{2}$. Since $|A \cap B| = 0$, we conclude

$$D(p, A \cup B) = D(p, A) + D(p, B) = \frac{1}{2} + \frac{1}{2} = 1.$$

This implies that $p \notin \partial^*(A \cup B)$, a contradiction. We conclude that $p \notin \partial^* A \cap \partial^* B$. We have proved that

$$\partial^*(A \cup B) \subseteq (\partial^* A \setminus \partial^* B) \cup (\partial^* B \setminus \partial^* A) \pmod{\mathcal{H}^{N-1}}.$$

To prove the opposite inclusion, assume that $p \in \partial^* A \setminus \partial^* B$. Then for $\mathcal{H}^{N-1}$-almost every p we may assume that

$$D(p, A) = D(p, R^N \setminus A) = \frac{1}{2}. \tag{2.3}$$

In particular, we have that $\overline{D}(p, A \cup B) > 0$. Assume that $\overline{D}(p, R^N \setminus (A \cup B)) = 0$. In this case, $D(p, A \cup B) = 1$. Using (2.3), we obtain $D(p, B) = \frac{1}{2}$. Hence, $p \in \partial^* B$, a contradiction. Thus, we also have $\overline{D}(p, R^N \setminus (A \cup B)) > 0$, and therefore $p \in \partial^*(A \cup B)$ for $\mathcal{H}^{N-1}$-almost every $p \in \partial^* A \setminus \partial^* B$. We conclude that $\partial^* A \setminus \partial^* B \subseteq \partial^*(A \cup B)$. Similarly we have that $\partial^* B \setminus \partial^* A \subseteq \partial^*(A \cup B)$. ∎

2.2. A generalized Green's formula. Let Ω be an open set in R^N. Following [11], let

$$X_2(\Omega) := \{z \in L^\infty(\Omega; R^N) : \operatorname{div} z \in L^2(\Omega)\},$$
$$X_{2,\mathrm{loc}}(\Omega) := \{z \in L^\infty(\Omega; R^N) : \operatorname{div} z \in L^2_{\mathrm{loc}}(\Omega)\}.$$

If $z \in X_{2,\mathrm{loc}}(\Omega)$ and $w \in L^2_{\mathrm{loc}}(\Omega) \cap BV_{\mathrm{loc}}(\Omega)$, we define the functional (z, Dw): $C_0^\infty(\Omega) \to R$ by the formula

$$\langle (z, Dw), \varphi \rangle := -\int_\Omega w\, \varphi \operatorname{div} z\, dx - \int_\Omega w\, z \cdot \nabla \varphi\, dx \qquad \forall \varphi \in C_0^\infty(\Omega). \tag{2.4}$$

Notice that

$$\langle (z, Dw), \varphi \rangle = \int_\Omega z \cdot \nabla w \, \varphi \, dx \qquad \forall w \in L^2_{\rm loc}(\Omega) \cap W^{1,1}_{\rm loc}(\Omega).$$

If $z \in X_2(\Omega)$ and $w \in L^2(\Omega) \cap BV(\Omega)$, then (z, Dw) is a Radon measure in Ω, and

$$\left| \int_B (z, Dw) \right| \le \int_B |(z, Dw)| \le \|z\|_\infty \int_B |Dw| \qquad \forall \text{ Borel set } B \subseteq \Omega.$$

We denote by $\theta(z, Dw) \in L^\infty_{|Dw|}(\Omega)$ the density of (z, Dw) with respect to $|Dw|$, that is,

$$(z, Dw)(B) = \int_B \theta(z, Dw) \, d|Dw| \qquad \forall \text{ Borel set } B \subseteq \Omega. \tag{2.5}$$

If $\Omega = R^N$, we have the following integration-by-parts formula [11] for $z \in X_2(R^N)$ and $w \in L^2(R^N) \cap BV(R^N)$:

$$\int_{R^N} w \, \text{div } z \, dx + \int_{R^N} (z, Dw) = 0. \tag{2.6}$$

In particular, if B is bounded and has finite perimeter in R^N, from (2.6) and (2.5) it follows that

$$\int_B \text{div } z \, dx = \int_{R^N} (z, -D\chi_B) = \int_{\partial^* B} \theta(z, -D\chi_B) \, d\mathcal{H}^{N-1}. \tag{2.7}$$

Notice also that if $z_1, z_2 \in X_2(R^N)$ and $z_1 = z_2$ almost everywhere on B, then $\theta(z_1, -D\chi_B)(x) = \theta(z_2, -D\chi_B)(x)$ for $\mathcal{H}^{N-1}$-almost every $x \in \partial^* B$.

We recall the following result proved in [11].

THEOREM 2.3. *Let $\Omega \subset R^N$ be a open set with Lipschitz boundary, $1 \le p \le N$, $p' = \frac{p}{p-1}$. Assume that either Ω or $R^N \setminus \overline{\Omega}$ is bounded. Let $u \in BV(\Omega) \cap L^{p'}(\Omega)$ and $z \in L^\infty(\Omega; R^N)$ with* div $z \in L^p(\Omega)$. *Then, using test functions $\varphi \in C^\infty_0(\Omega)$, (2.4) defines a Radon measure (z, Du) in Ω, there exists a function $[z \cdot \nu^\Omega] \in L^\infty(\partial\Omega)$ such that $\|[z \cdot \nu^\Omega]\|_{L^\infty(\partial\Omega)} \le \|z\|_{L^\infty(\Omega; R^N)}$, and*

$$\int_\Omega u \, \text{div} z \, dx + \int_\Omega (z, Du) = \int_{\partial\Omega} [z \cdot \nu^\Omega] u \, d\mathcal{H}^{N-1}.$$

In particular, if Ω or $R^N \setminus \overline{\Omega}$ is a bounded open set with Lipschitz boundary, then (2.7) has a meaning also if z is defined only on Ω and not on the whole of R^N, precisely when $z \in L^\infty(\Omega; R^N)$ with div $z \in L^1(\Omega)$. In this case we mean that $\theta(z, -D\chi_\Omega)$ coincides with $[z \cdot \nu^\Omega]$.

Remark 1. Let $\Omega \subset R^2$ be a bounded Lipschitz open set, and let $z_{\rm inn} \in L^\infty(\Omega; R^2)$ with $\text{div} z_{\rm inn} \in L^2(\Omega)$, and $z_{\rm out} \in L^\infty(R^2 \setminus \overline{\Omega}; R^2)$ with $\text{div} z_{\rm out} \in L^2(R^2 \setminus \overline{\Omega})$. Assume that

$$[z_{\rm inn} \cdot \nu^\Omega](x) = -[z_{\rm out} \cdot \nu^{R^2 \setminus \overline{\Omega}}](x) \qquad \text{for } \mathcal{H}^1\text{-a.e. } x \in \partial\Omega.$$

Then if we define $z := z_{\rm inn}$ on Ω and $z := z_{\rm out}$ on $R^2 \setminus \overline{\Omega}$, we have $z \in L^\infty(R^2; R^2)$ and $\text{div} z \in L^2(R^2)$.

2.3. The notion of solution, and existence and uniqueness results. Consider the energy functional $\Psi : L^2(R^N) \to (-\infty, +\infty]$ defined by

$$\Psi(u) := \begin{cases} \int_{R^N} |Du| & \text{if} \quad u \in L^2(R^N) \cap BV(R^N), \\ +\infty & \text{if} \quad u \in L^2(R^N) \setminus BV(R^N). \end{cases} \tag{2.8}$$

Since the functional Ψ is convex, lower semicontinuous, and proper, then $\partial\Psi$ is a maximal monotone operator with dense domain, generating a contraction semigroup in $L^2(R^N)$ (see [16]). Therefore, we have the following result.

THEOREM 2.4. *Let $u_0 \in L^2(R^N)$. Then there exists a unique strong solution in the semigroup sense u of* (1.2), (1.3) *in $[0,T]$ for every $T > 0$, i.e., $u \in C([0,T]; L^2(R^N)) \cap W^{1,2}_{\rm loc}(0,T; L^2(R^N))$, $u(0) = u_0$, $u(t) \in D(\partial\Psi)$ for almost every $t \in [0,T]$, and*

$$-u'(t) \in \partial\Psi(u(t)) \quad \text{for a.e. } t \in [0,T]. \tag{2.9}$$

Moreover, if u and v are the strong solutions of (1.2) *corresponding to the initial conditions $u_0, v_0 \in L^2(\Omega)$, respectively, then*

$$\|u(t) - v(t)\|_2 \le \|u_0 - v_0\|_2 \quad \text{for any } t > 0.$$

The semigroup theory immediately provides us with existence and uniqueness results for (1.2). The characterization of $\partial\Psi$ given in Lemma 2.5 below (see [8, 9, 12] for a proof) allows us to write Theorem 2.4 in more classical terms.

LEMMA 2.5. *The following assertions are equivalent:*

(a) $(u, v) \in \partial\Psi$*;*

(b)

$$u \in L^2(R^N) \cap BV(R^N),\ v \in L^2(R^N), \tag{2.10}$$

$$\exists z \in X_2(R^N) \text{ with } \|z\|_\infty \le 1, \text{ such that } v = -\mathrm{div} z \quad \text{in } \mathcal{D}'(R^N),$$

and

$$\int_{R^N} (z, Du) = \int_{R^N} |Du|. \tag{2.11}$$

Let us now give a more classical definition of solution for problem (1.2). As we shall notice below, this notion coincides with the notion of a strong solution in the sense of semigroups defined above.

DEFINITION 2.6. *A function $u \in C([0,T]; L^2(R^N))$ is called a strong solution of* (1.2) *if*

$$u \in W^{1,2}_{\rm loc}(0,T; L^2(R^N)) \cap L^1_w(]0,T[; BV(R^N)),$$

and there exists $z \in L^\infty\left(]0,T[\times R^N; R^N\right)$ with $\|z\|_\infty \le 1$ such that

$$u_t = \mathrm{div}\ z \qquad \text{in } \mathcal{D}'(]0,T[\times R^N)$$

and

$$\int_{R^N} (z(t), Du(t)) = \int_{R^N} |Du(t)| \quad \text{for a.e. } t > 0. \tag{2.12}$$

We have the following result [8, 9, 12].

THEOREM 2.7. *Let $u_0 \in L^2(R^N)$. A function $u \in C([0,T]; L^2(R^N))$ is a strong solution of* (1.2) *with $u(0) = u_0$ if and only if it is a strong solution of it in the semigroup sense. Hence there exists a unique strong solution u of* (1.2), (1.3) *in $[0,T] \times R^N$ for every $T > 0$. Moreover, if u and v are the strong solutions of* (1.2) *corresponding to the initial conditions $u_0, v_0 \in L^2(R^N)$, respectively, then*

$$\|(u(t) - v(t))^+\|_2 \le \|(u_0 - v_0)^+\|_2 \quad \text{for any } t > 0. \tag{2.13}$$

Obviously, using Lemma 2.5, a strong solution of (1.2) is a strong solution in the sense of semigroups. The converse implication would follow along the same lines, except for the measurability of $z(t,x)$. To ensure the joint measurability of z, one takes into account that, by the Crandall–Liggett theorem [19], semigroup solutions can be approximated by implicit-in-time discretizations of (2.9), and one constructs a function $z(t,x) \in L^\infty(]0,T[\times R^N)$ satisfying the requirements contained in Definition 2.6. For details we refer to [8, 10]. Let us finally recall that, by a suitable extension of the notion of solution, we have existence, uniqueness, and stability results with respect to convergence in $L^1_{\rm loc}(R^N)$ for initial conditions in $L^1_{\rm loc}(R^N)$ [12].

Theorem 2.7 can be complemented with the following result.

THEOREM 2.8. *Let $u_0 \in L^2(R^N) \cap L^N(R^N)$ with support contained in a ball B of radius $R > 0$, and let $u(t,x)$ be the unique solution of problem* (1.2). *Then* $\mathrm{supp}(u) \subseteq B$. *If $T^*(u_0) = \inf\{t > 0 \colon u(t) = 0\}$, then*

$$T^*(u_0) \le \frac{R\|u_0\|_\infty}{N}. \tag{2.14}$$

Let

$$w(t,x) := \begin{cases} \dfrac{u(t,x)}{T^*(u_0) - t} & \text{if} \quad 0 \le t < T^*(u_0), \\ 0 & \text{if} \quad t \ge T^*(u_0). \end{cases}$$

Then there exists an increasing sequence $t_n \to T^*(u_0)$ and a solution $v^* \ne 0$ of the eigenvalue problem

$$v \in \partial\Psi(v) \tag{2.15}$$

such that

$$\lim_{n\to\infty} w(t_n) = v^* \quad \text{in} \quad L^p(R^N)$$

for all $1 \le p < \infty$.

By Lemma 2.5, equation (1.1) can be understood in more classical terms. Let us write this definition in a more general context. We shall use the truncatures $T_k(r) := (-k) \wedge r \vee k$, $r \in R$, $k > 0$.

DEFINITION 2.9. *Let Ω be an open set in R^N and let $f \in L^2_{\rm loc}(\Omega)$. We say that a function $u \in L^1_{\rm loc}(\Omega)$ is a solution of*

$$-\mathrm{div}\left(\frac{Du}{|Du|}\right) = f \quad \text{in } \Omega \tag{2.16}$$

if

$$T_k(u) \in BV_{\rm loc}(\Omega) \qquad \forall k > 0, \tag{2.17}$$

$$\exists z \in L^\infty(\Omega; R^N) \text{ with } \|z\|_\infty \le 1, \text{ such that } -\mathrm{div} z = f \quad \text{in } \mathcal{D}'(\Omega),$$

and

$$\langle (z, DT_k(u)), \varphi \rangle = \int_\Omega |DT_k(u)|\, \varphi \quad \textit{for any } \varphi \in C_0^\infty(\Omega), \tag{2.18}$$

where the left-hand side is defined as in (2.4).

The above definition also makes sense if we assume that $f \in L^1_{\rm loc}(R^N)$. Since this will not be needed in what follows, and to avoid cumbersome statements in subsection 2.2, we have assumed that $L^1_{\rm loc}(R^N)$.

Remark 2. If u is a solution of (2.16) and $f \in L^p_{\rm loc}(\Omega)$ with $p \geq 2$, then $(z, D\chi_{\{u>t\}}) = |D\chi_{\{u>t\}}|$ (in the sense that $\langle (z, D\chi_{\{u>t\}}), \varphi \rangle = \langle |D\chi_{\{u>t\}}|, \varphi \rangle$ for any $\varphi \in C_0^\infty(\Omega)$) for almost any $t \in R$. Indeed, by [11, Proposition 2.7], we have

$$\langle (z, DT_k(u)), \varphi \rangle = \int_{-k}^{k} \langle (z, D\chi_{\{u>t\}}), \varphi \rangle \, dt, \qquad \varphi \in C_0^\infty(\Omega),\ k > 0.$$

Since $|DT_k(u)|(\varphi) = \int_{-k}^{k} |D\chi_{\{u>t\}}|(\varphi)$, we may write (2.18) as

$$\int_{-k}^{k} \langle (z, D\chi_{\{u>t\}}), \varphi \rangle \, dt = \int_{-k}^{k} |D\chi_{\{u>t\}}|(\varphi)\, dt, \qquad \varphi \in C_0^\infty(\Omega),\ k > 0,$$

and this implies our claim.

Remark 3. If $u \in L^\infty(\Omega)$, condition (2.18) can be replaced by $(z, Du) = |Du|$.

3. Properties of $L_{\rm loc}$-solutions. Throughout the paper, from now on we shall assume that $N = 2$.

PROPOSITION 3.1. *Let Ω be an open set in R^2, and let $u \in L^p_{\rm loc}(\Omega)$ for some $p \in]2, +\infty]$. Let u be a solution of* (1.1) *in Ω. The following assertions hold.*

(a) *If $p < +\infty$ (resp., $p = +\infty$), then for any $t \in R$ the sets $\{u > t\}$ and $\{u \geq t\}$ have boundary of class $C^{1,\alpha}$ in Ω for some $\alpha \in]0, 1[$ (resp., $\mathcal{C}^{1,1}$). Similar assertions hold for $\{u < t\}$ and $\{u \leq t\}$.*

(b) *If $u \geq a$ in Ω (resp., $u \leq a$ in Ω) for some $a \in R$, then $\kappa_{\Omega \cap \partial\{u>t\}} \geq a$ and $\kappa_{\Omega \cap \partial\{u \geq t\}} \geq a$ in the sense of distributions.*

Proof. Let us prove (a). Let t be such that $\{u > t\}$ is nonempty and has locally finite perimeter in Ω and $(z, D\chi_{\{u>t\}}) = |D\chi_{\{u>t\}}|$ (in particular, by Remark 2, for almost every t). Let E be a set of finite perimeter in R^2 such that $E \triangle \{u > t\} \subset\subset \Omega$. Take a bounded Lipschitz set Ω' with $E \triangle \{u > t\} \subset\subset \Omega' \subset \Omega$. Then, using (1.1), we have

$$\int_{\{u>t\} \cap \Omega'} {\rm div} z \, dx - \int_{E \cap \Omega'} {\rm div} z \, dx \leq P(E, \Omega') - P(\{u > t\}, \Omega'). \tag{3.1}$$

It follows that $\{u > t\}$ is a minimizer of the functional

$$P(E, \Omega) + \int_{E \cap \Omega} {\rm div}\ z \, dx, \qquad E \subseteq R^2, \tag{3.2}$$

with respect to perturbations with compact support in Ω. Since by assumption $-{\rm div} z = u \in L^p_{\rm loc}(\Omega)$ for some $p \in]2, +\infty]$, using the regularity results for prescribed curvature problems (see [7, 30]), it follows that $\Omega \cap \partial\{u > t\}$ is of class $C^{1,\alpha}$ for some $\alpha \in]0, 1[$ if $p < +\infty$, and of class $\mathcal{C}^{1,1}$ if $p = +\infty$. By the compactness property of minimizers for problem (3.2) (see, for instance, [4]) the above assertion holds for any t, and (a) follows for $\{u > t\}$.

Let us prove (b). Assume that $u \geq a$ in Ω (the case $u \leq a$ is analogous). Let $t \in R$ be such that $\{u > t\}$ is nonempty and has locally finite perimeter in Ω and $(z, D\chi_{\{u>t\}}) = |D\chi_{\{u>t\}}|$ as in Remark 2 (hence, for almost every t). Let E be a set of finite perimeter in R^2 such that $E \supseteq \{u > t\}$ and $E \setminus \{u > t\} \subset\subset \Omega' \subset \Omega$, Ω' being a bounded set with Lipschitz boundary. Then from (3.1) it follows that

$$\begin{aligned} P(\{u > t\}, \Omega') &\leq P(E, \Omega') + \int_{(E\setminus\{u>t\})\cap\Omega'} \mathrm{div} z \; dx \\ &\leq P(E, \Omega') - a(|E \cap \Omega'| - |\{u > t\} \cap \Omega'|). \end{aligned}$$

It follows that $\{u > t\}$ is a minimizer of the functional

$$P(E, \Omega) - a|E \cap \Omega|, \qquad \{u > t\} \subseteq E \subseteq R^2,$$

with respect to perturbations with compact support in Ω. This concludes the proof of (b) [7, 30].

The corresponding assertions for the sets $\{u \geq t\}$ can be proved in a similar way. □

In what follows, given a function u as in Proposition 3.1 and $t \in R$, we always identify the set $\{u > t\}$ (resp., $\{u < t\}$) with its points of density one, which is an open set. We accordingly define $\{u \geq t\}$ as the complement of $\{u < t\}$.

4. Properties of $W^{1,1}_{\mathrm{loc}}$-solutions.

PROPOSITION 4.1. *Let u be a solution of* (1.1). *Assume that $u \in W^{1,1}_{\mathrm{loc}}(\Omega) \cap L^\infty_{\mathrm{loc}}(\Omega)$ for some open set $\Omega \subseteq R^2$. Then for any $t \in R$ every connected component of $\Omega \cap \partial\{u > t\}$ is contained in the boundary of a ball of radius $1/t$.*

Proof. Let $t \in R$, $\gamma := \Omega \cap \partial\{u > t\}$, and $\epsilon > 0$. By Proposition 3.1 the curve γ and the two curves $\gamma^-_\epsilon := \Omega \cap \partial\{u > t - \epsilon\}$, $\gamma^+_\epsilon := \Omega \cap \partial\{u < t + \epsilon\}$ are of class $\mathcal{C}^{1,1}$. Moreover, since $u \in W^{1,1}_{\mathrm{loc}}(\Omega)$, the two sets $\gamma^-_\epsilon \cap \gamma$ and $\gamma^+_\epsilon \cap \gamma$ are closed sets of zero $\mathcal{H}^1$-measure. Then the curve $\gamma \setminus (\gamma^-_\epsilon \cup \gamma^-_\epsilon)$ is contained in $\Omega \cap \{|u - t| < \epsilon\}$. Since γ is of class $\mathcal{C}^{1,1}$, by (b) of Proposition 3.1 it follows that γ has curvature belonging to $(t - \epsilon, t + \epsilon)$. The thesis follows by letting $\epsilon \to 0^+$. □

Note that if u is as in Proposition 4.1, then the set $\{u > t\}$ is a disjoint union of balls of radius $\frac{1}{t}$ for any $t \in R$ such that the boundary of $\{u > t\}$ is contained in Ω.

LEMMA 4.2. *Let $u \in W^{1,1}_{\mathrm{loc}}(R^2) \cap L^\infty_{\mathrm{loc}}(R^2)$ be a solution of* (1.1). *Then $u \equiv 0$.*

Proof. Assume by contradiction that $\lambda := \operatorname{ess\,sup}_{R^2} u > 0$ (the case $\operatorname{ess\,inf} u_{R^2} < 0$ can be treated in a similar way). Using Proposition 4.1 it follows that the set $\{u > t\}$ contains an open ball B_t of radius $\frac{1}{t}$ for any $t \in (0, \lambda)$. Fix $t \in (0, \lambda)$ and let $t^* := \operatorname{ess\,sup}_{B_t} u > t$. Then the closure of a connected component of the set $B_t \cap \{u = t^*\} = B_t \cap \{u \geq t^*\}$ is a closed ball $D_{t^*} \subset B_t$ of radius $\frac{1}{t^*}$. Using (1.1) we get

$$t^* = \frac{(t^*)^2}{\pi} \int_{D_{t^*}} u \; dx = -\frac{(t^*)^2}{\pi} \int_{D_{t^*}} \mathrm{div} z \; dx = 2t^*,$$

which is a contradiction. □

Loosely speaking, the following proposition classifies solutions with no jumps.

PROPOSITION 4.3. *Assume that u is a solution of* (1.1) *satisfying the following assumption:*

$$\forall t \in R \;\; \exists \text{ an open set } U_t \supset \partial\{u > t\} \text{ such that } u \in L^\infty_{\mathrm{loc}}(U_t).$$

Assume also that $T_k(u) \in W^{1,1}_{\rm loc}(R^2)$ for any $k > 0$. Then one of the following possibilities holds:

- *$u \equiv 0$;*
- *u is positive and the set $\{u > t\}$ is a ball of radius $\frac{1}{t}$ for any $t > 0$;*
- *u is negative and the set $\{u < t\}$ is a ball of radius $-\frac{1}{t}$ for any $t < 0$;*
- *u is nonnegative, $\{u > 0\}$ is a halfspace, and the set $\{u > t\}$ is a ball of radius $\frac{1}{t}$ for any $t > 0$;*
- *u is nonpositive, $\{u < 0\}$ is a halfspace, and the set $\{u < t\}$ is a ball of radius $-\frac{1}{t}$ for any $t < 0$;*
- *both $\{u > 0\}$ and $\{u < 0\}$ are halfspaces, the set $\{u > t\}$ is a ball of radius $\frac{1}{t}$ for any $t > 0$, and the set $\{u < \tau\}$ is a ball of radius $-\frac{1}{\tau}$ for any $\tau < 0$.*

Proof. Assume that $\lambda := \operatorname{ess\,sup} u > 0$ (the case $\operatorname{ess\,inf} u < 0$ being similar). From Proposition 4.1 we get that $\{u > t\}$ is the disjoint union of balls of radius $\frac{1}{t}$ for any $t \in (0, \lambda)$. Reasoning as in the proof of Lemma 4.2 we deduce that $\lambda = +\infty$. Observe that, given $0 < t_1 < t_2$, to each ball $B_1 \subseteq \{u > t_1\}$ (of radius $1/t_1$) there corresponds one and only one ball $B_2 \subseteq \{u > t_2\}$ (of radius $1/t_2$) such that $B_2 \subset B_1$, and vice versa. Hence there is a pairwise correspondence between the balls of $\{u > t_1\}$ and those of $\{u > t_2\}$. Letting $t \to 0^+$, $\{u > t\}$ consists of at most two balls, since given any three disjoint balls whose radius goes to infinity, at least one of them has a distance from a fixed point which goes to infinity. Hence $u > 0$ may consist of either one halfspace, two halfspaces, or the whole of R^2.

Claim. The set $\{u > t\}$ consists of exactly one ball of radius $\frac{1}{t}$ for any $t > 0$.

Observe that, once the claim is proved, all assertions of the proposition follow, since $\{u > 0\} = \bigcup_{t>0}\{u > t\}$ can only be a halfspace or the whole of R^2. Assume by contradiction that $\{u > t\}$ is the union of two balls (of radius $\frac{1}{t}$); hence $u \geq 0$ is the union of two halfspaces of R^2. Given $\tau < 0$, the set $\{u < \tau\}$ is either empty or contains a ball of radius $-\frac{1}{\tau}$; however, by the above argument there is no place for such a ball. Hence $u \geq 0$. Then $\{u = 0\}$ is either a line or a stripe. Without loss of generality, we may assume that $\{u = 0\} = [-l, l] \times R$ for some $l \geq 0$. Let $L > l$ and, for $t > 0$ small enough and such that $(z, D\chi_{\{u>t\}}) = |D\chi_{\{u>t\}}|$, set $S_{t,L} := \{u < t\} \cap]-L, L[^2$. Since $-\mathrm{div} z = u$ is bounded in $S_{t,L}$, we have

$$
\begin{aligned}
0 \geq -\int_{S_{t,L}} u \, dx = \int_{S_{t,L}} \mathrm{div}\, z \, dx = \int_{\partial S_{t,L}} [z, \nu^{S_{t,L}}] \, d\mathcal{H}^1 \\
\geq \mathcal{H}^1(\partial S_{t,L} \cap \partial\{u < t\}) - \mathcal{H}^1(\partial S_{t,L} \cap \{u < t\}) \geq 4L - \mathcal{H}^1(\partial S_{t,L} \cap \{u < t\}).
\end{aligned}
$$

Letting $t \to 0^+$ and using the fact that $\{u > t\}$ is the union of two balls of radius $1/t$, we obtain $4L - 4l \leq 0$, a contradiction. Our claim is proved and the proposition follows. ∎

5. Solutions of div z = constant in bounded domains. In the following, $m \geq 1$ is an integer, and we denote by $C_0, C_1, \ldots, C_m$ bounded open sets of R^2 with boundary of class $\mathcal{C}^{1,1}$ having the following properties:

- $\overline{C_l} \subset C_0$ for any $l \in \{1, \ldots, m\}$;
- $\overline{C_l} \cap \overline{C_h} = \emptyset$ for any $l, h \in \{1, \ldots, m\}$, $l \neq h$.

We define

$$
F := C_0 \setminus \bigcup_{l=1}^{m} \overline{C_l},
$$

$$J_0 := \frac{1}{|F|}\left(\sum_{i=0}^{k} P(C_i) - \sum_{j=k+1}^{m} P(C_j)\right), \tag{5.1}$$

where $0 \le k < m$ is a fixed integer.

Given a set $E \subseteq F$ of finite perimeter in F, we also let

$$\mathcal{F}_F(E) := P(E,F) + \sum_{i=0}^{k} \mathcal{H}^1(\partial^* E \cap \partial C_i) - \sum_{j=k+1}^{m} \mathcal{H}^1(\partial^* E \cap \partial C_j) - J_0|E|.$$

Remark 4. It is clear that $\mathcal{F}_F(\emptyset) = 0$. Observe also that, thanks to the definition of J_0, $\mathcal{F}_F(F) = 0$.

We now define a class $\mathcal{A}$ of subsets of F.

DEFINITION 5.1. *Let $E \subseteq F$ be a finite perimeter set and let $J_0 > 0$. We say that $E \in \mathcal{A}$ if either $E \in \{\emptyset, F\}$ or the following conditions hold: $F \cap \partial^* E$ consists of disjoint arcs Γ of circles of radius $1/J_0$, with $\partial F \cap \overline{\Gamma} \neq \emptyset$, and*

$$\nu^E = \nu^{C_0} \quad \text{on } \overline{\Gamma} \cap \partial C_0, \tag{5.2}$$

$$\nu^E = -\nu^{C_i} \quad \text{on } \overline{\Gamma} \cap \partial C_i,\ i \in \{1,\dots,k\}, \tag{5.3}$$

$$\nu^E = \nu^{C_j} \quad \text{on } \overline{\Gamma} \cap \partial C_j,\ j \in \{k+1,\dots,m\}. \tag{5.4}$$

In (5.2), (5.3), and (5.4) we keep the notation ν^E to indicate the extension of the outer unit normal vector to ∂E at the points of $\overline{\Gamma}$.

The following result can be essentially found in [25, Theorem 1] and [26, Theorem 6.10]. Indeed, the results in [25, 26] cover the case of equalities (5.2) and (5.3), but they can be adapted to prove (5.4).

THEOREM 5.2. *Let $E \subseteq F$ be a finite perimeter set and assume that $\mathcal{F}_F(E) = \min\{\mathcal{F}_F(B) : B \subseteq F\}$. Then $E \in \mathcal{A}$.*

The equivalence (a) $\Longleftrightarrow$ (c) of the next theorem in the crystalline case has been investigated in [13].

THEOREM 5.3. *The following conditions are equivalent:*

(a) *There exists a vector field $z : F \to R^2$ satisfying*

$$z \in L^\infty(F;R^2), \quad \begin{cases} -\operatorname{div} z = J_0 & \text{in } \mathcal{D}'(F), \\ \|z\|_\infty \le 1, & \\ [z,\nu^F] = -1 & \mathcal{H}^1\text{-a.e. on } \partial C_i,\ i \in \{0,\dots,k\}, \\ [z,\nu^F] = 1 & \mathcal{H}^1\text{-a.e. on } \partial C_j,\ j \in \{k+1,\dots,m\}. \end{cases} \tag{5.5}$$

(b) *We have*

$$J_0 \int_F w \le \int_F |Dw| + \sum_{i=0}^{k} \int_{\partial C_i} w - \sum_{j=k+1}^{m} \int_{\partial C_j} w \qquad \forall w \in BV(F). \tag{5.6}$$

(c) *For any set $E \subseteq F$ of finite perimeter in F we have $\mathcal{F}_F(E) \ge 0$.*

(d) *We have*

$$\min_{E \in \mathcal{A}} \mathcal{F}_F(E) = 0. \tag{5.7}$$

Proof. We divide the proof into several steps.

Step 1. Let Ω be an open bounded connected subset of R^2 with $\mathcal{C}^{1,1}$ boundary, $f \in L^2(\Omega)$, $g \in L^\infty(\partial\Omega)$, and $\lambda > 0$. Assume that $\|g\|_\infty < 1$. A function $u \in BV(\Omega) \subset L^2(\Omega)$ is a solution of

$$\min_{w\in BV(\Omega)} \mathcal{E}(w), \qquad \mathcal{E}(w) := \int_\Omega |Dw| + \frac{1}{2\lambda}\int_\Omega (w-f)^2\,dx - \int_{\partial\Omega} gw \tag{5.8}$$

if and only if there exists $z \in X_2(\Omega)$, with $\|z\|_\infty \le 1$, satisfying $(z, Du) = |Du|$ as measures in Ω, $[z, \nu^\Omega] = g\,\mathcal{H}^1$-almost everywhere on $\partial\Omega$ and $-\lambda \mathrm{div} z = f - u$ in $\mathcal{D}'(\Omega)$.

We observe that the functional $\mathcal{E}$ is convex and L^1-lower semicontinuous. Moreover, since $\|g\|_\infty < 1$ and $\partial\Omega$ is of class $\mathcal{C}^{1,1}$, using the results of Giusti [28] we get that $\mathcal{E}$ is coercive. Therefore it attains its minimum, which is also unique. Hence $u = \mathrm{argmin}\,\mathcal{E}$ if and only if $0 \in \partial\mathcal{E}(u)$, where ∂ denotes the subdifferential in L^2.

We now define the operator $\mathcal{A}_g$ in $L^2(\Omega)\times L^2(\Omega)$ as follows: $(w, v) \in \mathcal{A}_g$ if and only if $w \in BV(\Omega)$, $v \in L^2(\Omega)$, and there is a vector field $z \in L^\infty(\Omega, R^2)$ with $\|z\|_\infty \le 1$ such that $(z, Dw) = |Dw|$, $-\mathrm{div}\, z = v$ in $\mathcal{D}'(\Omega)$, and $[z, \nu^\Omega] = g\,\mathcal{H}^1$-almost everywhere on $\partial\Omega$. Let us prove that the operator $\mathcal{A}_g$ is maximal monotone. As a consequence, since $\mathcal{A}_g \subseteq \partial\mathcal{E}$ and both are maximal monotone, we conclude that $\mathcal{A}_g = \partial\mathcal{E}$. This will prove Step 1.

The monotonicity of $\mathcal{A}_g$ follows by an integration by parts. To prove the maximal monotonicity, we have to solve

$$f \in u + \mathcal{A}_g u \qquad \forall f \in L^2(\Omega). \tag{5.9}$$

First, we assume that $f \in L^\infty(\Omega)$. Let us approximate (5.9) by

$$\begin{cases} u - \mathrm{div}(\mathcal{T}_\epsilon u) = f & \text{in } \Omega, \\ [\mathcal{T}_\epsilon u, \nu^\Omega] = g & \text{in } \partial\Omega, \end{cases} \qquad \mathcal{T}_\epsilon u := \frac{Du}{\sqrt{\epsilon^2 + |Du|^2}}. \tag{5.10}$$

Following [28], we have that (5.10) has a unique solution $u_\epsilon \in BV(\Omega)$. If we further assume that $f \in W^{1,\infty}(\Omega)$, we have $u_\epsilon \in W^{1,1}(\Omega)$ (actually $u_\epsilon \in C^{2,\alpha}(\overline{\Omega})$; see [28]).

Let us prove the basic estimates required to pass to the limit as $\epsilon \to 0$.

(i) L^2 and bounded variation estimates on u_ϵ when $f \in L^\infty(\Omega)$: multiplying (5.10) by u_ϵ, after integration by parts, we get

$$\int_\Omega u_\epsilon^2 + \int_\Omega \mathcal{T}_\epsilon u_\epsilon \cdot Du_\epsilon = \int_\Omega f u_\epsilon + \int_{\partial\Omega} g u_\epsilon.$$

Since $\frac{x^2}{\sqrt{\epsilon^2+x^2}} \ge |x| - \epsilon$ for all $x \in R$, from the above estimate we have

$$\int_\Omega u_\epsilon^2 + \int_\Omega |Du_\epsilon| \le \epsilon|\Omega| + \int_\Omega f u_\epsilon + \int_{\partial\Omega} g u_\epsilon. \tag{5.11}$$

Now, using [28, Lemma 1.2] and $\|g\|_\infty =: 1 - 2\sigma < 1$, there is a constant c depending on σ, g, Ω, such that

$$\left|\int_{\partial\Omega} gw\right| \le (1-\sigma)\int_\Omega |Dw| + c\int_\Omega |w| \qquad \forall w \in BV(\Omega). \tag{5.12}$$

Inserting (5.12) in (5.11) we obtain the estimate

$$\frac{1}{2}\int_\Omega u_\epsilon^2 + \sigma\int_\Omega |Du_\epsilon| \le (\epsilon + c^2)|\Omega| + \|f\|_2^2.$$

Thus, by extracting a subsequence, if necessary, we may assume that $u_\epsilon \to u$ in $L^p(\Omega)$ for any $1 \le p < 2$ and weakly in $L^2(\Omega)$, where $u \in BV(\Omega)$.

(ii) L^3 estimate on u_ϵ when $f \in W^{1,\infty}(\Omega)$. We multiply (5.10) by $|T_k(u_\epsilon)|u_\epsilon$. After integrating by parts we obtain

$$\int_\Omega u_\epsilon^2 |T_k(u_\epsilon)| + \int_\Omega \mathcal{T}_\epsilon u_\epsilon \cdot D(|T_k(u_\epsilon)|u_\epsilon) = \int_\Omega f|T_k(u_\epsilon)|u_\epsilon + \int_{\partial\Omega} g|T_k(u_\epsilon)|u_\epsilon.$$

Using (5.12) and

$$\int_\Omega \mathcal{T}_\epsilon u_\epsilon \cdot D(|T_k(u_\epsilon)|u_\epsilon) \geq \int_\Omega |D(|T_k(u_\epsilon)|u_\epsilon)| - \epsilon \int_\Omega [|u_\epsilon| + |T_k(u_\epsilon)|]$$

we obtain

$$\int_\Omega u_\epsilon^2 |T_k(u_\epsilon)| + \sigma \int_\Omega |D(|T_k(u_\epsilon)|u_\epsilon| \leq (\|f\|_\infty + c) \int_\Omega |T_k(u_\epsilon)||u_\epsilon| + \epsilon \int_\Omega |u_\epsilon| + \epsilon \int_\Omega |T_k(u_\epsilon)|.$$

Since u_ϵ is bounded in $L^2(\Omega)$, letting $k \to \infty$, we deduce that u_ϵ is bounded in $L^3(\Omega)$. Thus also $u \in L^3(\Omega)$.

Now,

$$\int_\Omega (u_\epsilon - u)^2\, dx \leq \left(\int_\Omega |u_\epsilon - u|^3\, dx\right)^{1/2} \left(\int_\Omega |u_\epsilon - u|\, dx\right)^{1/2} \to 0 \quad \text{as } \epsilon \to 0.$$

Thus we may extract a sequence u_ϵ converging in $L^2(\Omega)$ to some function $u \in BV(\Omega)$. Moreover, we may assume that $\mathcal{T}_\epsilon u_\epsilon \to z$ weakly* in $L^\infty(\Omega, R^2)$. Letting $\epsilon \to 0$ in (5.10) we have

$$u - \operatorname{div} z = f \quad \text{in } \mathcal{D}'(\Omega). \tag{5.13}$$

Still we have to prove that $(z, Du) = |Du|$ and $[z, \nu^\Omega] = g$.

Let φ be a smooth function in Ω, continuous up to $\partial\Omega$. We multiply (5.10) by φ and integrate by parts to obtain

$$\int_\Omega u_\epsilon \varphi + \int_\Omega \mathcal{T}_\epsilon u_\epsilon \cdot \nabla\varphi - \int_{\partial\Omega} [\mathcal{T}_\epsilon u_\epsilon, \nu^\Omega]\varphi = \int_\Omega f\varphi. \tag{5.14}$$

Letting $\epsilon \to 0$ and using that $[\mathcal{T}_\epsilon u_\epsilon, \nu^\Omega] = g$, we obtain

$$\int_\Omega u\varphi + \int_\Omega z \cdot \nabla\varphi - \int_{\partial\Omega} g\varphi = \int_\Omega f\varphi. \tag{5.15}$$

Integrating by parts the second term of the above equality, we get

$$\int_\Omega u\varphi - \int_\Omega \operatorname{div} z\, \varphi + \int_{\partial\Omega} ([z, \nu^\Omega] - g)\varphi = \int_\Omega f\varphi. \tag{5.16}$$

Now, using (5.13) it follows that $\int_{\partial\Omega}([z, \nu^\Omega] - g)\varphi = 0$ for all test functions φ. This implies that $[z, \nu^\Omega] = g$ on $\partial\Omega$.

To prove that $(z, Du) = |Du|$, we observe that from the lower semicontinuity of $\mathcal{E}$ and the convergence $\int_\Omega (u_\epsilon - f)^2\, dx \to \int_\Omega (u - f)^2\, dx$ as $\epsilon \to 0$, we have

$$\begin{aligned}\int_\Omega |Du| - \int_{\partial\Omega} gu &\leq \liminf_\epsilon \left(\int_\Omega |Du_\epsilon| - \int_{\partial\Omega} gu_\epsilon\right) = \liminf_\epsilon \left(\int_\Omega (\mathcal{T}_\epsilon u_\epsilon, Du_\epsilon) - \int_{\partial\Omega} gu_\epsilon\right) \\ &= \liminf_\epsilon - \int_\Omega \operatorname{div} \mathcal{T}_\epsilon u_\epsilon\, u_\epsilon = - \int_\Omega \operatorname{div} z\, u \\ &= \int_\Omega (z, Du) - \int_{\partial\Omega} gu \leq \int_\Omega |Du| - \int_{\partial\Omega} gu.\end{aligned}$$

We conclude that $\int_\Omega (z, Du) = \int_\Omega |Du|$.

We have proved that there is a solution of (5.9) for each $f \in W^{1,\infty}(\Omega)$. Our next goal is to prove that the operator $\mathcal{A}_g$ is closed. As a consequence we obtain that (5.9) has a solution for each $f \in L^2(\Omega)$. To prove the closedness of $\mathcal{A}_g$, let $(u_n, v_n) \in \mathcal{A}_g$ be such that $(u_n, v_n) \to (u, v)$ in $L^2(\Omega) \times L^2(\Omega)$. Then there is a vector field $z_n \in L^\infty(\Omega, R^2)$ with $\|z_n\|_\infty \le 1$ such that $v_n = -\mathrm{div}\, z_n$, $(z_n, Du_n) = |Du_n|$ and $[z_n, \nu^\Omega] = g$. Modulo a subsequence, we may assume that $z_n \to z$ weakly* in $L^\infty(\Omega, R^2)$ with $\|z\|_\infty \le 1$. Since $v_n = -\mathrm{div}\, z_n \to -\mathrm{div}\, z$ in $\mathcal{D}'(\Omega)$, we have $v = -\mathrm{div}\, z$. The proofs of the facts $[z, \nu^\Omega] = g$ and $(z, Du) = |Du|$ follow the same arguments as those in the corresponding proofs above, and we shall omit the details. We conclude that $\mathcal{A}_g$ is closed in $L^2(\Omega)$. This ends the proof that $\mathcal{A}_g$ is maximal monotone and $\partial\mathcal{E} = \mathcal{A}_g$.

Step 2. The function $u \equiv 0$ is the solution of (5.8) if and only if f and g satisfy

$$\int_\Omega |Dw| \ge \frac{1}{\lambda} \int_\Omega wf\, dx + \int_{\partial\Omega} gw \qquad \forall w \in BV(\Omega). \tag{5.17}$$

The proof follows along the same lines as the proof of [12, Lemma 1]. Clearly $u \equiv 0$ is the solution of (5.8) if and only if

$$\int_\Omega |Dw| + \frac{1}{2\lambda} \int_\Omega (w - f)^2\, dx - \int_{\partial\Omega} gw \ge \frac{1}{2\lambda} \int_\Omega f^2\, dx \qquad \forall w \in BV(\Omega). \tag{5.18}$$

Replacing w by ϵw (where $\epsilon > 0$), expanding the L^2-norm, dividing by $\epsilon > 0$, and letting $\epsilon \to 0^+$, we have (5.17).

On the other hand, if (5.17) holds, (5.18) also holds. Finally note that, replacing w by $-w$, we see that we may replace the right-hand side of (5.17) by its absolute value.

Step 3. Problem (5.5) has a solution if and only if (5.6) holds.

Note that it is enough to prove inequality (5.6) only for functions $w \in BV(F)$, which do not change sign, i.e., $w \ge 0$ or $w \le 0$.

Suppose that (5.5) has a solution z. Let $w \in BV(F)$. Multiplying $-\mathrm{div} z = J_0$ on F by w and integrating by parts, we obtain that (5.6) holds.

Assume now that (5.6) holds. Multiplying (5.6) by $1 - \epsilon > 0$ we deduce that

$$(1-\epsilon)J_0 \int_F w \le \int_F |Dw| + (1-\epsilon) \sum_{i=0}^{k} \int_{\partial C_i} w - (1-\epsilon) \sum_{j=k+1}^{m} \int_{\partial C_j} w$$
$$\forall w \in BV(F).$$

Thus, by Step 2 with $\lambda = 1$ we deduce that $u = 0$ is a solution of (5.8) with $f = (1-\epsilon)J_0\chi_F$, and $g \equiv -(1-\epsilon)$ in ∂C_i, $i \in \{0, \dots, k\}$, and $g \equiv 1 - \epsilon$ in ∂C_j, $j \in \{k+1, \dots, m\}$, for all $\epsilon \in]0, 1[$. Then by Step 1, we know that there exists a solution $\xi_\epsilon \in L^\infty(F, R^2)$ such that $\|\xi_\epsilon\|_\infty \le 1$, $-\mathrm{div}\xi_\epsilon = (1-\epsilon)J_0\chi_F$, $[\xi_\epsilon, \nu^F] = g$. Letting $\epsilon \to 0$, we find a vector field z satisfying (5.5).

Step 4. Conditions (b) and (c) are equivalent.

(c) follows from (b) by taking $w = \chi_E$ in (5.6) for any set of finite perimeter $E \subseteq F$. (b) follows from (c) by means of the coarea formula. Indeed, let $w \in BV(F)$, $w \ge 0$. We have

$$J_0\int_F w\,dx = J_0\int_0^\infty\int_F \chi_{\{w\ge t\}}\chi_F\,dx\,dt = J_0\int_0^\infty |\{w\ge t\}\cap F|\,dt$$
$$\le \int_0^\infty P(\{w\ge t\},F)\,dt + \sum_{i=0}^{k}\int_0^\infty \mathcal{H}^1(\partial^*\{w\ge t\}\cap \partial C_i)\,dt$$
$$- \sum_{j=k+1}^{m}\int_0^\infty \mathcal{H}^1(\partial^*\{w\ge t\}\cap \partial C_j)\,dt$$
$$= \int_F |Dw| + \sum_{i=0}^{k}\int_{\partial C_i} w - \sum_{j=k+1}^{m}\int_{\partial C_j} w.$$

Let us prove the corresponding inequality for $w\in BV(F)$, $w\le 0$. First, we observe that, writing $F\setminus E$ instead of E in (c), we obtain

$$P(F\setminus E,F) + \sum_{i=0}^{k}\mathcal{H}^1(\partial^*(F\setminus E)\cap\partial C_i) - \sum_{j=k+1}^{m}\mathcal{H}^1(\partial^*(F\setminus E)\cap\partial C_j) - J_0|F\setminus E|\ge 0.$$

Since $P(F\setminus E,F) = P(E,F)$ and $\mathcal{H}^1(\partial^*(F\setminus E)\cap\partial C_l) = P(C_l) - \mathcal{H}^1(\partial^* E\cap\partial C_l)$, using (5.1), we may write the last equation as

$$P(E,F) + \sum_{j=k+1}^{m}\mathcal{H}^1(\partial^*E\cap\partial C_j) - \sum_{i=0}^{k}\mathcal{H}^1(\partial^*E\cap\partial C_i) + J_0|E|\ge 0. \tag{5.19}$$

Now, we may proceed as in the case where $w\ge 0$ but using (5.19) instead of (c). Indeed,

$$J_0\int_F w\,dx = -J_0\int_{-\infty}^0\int_F \chi_{\{w\le t\}}\chi_F\,dx\,dt = -J_0\int_{-\infty}^0 |\{w\le t\}\cap F|\,dt$$
$$\le \int_{-\infty}^0 P(\{w\le t\},F)\,dt - \sum_{i=0}^{k}\int_{-\infty}^0 \mathcal{H}^1(\partial^*\{w\le t\}\cap\partial C_i)\,dt$$
$$+ \sum_{j=k+1}^{m}\int_{-\infty}^0 \mathcal{H}^1(\partial^*\{w\le t\}\cap\partial C_j)\,dt$$
$$= \int_F |Dw| + \sum_{i=0}^{k}\int_{\partial C_i} w - \sum_{j=k+1}^{m}\int_{\partial C_j} w.$$

Finally, if $w\in BV(F)$, we decompose $w = w^+ + w^-$, write the corresponding inequalities (5.6) for w^+ and w^-, and add them to obtain that (5.6) holds for w.

Step 5. Condition (c) is equivalent to

$$\min_{E\subseteq F}\mathcal{F}_F(E) = \mathcal{F}_F(\emptyset) = \mathcal{F}_F(F) = 0, \tag{5.20}$$

where the minimum is taken on the sets $E\subseteq F$ of finite perimeter. Moreover, any set $E\subseteq F$ of finite perimeter minimizing the left-hand side of (5.20) belongs to $\mathcal{A}$ by Theorem 5.2; therefore condition (c) is equivalent to condition (d). □

Given a set $E\subseteq R^2$, of finite perimeter in R^2, we define the functional $\mathcal{G}$ as

$$\mathcal{G}(E) := P(E) - \sum_{j=k+1}^{m} P(C_j) - J_0|E\cap F|.$$

Remark 5. Recalling the definition of J_0, we have $\mathcal{G}(F \cup (\bigcup_{j=k+1}^m C_j)) = 0$.

PROPOSITION 5.4. *The following conditions are equivalent:*

(a) *The set* $F \cup (\bigcup_{j=k+1}^m C_j)$ *is a solution of the variational problem*

$$\min\left\{\mathcal{G}(E) : \bigcup_{j=k+1}^m C_j \subseteq E \subseteq C_0 \setminus \bigcup_{i=1}^k \overline{C}_i\right\}. \tag{5.21}$$

(b) *There exists a vector field* z *satisfying* (5.5).

Remark 6. If $k = 0$ in Proposition 5.4, the last inclusion in (5.21) must be understood as $E \subseteq C_0$.

Proof of Proposition 5.4. Assume that there exists a vector field z satisfying (5.5). Given a finite perimeter set $E \subset R^2$ such that $\bigcup_{j=k+1}^m C_j \subseteq E \subseteq C_0 \setminus \bigcup_{i=1}^k \overline{C}_i$, we integrate the divergence of z on $E \cap F$ and obtain

$$\begin{aligned} J_0|E \cap F| &= -\int_{E\cap F} \operatorname{div} z \, dx \\ &\le P(E \cap F, F) + \sum_{i=0}^k \mathcal{H}^1(\partial^*(E \cap F) \cap \partial C_i) - \mathcal{H}^1\left(\partial^*(E \cap F) \cap \left(\bigcup_{j=k+1}^m C_j\right)\right) \\ &= P(E) - \sum_{j=k+1}^m P(C_j). \end{aligned}$$

It follows that $\mathcal{G}(E) \ge 0$, and (a) follows.

Let us now assume that (a) holds. Let $D \subset F$ be a set of finite perimeter. By Theorem 5.3 (see condition (c)), to obtain a vector field satisfying (5.5) it is enough to prove that

$$P(D) - 2\sum_{j=k+1}^m \mathcal{H}^1(\partial^* D \cap \partial C_j) \ge J_0|D|. \tag{5.22}$$

Set $A := D \cup \bigcup_{j=k+1}^m C_j$. By assumption we have

$$\begin{aligned} 0 \le \mathcal{G}(A) &= P(A) - \sum_{j=k+1}^m P(C_j) - J_0|D| \\ &= P(D) - 2\sum_{j=k+1}^m \mathcal{H}^1(\partial^* D \cap \partial C_j) + \sum_{j=k+1}^m P(C_j) - \sum_{j=k+1}^m P(C_j) - J_0|D|, \end{aligned}$$

which is (5.22). □

Remark 7. If we consider the case in which $k = 0$, then J_0 tends to zero as C_0 tends to R^2; in this case, the minimum problem (5.21) reduces to the problem considered in [12, Theorem 6].

5.1. Characterization through the curvature of the boundaries. The aim of this subsection is to prove Theorem 5.10, which is a characterization of the solvability of problem (5.5) through pointwise curvature conditions on the boundaries of the sets C_i. We begin with some preliminaries. The next definition is taken from [18, Theorem 4.1].

DEFINITION 5.5. *Let $\Omega \subseteq R^2$ be an open set with boundary of class $\mathcal{C}^{1,1}$ and $\rho > 0$. We say that Ω satisfies the ρ-ball condition if an open ball of radius ρ can be rotated along $\partial\Omega$ in Ω in such a way that no antipods of the ball lie on $\partial\Omega$.*

It is clear that if Ω satisfies the ρ-ball condition, then it satisfies the σ-ball condition for any $\sigma \in]0, \rho]$.

LEMMA 5.6. *Let $\Omega \subseteq R^2$ be an open set satisfying the ρ-ball condition for some $\rho > 0$. Then $\operatorname{ess\,sup}_{\partial\Omega} \kappa_{\partial\Omega} \leq \frac{1}{\rho}$. Moreover, given an open ball $B_\rho \subset \Omega$ of radius ρ and tangent to $\partial\Omega$, the set $\gamma \cap \partial B_\rho$ is connected for any connected component γ of $\partial\Omega$ and spans an angle strictly less than π.*

Proof. The inequality $\operatorname{ess\,sup}_{\partial\Omega} \kappa_{\partial\Omega} \leq \frac{1}{\rho}$ is immediate. Now let $p, q \in \partial B_\rho \cap \partial\Omega$, and denote by $\gamma \subset \partial B_\rho$ the shortest of the two circular arcs in ∂B_ρ having p and q as boundary points (such a γ is uniquely determined since p and q cannot be antipodal by the ρ-ball condition). If $\gamma \not\subset \partial\Omega$, we slightly rotate B_ρ along $\partial\Omega$ around p towards q, and denoting by B' such a rotated ball, one verifies that q belongs to the interior of B', thus violating the ρ-ball condition. Hence $\gamma \subseteq \partial B_\rho \cap \partial\Omega$, and γ spans an angle strictly less than π. □

Remark 8. In general, the inequality $\operatorname{ess\,sup}_{\partial\Omega} \kappa_{\partial\Omega} \leq \frac{1}{\rho}$ does not imply the ρ-ball condition for the set Ω. However, if Ω is a convex set with boundary of class $\mathcal{C}^{1,1}$ such that $\operatorname{ess\,sup}_{\partial\Omega} \kappa_{\partial\Omega} < \frac{1}{\rho}$, then Ω satisfies the ρ-ball condition.

Remark 9. If C_l is convex for any $l \in \{0, \dots, m\}$, $\operatorname{ess\,sup}_{\partial C_0} \kappa_{\partial C_0} < J_0$ (in particular $J_0 > 0$), and

$$\operatorname{dist}(\partial C_l, \partial C_h) > \frac{2}{J_0} \qquad \forall (l, h) \in \{0, \dots, m\},\ l \neq h,$$

then F satisfies the $\frac{1}{J_0}$-ball condition.

Given a function $f \in W^{1,1}(]a, b[) \cap C^{1,1}(]a, b[)$, we denote by $\kappa(x, f(x))$ the curvature of the graph of f at the point $(x, f(x))$, i.e.,

$$\kappa(x, f(x)) := -\frac{f''(x)}{(1 + f'^2(x))^{3/2}} \quad \text{for a.e. } x \in\ ]a, b[.$$

LEMMA 5.7. *Let $f, g \in W^{1,1}(]a, b[) \cap C^{1,1}(]a, b[)$ be such that $f \leq g$ on $[a, b]$, and $f(a) = g(a)$, $f(b) = g(b)$. Assume that $\operatorname{ess\,inf}_{]a,b[} \kappa(x, f(x)) \geq \operatorname{ess\,sup}_{]a,b[} \kappa(x, g(x)) \geq 0$. Then $f = g$.*

Proof. By a smoothing argument we can assume that $f, g \in C^2(]a, b[)$. Suppose by contradiction that $f \neq g$, and let $c := \max_{[a,b]}(g - f) > 0$. Let us fix $\epsilon > 0$ and consider the function $f_\epsilon(x) := (1 - \epsilon) f(x/(1 - \epsilon))$, $x \in [(1 - \epsilon)a, (1 - \epsilon)b]$. Then, for ϵ small enough, the function $g - f_\epsilon$ attains its maximum at a point $\overline{x} \in]a, b[\cap](1 - \epsilon)a, (1 - \epsilon)b[$. Hence $g'(\overline{x}) = f'_\epsilon(\overline{x})$, $g''(\overline{x}) \leq f''_\epsilon(\overline{x})$. It follows that

$$\kappa(\overline{x}, g(\overline{x})) \geq \kappa(\overline{x}, f_\epsilon(\overline{x})) = \frac{1}{1 - \epsilon} \kappa\left(\frac{\overline{x}}{1 - \epsilon}, f\left(\frac{\overline{x}}{1 - \epsilon}\right)\right) > \kappa\left(\frac{\overline{x}}{1 - \epsilon}, f\left(\frac{\overline{x}}{1 - \epsilon}\right)\right),$$

which gives a contradiction. □

LEMMA 5.8. *Let K_0 and K_1 be two bounded strictly convex sets of class $\mathcal{C}^{1,1}$ in the plane, with $K_1 \subseteq K_0$ and $K_1 \neq K_0$. Assume that $\operatorname{ess\,sup}_{\partial K_0} \kappa_{\partial K_0} \leq \operatorname{ess\,inf}_{\partial K_1} \kappa_{\partial K_1}$. Then either $\partial K_0 \cap \partial K_1 = \emptyset$ or $\partial K_0 \cap \partial K_1$ is a connected arc which spans an angle strictly less than π.*

Proof. Let Γ be a connected component of $\partial K_0 \setminus \partial K_1$, and assume that $\Gamma \neq \partial K_0$. It is enough to prove that Γ spans an angle strictly greater than π. Assume by

contradiction that Γ spans an angle less than or equal to π. Then there exists an arc $\Gamma' \subset \partial K_1 \setminus \partial K_0$ which also spans an angle less than or equal to π and has the same endpoints as Γ. By the strict convexity of K_0 and K_1, and with a proper choice of a coordinate system, we may assume that Γ' and Γ are, respectively, the graphs of two functions f and g, which satisfy the assumptions of Lemma 5.7. We get a contradiction from that lemma. □

We recall the following result, which follows from [34, (6.52)].

PROPOSITION 5.9. *Let K_0 and K_1 be two bounded convex sets of class $\mathcal{C}^{1,1}$ in the plane, with $K_1 \subseteq K_0$. Assume that* $\operatorname{ess\,sup}_{\partial K_0} \kappa_{\partial K_0} \leq \operatorname{ess\,inf}_{\partial K_1} \kappa_{\partial K_1}$. *Then*

$$2\pi(|K_0| + |K_1|) - P(K_0)P(K_1) \geq 0.$$

Moreover the inequality is strict if $K_1 \subset\subset K_0$.

Remark 10. Let $\lambda > 0$. Then the function

$$\rho \to P(B_\rho) - \lambda |B_\rho| = \pi(2\rho - \lambda\rho^2)$$

attains its maximum at $\rho = 1/\lambda$.

We are now in a position to prove the main result of this section.

THEOREM 5.10. *Assume that there exists a vector field $z : F \to R^2$ satisfying* (5.5). *Then*

$$\operatorname*{ess\,sup}_{\partial C_0} \kappa_{\partial C_0} \leq J_0, \tag{5.23}$$

$$\operatorname*{ess\,inf}_{\partial C_i} \kappa_{\partial C_i} \geq -J_0, \qquad i \in \{1, \dots, k\}, \tag{5.24}$$

$$\operatorname*{ess\,inf}_{\partial C_j} \kappa_{\partial C_j} \geq J_0, \qquad j \in \{k+1, \dots, m\}. \tag{5.25}$$

Conversely, assume that

(a) *the inequality* (5.25) *holds;*

(b) *$F \cup (\bigcup_{j=k+1}^m C_j)$ satisfies the $\frac{1}{J_0}$-ball condition;*

(c) $\operatorname{dist}(\partial C_l, \partial C_h) > \frac{2}{J_0}$ *for all $(l,h) \in \{0, \dots, k\}^2 \cup \{k+1, \dots, m\}^2$, $l \neq h$.*

Then there exists a vector field $z : F \to R^2$ satisfying (5.5).

Remark 11. If $k = 0$ in Theorem 5.10, then condition (5.24) does not appear.

Proof of Theorem 5.10. Assume that problem (5.5) has a solution. Fix $j \in \{k+1, \dots, m\}$ and $x \in \partial C_j$. Let A be an open neighborhood of x where ∂C_j can be written as a graph; we can assume that $\overline{A} \subset C_0$ and $\overline{A} \cap (\cup_{l \in \{1,\dots,m,\}, l \neq j} C_l) = \emptyset$. We claim that

$$P(C_j) - J_0|C_j| \leq P(C_j \cup B) - J_0|C_j \cup B| \qquad \forall B \text{ Borel}, \ \overline{B} \subset A. \tag{5.26}$$

Let B be a Borel set with $\overline{B} \subseteq A$. We can assume that $P(B) < +\infty$. Define $E := B \cup \bigcup_{l=k+1}^m C_l$. Using Proposition 5.4 we have

$$0 \leq \mathcal{G}(E) = P(E) - \sum_{l=k+1}^m P(C_l) - J_0|E \cap F|.$$

Since $E \cap F = B \setminus C_j$ and $P(E) = P(C_j \cup B) + \sum_{l=k+1, l\neq j}^m P(C_l)$, we have

$$P(C_j \cup B) - J_0|B \setminus C_j| \geq P(C_j). \tag{5.27}$$

By substracting $J_0|C_j|$ to (5.27) we obtain (5.26). Then (5.25) is a consequence of Proposition 2.1.

Similarly, fix $x \in \partial C_0$ (resp., $x \in \partial C_i$ for some $i \in \{1, \dots, k\}$), and let A be an open neighborhood of x where ∂C_0 (resp., ∂C_i) can be written as a graph; we can assume that $\overline{A} \cap (\cup_{l \in \{1,\dots,m\}} C_l) = \emptyset$ (resp., $\overline{A} \subset C_0$, $\overline{A} \cap (\cup_{l \in \{1,\dots,m\}, l \neq i} C_i) = \emptyset$). Then

$$P(C_0) - J_0|C_0| \leq P(C_0 \setminus B) - J_0|C_0 \setminus B| \tag{5.28}$$

$$\text{(resp., } P(C_i) + J_0|C_i| \leq P(C_i \cup B) + J_0|C_i \cup B|) \tag{5.29}$$

for any Borel set B with $\overline{B} \subset A$. Indeed, define $E := (F \setminus B) \cup \bigcup_{j=k+1}^{m} C_j$. Using Proposition 5.4 and the equality $P(E) = P(C_0 \setminus B) + \sum_{i=1}^{k} P(C_i)$, we have

$$\begin{aligned} 0 \leq \mathcal{G}(E) &= P(E) - \sum_{j=k+1}^{m} P(C_j) - J_0|E \cap F| \\ &= P(C_0 \setminus B) - P(C_0) + \sum_{i=0}^{k} P(C_i) - \sum_{j=k+1}^{m} P(C_j) - J_0|E \cap F| \\ &= P(C_0 \setminus B) - P(C_0) + J_0|F| - J_0|E \cap F|, \end{aligned}$$

where in the last equality we have used the definition of J_0. We then get

$$P(C_0) - J_0|C_0| \leq P(C_0 \setminus B) - J_0(|E \cap F| + |C_0| - |F|) = P(C_0 \setminus B) - J_0|C_0 \setminus B|,$$

which is (5.28). Then (5.23) is a consequence of Proposition 2.1.

Eventually, in the case where $x \in \partial C_i$ for some $i \in \{1, \dots, k\}$, and A has been chosen as described above, we define again $E := (F \setminus B) \cup \left(\bigcup_{l=k+1}^{m} C_l\right)$. Then

$$\begin{aligned} 0 \leq \mathcal{G}(E) &= P(E) - \sum_{l=k+1}^{m} P(C_l) - J_0|E \cap F| \\ &= P(C_i \cup B) - P(C_i) + \sum_{l=0}^{k} P(C_l) - \sum_{l=k+1}^{m} P(C_l) - J_0|E \cap F| \\ &= P(C_i \cup B) - P(C_i) + J_0|F| - J_0|E \cap F|, \end{aligned}$$

which implies

$$\begin{aligned} P(C_i) + J_0|C_i| &\leq P(C_i \cup B) + J_0(|F| - |E \cap F| + |C_i|) \\ &= P(C_i \cup B) + J_0|C_i \cup B|, \end{aligned} \tag{5.30}$$

and, by Proposition 2.1, (5.24) follows.

Assume now that (a)–(c) hold. Notice that condition (b) implies (5.23), which, in turn, implies $J_0 > 0$. Observe also that, by (5.25), the sets $C_{k+1}, \dots, C_m$ are strictly convex.

Denote by $E_{\min} \in \mathcal{A}$ a solution of the minimum problem (5.7), with $E_{\min} \notin \{\emptyset, F\}$. By Theorem 5.3 and Remark 4, it is enough to prove that

$$\mathcal{F}_F(E_{\min}) \geq 0. \tag{5.31}$$

We can assume that $E_{\min}$ is connected, since the functional $\mathcal{F}_F$ is additive on connected components [5]. Recall that, by the definition of $\mathcal{A}$, the closure of (any connected component of) $E_{\min}$ must intersect ∂F.

Let Γ be a connected component of $F \cap \partial E_{\min}$, and let p, q be the endpoints of Γ, with $p \in \partial C_j$ and $q \in \partial C_i$, for some $i, j \in \{0, \dots, m\}$ (p not necessarily different from q). We recall that Γ meets tangentially ∂F (see conditions (5.2)–(5.4)) and, by assumption (b), Γ is contained in the boundary of an open ball $B \subseteq F \cup (\bigcup_{n=k+1}^{m} C_n)$ of radius $\frac{1}{J_0}$. We now divide the proof into three steps. We first show that p and q cannot belong both to the same ∂C_i when $i \le k$.

Step 1. If $i \le k$, then $i \neq j$.

Assume by contradiction that $i = j \le k$. Using assumption (b), by Lemma 5.6 (applied with $\Omega := F \cup (\bigcup_{l=k+1}^{m} C_l)$) it follows that p and q are the extrema of an arc $\gamma \subseteq \partial B \cap \partial C_i$ which spans an angle strictly less than π. Notice that $\partial B = \gamma \cup \Gamma$; moreover, recalling that the curvature (which is equal to $1/J_0$) of $E_{\min}$ inside F is positive, either $E_{\min} = B$ or $E_{\min} = B \setminus C_{\bar{j}}$ for some index $\bar{j} \ge k+1$. Observe that, in the latter case, $C_{\bar{j}} \subset\subset B$ and, by condition (c), there cannot be any other C_l, with $l \ge k+1$ and $l \neq \bar{j}$, with $C_l \subseteq B$. Let us consider a new set $E' := B' \setminus B$ if $E_{\min} = B$ (resp., $E' := B' \setminus C_{\bar{j}}$ if $E_{\min} = B \setminus C_{\bar{j}}$), where B' is a ball obtained by slightly translating B towards the interior of F, and slightly modifying its radius. By Remark 10 we have $\mathcal{F}_F(E') < \mathcal{F}_F(E_{\min})$, which contradicts the minimality of $E_{\min}$. We now show that either $i \le k$ and $j \ge k+1$ or vice versa.

Step 2. The cases $i, j \le k$ and $i, j \ge k+1$ cannot happen.

By assumption (c) and Step 1 it is clear that the case $i, j \le k$ cannot happen, nor can the case $i, j \ge k+1$ with $i \neq j$. We have to exclude the case $i = j \ge k+1$. Recalling that (5.25) implies the strict convexity of C_j, using (a) and (5.4), we have that $C_j \subseteq B$. Using again the strict convexity of C_j, Lemma 5.8 implies that $\partial C_j \cap \partial B$ is a connected arc which spans an angle strictly less than π. Hence we get a contradiction by slightly modifying $E_{\min}$ as in Step 1.

By Steps 1 and 2 we conclude that there exists an arc Γ of $F \cap \partial E_{\min}$ whose endpoints p, q satisfy $p \in \partial C_j$, $q \in \partial C_i$, and $i \in \{0, \dots, k\}$, $j \in \{k+1, \dots, m\}$.

In the following, we write C_i for $i \le k$, but we mean $R^2 \setminus \overline{C_0}$ when $i = 0$.

Let us call the inner (resp., outer) side of Γ the side of Γ inside (resp., outside) $E_{\min}$. Notice that from conditions (5.2)–(5.4) C_i cannot lie in the inner side of Γ and C_j cannot lie in the outer side of Γ. Moreover, since $J_0 > 0$ the inner (resp., outer) side of Γ is also the side of Γ inside (resp., outside) B.

Step 3. We have $B = E_{\min} \cup C_j$.

Let $p' \in \partial C_j$ be the endpoint of an arc $\Gamma' \subseteq F \cap \partial E_{\min}$. Then Γ' is contained in the boundary of an open ball $B' \subseteq F \cup \left(\bigcup_{l=k+1}^{m} C_l\right)$ of radius $\frac{1}{J_0}$. By assumption (c) Γ' cannot meet another set $C_{j'}$ with $j' \ge k+1$, $j' \neq j$. On the other hand, the above discussion implies $C_j \subseteq B'$. Let us suppose that the other endpoint q' of Γ' (different from p') belongs to $\partial C_{i'}$ for some $i' \le k$. Observe that $B' \cap C_{i'} = \emptyset$. If $i \neq i'$, then $B \neq B'$ (if $B = B' \ni \{q, q'\}$, then $\operatorname{dist}(C_i, C_{i'}) \le \frac{2}{J_0}$, a contradiction with assumption (c)). Since B and B' contain C_j, we have $B \cap B' \neq \emptyset$. Now, Γ is an arc of ∂B joining $p \in \overline{B} \cap \overline{B'}$ to $q \in \partial C_i \cap \overline{B}$, $q \notin \overline{B'}$, whereas Γ' is joining $p' \in \overline{B} \cap \overline{B'}$ to $q' \in \partial C_{i'} \cap \overline{B'}$, $q \notin \overline{B}$. It follows that either $\Gamma \cap \Gamma' \neq \emptyset$ or there exists another arc of $\partial B \cap \partial E_{\min} \cap F$ different from Γ intersecting Γ'; see Figure 5.1. Since these arcs intersect transversally, this contradicts the fact that $\partial E_{\min}$ is smooth. It follows that $i = i'$. Moreover, since $(B \cup B') \cap C_i = \emptyset$, for the same reason (i.e., the fact that $\partial E_{\min}$ is smooth) we also get $B = B'$.

The ball B cannot meet, nor contain, any other set $C_{i'}$ with $i' \le k$, $i \neq i'$, nor any other set $C_{j'}$ with $j' \ge k+1$, $j' \neq j$. Thus $B = E_{\min} \cup C_j$ (see Figure 5.2) and Step 3 is proved.

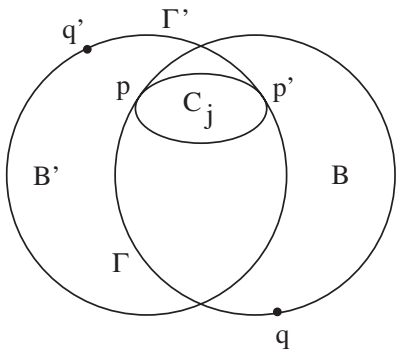

FIG. 5.1. *The two intersecting balls B and B'.*

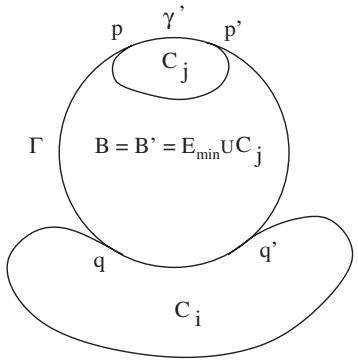

FIG. 5.2. *The minimizing set $E_{\min}$ $(i \geq 1)$.*

We now conclude the proof. Applying Proposition 5.9 with $K_1 = C_j$ and $K_0 = B$, we compute (see Figure 5.2)

$$\begin{aligned}\mathcal{F}_F(E_{\min}) &= P(E_{\min}, F) + \mathcal{H}^1(\partial C_i \cap \partial E_{\min}) - \mathcal{H}^1(\partial C_j \cap \partial E_{\min}) - J_0|E_{\min}| \\ &= \frac{2\pi}{J_0} - P(C_j) - J_0\left(\frac{\pi}{J_0^2} - |C_j|\right) = \frac{\pi}{J_0} - P(C_j) + J_0|C_j| \geq 0,\end{aligned}$$

which gives (5.31) and hence the thesis. □

PROPOSITION 5.11. *Let K_0, K_1 be two bounded open convex sets of R^2 with boundary of class $\mathcal{C}^{1,1}$ such that $\overline{K_1} \subseteq K_0$. Let $F := K_0 \setminus \overline{K_1}$. Let*

$$J := \frac{P(K_0) - P(K_1)}{|F|} > 0.$$

If

$$\operatorname{ess\,sup}_{\partial K_0} \kappa_{\partial K_0} \leq J, \tag{5.32}$$

$$\operatorname{ess\,inf}_{\partial K_1} \kappa_{\partial K_1} \geq J, \tag{5.33}$$

then there exists a vector field $z \in L^\infty(F, R^2)$ *with* $\|z\|_\infty \le 1$ *such that*

$$\begin{cases} -\operatorname{div} z = J & \text{in } \mathcal{D}'(F), \\ [z, \nu^F] = -1 & \mathcal{H}^1\text{-a.e. on } \partial K_0, \\ [z, \nu^F] = 1 & \mathcal{H}^1\text{-a.e. on } \partial K_1. \end{cases} \tag{5.34}$$

Remark 12. Proposition 5.11 admits a direct proof along the lines of [27]. Notice also that, thanks to Remark 9, Proposition 5.11 would be a consequence of Theorem 5.10 (in the case $k = 0$ and $m = 1$) if the strict inequality in (5.32) were valid.

Proof of Proposition 5.11. Let us prove that assumptions (a) and (b) of Theorem 5.10 hold for $F_\lambda := K_{0\lambda} \setminus \overline{K_{1\lambda}}$, where $K_{0\lambda} := (1+\lambda)K_0$, $K_{1\lambda} := (1-\lambda)K_1$, $\lambda > 0$ being small enough. We observe that $P(K_{0\lambda}) = (1+\lambda)P(K_0)$, $P(K_{1\lambda}) = (1-\lambda)P(K_1)$, $|K_{0\lambda}| = (1+\lambda)^2|K_0|$, and $|K_{1\lambda}| = (1-\lambda)^2|K_1|$; hence

$$J_\lambda := \frac{P(K_{0\lambda}) - P(K_{1\lambda})}{|F_\lambda|} = J + \frac{\lambda}{|F|}(P(K_0) + P(K_1) - 2J(|K_0| + |K_1|)) + o(\lambda).$$

Since

$$\operatorname*{ess\,sup}_{\partial K_{0\lambda}} \kappa_{\partial K_{0\lambda}} = \frac{1}{1+\lambda} \operatorname*{ess\,sup}_{\partial K_0} \kappa_{\partial K_0} \le \frac{1}{1+\lambda} J,$$

it suffices to prove that $\frac{1}{1+\lambda} J < J_\lambda$ to conclude that

$$\operatorname*{ess\,sup}_{\partial K_{0\lambda}} \kappa_{\partial K_{0\lambda}} < J_\lambda. \tag{5.35}$$

By Remark 8, this implies that $K_{0\lambda}$ satisfies the $\frac{1}{J_\lambda}$-ball condition. Now, $\frac{1}{1+\lambda} J < J_\lambda$ for λ small enough if and only if

$$2P(K_0)|K_1| < P(K_1)(|K_0| + |K_1|). \tag{5.36}$$

Since $\overline{K_1} \subset K_0$, using Proposition 5.9 and the isoperimetric inequality, we deduce

$$|K_0| + |K_1| > \frac{1}{2\pi} P(K_0)P(K_1) \ge 2\frac{P(K_0)|K_1|}{P(K_1)},$$

and we obtain (5.36), and therefore also (5.35).

Let us prove that condition (b) of Theorem 5.10 holds. Since

$$\operatorname*{ess\,inf}_{\partial K_{1\lambda}} \kappa_{\partial K_{1\lambda}} = \frac{1}{1-\lambda} \operatorname*{ess\,inf}_{\partial K_1} \kappa_{\partial K_1} \le \frac{1}{1-\lambda} J,$$

to conclude that

$$\operatorname*{ess\,sup}_{\partial K_{1\lambda}} \kappa_{\partial K_{1\lambda}} \ge J_\lambda, \tag{5.37}$$

it suffices to prove that $\frac{1}{1-\lambda} J \ge J_\lambda$. Now, $\frac{1}{1-\lambda} J > J_\lambda$ for λ small enough if and only if

$$2P(K_1)|K_0| < P(K_0)(|K_0| + |K_1|). \tag{5.38}$$

Again, since $\overline{K_1} \subseteq K_0$, using Proposition 5.9 and the isoperimetric inequality, we deduce

$$|K_0| + |K_1| > \frac{1}{2\pi} P(K_0)P(K_1) \ge 2\frac{P(K_1)|K_0|}{P(K_0)},$$

and we conclude that (5.38), and therefore also (5.37), holds.

By Theorem 5.10, there exists a vector field $z_\lambda \in L^\infty(F_\lambda, R^2)$ such that $\|z_\lambda\|_\infty \le 1$, satisfying

$$\begin{cases} -\operatorname{div} z_\lambda = J_\lambda & \text{in } \mathcal{D}'(F_\lambda), \\ [z_\lambda, \nu^{F_\lambda}] = -1 & \mathcal{H}^1\text{-a.e. on } \partial K_{0\lambda}, \\ [z_\lambda, \nu^{F_\lambda}] = 1 & \mathcal{H}^1\text{-a.e. on } \partial K_{1\lambda}. \end{cases}$$

Letting $\lambda \to 0^+$ we obtain a solution of (5.34). ∎

6. Solutions of div $z = 0$ in an unbounded domain. In this section we assume that $C_0 = R^2$, $k \ge 1$, we let $C_1, \dots, C_m$ be as in section 5, and we let $F := R^2 \setminus \bigcup_{i=1}^m \overline{C}_i$. We are concerned with the existence of a vector field $z : F \to R^2$ such that

$$(6.1)\quad z \in L^\infty(F, R^2), \quad \begin{cases} -\operatorname{div} z = 0 & \text{in } \mathcal{D}'(F), \\ \|z\|_\infty \le 1, & \\ [z, \nu^F] = -1 & \mathcal{H}^1\text{-a.e. on } \partial C_i,\ i \in \{1, \dots, k\}, \\ [z, \nu^F] = 1 & \mathcal{H}^1\text{-a.e. on } \partial C_j,\ j \in \{k+1, \dots, m\}. \end{cases}$$

THEOREM 6.1. *The following conditions are equivalent:*

(i) *Problem* (6.1) *has a solution.*

(ii) *We have*

$$(6.2)\qquad 0 \le \int_F |Dw| + \sum_{i=1}^k \int_{\partial C_i} w - \sum_{j=k+1}^m \int_{\partial C_j} w \qquad \forall w \in BV(F).$$

(iii) *For any $E \subseteq F$ of finite perimeter, we have*

$$(6.3)\qquad P(E, F) \ge \left| \sum_{j=k+1}^m \mathcal{H}^1(\partial^* E \cap \partial C_j) - \sum_{i=1}^k \mathcal{H}^1(\partial^* E \cap \partial C_i) \right|.$$

(iv) *Let E_1 be a solution of the variational problem*

$$(6.4)\qquad \min \left\{ P(E) : \bigcup_{j=k+1}^m C_j \subseteq E \subseteq R^2 \setminus \bigcup_{i=1}^k C_i \right\}.$$

Then we have

$$(6.5)\qquad P(E_1) = \sum_{j=k+1}^m P(C_j).$$

Let E_2 be a solution of the variational problem

$$(6.6)\qquad \min \left\{ P(E) : \bigcup_{i=1}^k C_i \subseteq E \subseteq R^2 \setminus \bigcup_{j=k+1}^m C_j \right\}.$$

Then we have

$$(6.7)\qquad P(E_2) = \sum_{i=1}^k P(C_i).$$

Remark 13. Notice that (iv) implies that each C_l is a convex set. Moreover, since any minimizer of problems (6.4) and (6.6) has boundary (lying inside F) made of a finite number of segments which intersect tangentially ∂F (and there are only a finite number of such segments), the number of such minimizers is finite. Finally, conditions (6.5) and (6.7) are essentially distance conditions between sets C_i of the same type; for example, they are satisfied if $\mathrm{dist}(\partial C_i, \partial C_j) > P(C_l)$ for any $(i,j,l) \in \{1,\dots,k\}^3 \cup \{k+1,\dots,m\}^3$, $i \neq j$.

Proof. We divide the proof into four steps.

Step 1. Let $f \in L^2(F)$, $g \in L^\infty(\partial F)$, $\lambda > 0$. The following hold:

(a) Assume that $\|g\|_\infty < 1$. The function u is the solution of

$$\min_{w \in BV(F)} Q(w), \qquad Q(w) := \int_F |Dw| + \frac{1}{2\lambda} \int_F (w-f)^2 \, dx - \int_{\partial F} gw \, d\mathcal{H}^1 \tag{6.8}$$

if and only if there exists $z \in X_2(F)$ with $\|z\|_\infty \leq 1$ satisfying $(z, Du) = |Du|$ as measures in F, $[z, \nu^F] = g$ $\mathcal{H}^1$-almost everywhere on ∂F and $-\lambda \mathrm{div} z = f - u$ in $\mathcal{D}'(F)$.

(b) The function $u \equiv 0$ is the solution of (6.8) if and only if

$$\int_F |Dw| \geq \frac{1}{\lambda} \int_F wf \, dx - \int_{\partial F} gw \qquad \forall w \in BV(F).$$

Let us prove both assertions. Let $R > 0$ be such that $R^2 \setminus F \subset\subset B_R = B_R(0)$. We consider the functional

$$Q_R(w) := \int_{B_R \cap F} |Dw| + \frac{1}{2\lambda} \int_{B_R \cap F} (w-f)^2 \, dx - \int_{\partial F} gw \, d\mathcal{H}^1 \tag{6.9}$$

defined for $w \in BV(B_R \cap F)$. Now, since $\|g\|_\infty < 1$ and ∂F is of class $\mathcal{C}^{1,1}$, using the results of Giusti [28], we know that the convex functional Q_R is lower semicontinuous and proper, and it attains its infimum in $BV(B_R \cap F)$. Let $w_n \to w$ in $L^2(B_R \cap F)$. Then $Q_R(w) \leq \liminf_n Q_R(w_n) \leq \liminf_n Q(w_n)$. Since this is true for all $R > 0$, we deduce that $Q(w) \leq \liminf_n Q(w_n)$. Thus, Q is convex, lower semicontinuous, and proper. As we shall note below, Q attains its infimum in $BV(F)$. Hence $u = \mathrm{argmin}\, Q$ if and only if $0 \in \partial Q(u)$.

Now, we define the operator $\mathcal{A}'_g$ in $L^2(F) \times L^2(F)$ as follows: $(w,v) \in \mathcal{A}'_g$ if and only if $w \in BV(F)$, $v \in L^2(F)$, and there is a vector field $z \in L^\infty(F, R^2)$ with $\|z\|_\infty \leq 1$ such that $(z, Dw) = |Dw|$ and $-\mathrm{div}\, z = v$ in $\mathcal{D}'(F)$, $[z, \nu^F] = g$ $\mathcal{H}^1$-almost everywhere on ∂F. We claim that the operator $\mathcal{A}'_g$ is maximal monotone. The monotonicity of $\mathcal{A}'_g$ follows by an integration by parts. To prove the maximal monotonicity we have to solve the equation

$$f \in u + \mathcal{A}'_g u \qquad \forall f \in L^2(F). \tag{6.10}$$

First, we assume that $f \in L^p(F)$ for any $p \in [1, \infty]$. Let us approximate (6.10) by

$$\begin{cases} u - \mathrm{div}\, z = f & \text{in } \mathcal{D}'(B_R \cap F), \\ [z, \nu^{B_R \cap F}] = g & \mathcal{H}^1\text{-a.e. in } \partial F, \\ [z, \nu^{B_R \cap F}] = 0 & \mathcal{H}^1\text{-a.e. in } \partial B_R, \end{cases} \tag{6.11}$$

where $z \in L^\infty(B_R \cap F, R^2)$ is such that $\|z\|_\infty \leq 1$ and $(z, Du) = |Du|$. Then, by Step 1 of the proof of Theorem 5.3, equation (6.11) has a unique solution u_R. Let z_R

denote the associated vector field. Let us comment on the basic estimates required to pass to the limit as $R \to \infty$.

(i) L^2 and bounded variation estimates on u_R: multiplying (6.11) by u_R, after integration by parts, we get

$$\int_{B_R\cap F} u_R^2 + \int_{B_R\cap F} |Du_R| = \int_{B_R\cap F} f u_R + \int_{\partial F} g u_R. \tag{6.12}$$

Now, using [28, Lemma 2.2], there exists $\epsilon_0 > 0$ such that, for each $\delta > 0$, there is $c(\delta) > 0$ such that

$$\left|\int_{\partial F} gw\right| \le \Big(1 - \frac{\epsilon_0}{2}\Big) \int_{S_\delta} |Dw| + c(\delta) \int_{S_\delta} |w| \qquad \forall w \in BV(B_R \cap F), \tag{6.13}$$

where $S_\delta := \{x \in B_R \cap F : \operatorname{dist}(x, \partial F) < \delta\}$, where the constant $c(\delta)$ does not depend on $R > 0$. Using (6.13) in (6.12) we obtain the estimate

$$\frac{1}{4} \int_{B_R\cap F} u_R^2 + \epsilon_0 \int_{B_R\cap F} |Du_R| \le \frac{1}{2}\|f\|_2^2 + C|S_\delta|.$$

Thus, by extracting a subsequence, if necessary, we may assume that $u_R \to u$ in $L^p_{\rm loc}$ for any $1 \le p < 2$ and weakly in $L^2(F)$ where $u \in L^2(F)$ and $\int_F |Du| < \infty$.

Let us mention that, as a consequence of (6.13), if $Q(u_n)$ is bounded, we obtain that $\int_F |u_n|^2$ and $\int_F |Du_n|$ are bounded and, therefore, Q attains its infimum.

(ii) L^p estimate on u_R: let $\eta_p : R \to R$ be a smooth function such that $\eta_p'(r) > 0$ for all $r \in R$, $\eta_p(0) = 0$, and $\operatorname{sign}(r)\eta_p(r)$ behaves as $|r|^{p-1}$ as $r \to \infty$. We multiply (6.11) by $\eta_p(u_R)$. Integrating by parts and using (6.13), we obtain

$$\int_{B_R\cap F} u_R \eta_p(u_R) \le \int_{B_R\cap F} |f||\eta_p(u_R)| + c(\delta) \int_{S_\delta} |\eta_p(u_R)|. \tag{6.14}$$

Let $p = 1$, and assume that $|\eta_1(r)| \le 1$ for any $r \in R$. We obtain

$$\int_{B_R\cap F} u_R \eta_1(u_R) \le \int_{B_R\cap F} |f| + c(\delta)|S_\delta|. \tag{6.15}$$

Take a sequence $\eta_{1,n}(r)$ such that $\eta_{1,n}(r) \to \operatorname{sign}(r)$ for any $r \ne 0$. Using Fatou's theorem we deduce that

$$\int_{B_R\cap F} |u_R| \le \int_{B_R\cap F} |f| + c(\delta)|S_\delta|.$$

Assume that u_R is bounded in L^q. Using $p = q$ in (6.14) and proceeding in the same way, we deduce that u_R is bounded in L^{q+1}. This implies that u_R is bounded in L^p for all $p < \infty$. Thus $u \in L^p(F)$ for any $1 \le p < \infty$.

Now, let $R > M > 0$, where M is such that all sets C_i are contained in $B_{M/4}(0)$. Let $\varphi \in W^{1,\infty}(R^2)$ be such that $\varphi = 0$ on $B_{M/2}(0)$, $\varphi = 1$ outside $B_M(0)$, and it increases linearly along the rays from 0 to 1 in $B_M(0) \setminus B_{M/2}(0)$. We multiply (6.11) by $u_R\varphi^2$ and integrate by parts to obtain

$$\int_{B_R\cap F} u_R^2\varphi^2 + \int_{B_R\cap F} |Du_R|\varphi^2 = \int_{B_R\cap F} f u_R\varphi^2 - \int_{B_R\cap F} u_R z_R \cdot \nabla(\varphi^2).$$

Hence

$$\int_{B_R\cap F} u_R^2\varphi^2 \le \int_{B_R\cap F} |f||u_R|\varphi^2 + \int_{B_R\cap F} |u_R|\varphi|\nabla\varphi| \le \frac{1}{2}\int_{B_R\cap F} |f|^2\varphi^2 + \frac{1}{2}\int_{B_R\cap F} |u_R|^2\varphi^2 + \|u_R\varphi\|_{3/2}\|\nabla\varphi\|_3.$$

As $|\nabla\varphi| \le \frac{2}{M}$ we have

$$\|\nabla\varphi\|_3 \le \frac{2}{M}\left(\frac{3}{4}\pi M^2\right)^{1/3} \le \frac{C}{M^{1/3}}.$$

Since $\|u_R\varphi\|_{3/2}$ is bounded independently of R and M, we have

$$\int_{B_R\cap F} u_R^2\varphi^2 \le C\int_{B_R\cap F} |f|^2\varphi^2 + \frac{C}{M^{1/3}}.$$

Thus, given $\epsilon > 0$ we find M large enough so that

$$\int_{B_R\cap F} u_R^2\varphi^2 \le \epsilon$$

for any $R > M$. Assume that u_R is extended by 0 outside B_R. Thus u_R is equi-integrable near infinity. Thus, to prove that $u_R \to u$ in $L^2(F)$ it is sufficient to prove that $u_R \to u$ in $L^2_{\rm loc}(F)$. For that, let $\varphi \in C_0^\infty(R^2)$. Then

$$\int_F |u_R - u|^2\varphi^2 \le \left(\int_F |u_R-u|^3\varphi^2\right)^{1/2}\left(\int_F |u_R-u|\varphi^2\right)^{1/2} \to 0 \quad \text{as } R\to\infty,$$

since the first integral is bounded independently of R and the second tends to 0 as $R\to\infty$.

The two previous estimates imply that we may extract a subsequence u_R converging in $L^2(F)$ to some function $u \in BV(F)$. Moreover, we may assume that $z_R \to z$ weakly* in $L^\infty(F, R^2)$. Letting $R\to\infty$ in (6.11) we have

$$u - \operatorname{div} z = f \quad \text{in } \mathcal{D}'(F). \tag{6.16}$$

We still have to prove that $(z, Du) = |Du|$ and $[z, \nu^F] = g$.

Let φ be a smooth function in F, continuous up to ∂F and vanishing for large values of $|x|$. We multiply (6.11) by φ and integrate by parts to obtain

$$\int_{B_R\cap F} u_R\varphi + \int_{B_R\cap F} z_R\cdot\nabla\varphi - \int_{\partial F}[z_R, \nu^{B_R\cap F}]\varphi = \int_{B_R\cap F} f\varphi. \tag{6.17}$$

Letting $R\to\infty$ and using that $[z_R, \nu^{B_R\cap F}] = g$, we obtain

$$\int_F u\varphi + \int_F z\cdot\nabla\varphi - \int_F g\varphi = \int_F f\varphi. \tag{6.18}$$

Integrating by parts the second term of the above equality, we get

$$\int_F u\varphi - \int_F \operatorname{div} z\,\varphi + \int_{\partial F}([z,\nu^F] - g)\varphi = \int_F f\varphi. \tag{6.19}$$

Using (6.16) it follows that $\int_{\partial F}([z,\nu^F]-g)\varphi = 0$ for any φ. This implies that $[z,\nu^F]=g$ on ∂F. To prove that $(z,Du)=|Du|$, we observe that from the lower semicontinuity of Q and the convergence $\int_{B_R\cap F}(u_R-f)^2\,dx \to \int_F (u-f)^2\,dx$ as $R\to\infty$, we have

$$\begin{aligned}\int_F |Du| - \int_{\partial F} gu &\le \liminf_R \left(\int_{B_R\cap F}|Du_R| - \int_{\partial F} gu_R\right)\\ &= \liminf_R\left(\int_{B_R\cap F}(z_R,Du_R) - \int_{\partial F} gu_R\right) = \liminf_R -\int_{B_R\cap F}\operatorname{div} z_R\, u_R = -\int_F \operatorname{div} z\, u\\ &= \int_F (z,Du) - \int_{\partial F} gu \le \int_F |Du| - \int_{\partial F} gu.\end{aligned}$$

We conclude that $\int_F (z,Du) = \int_F |Du|$.

We have proved that there is a solution of (6.10) for each $f\in L^\infty(F)\cap L^2(F)$. Our next purpose is to prove that the operator $\mathcal{A}'_g$ is closed. As a consequence we obtain that (6.10) has a solution for any $f\in L^2(B_R\cap F)$. To prove the closedness of $\mathcal{A}'_g$, let $(u_n,v_n)\in\mathcal{A}'_g$ be such that $(u_n,v_n)\to(u,v)$ in $L^2(F)\times L^2(F)$. Then there is a vector field $z_n\in L^\infty(F,R^2)$ with $\|z_n\|_\infty\le 1$ such that $v_n=-\operatorname{div} z_n$, $(z_n,Du_n)=|Du_n|$ and $[z_n,\nu^F]=g$. Up to a subsequence, we may assume that $z_n\to z$ weakly* in $L^\infty(F,R^2)$ with $\|z\|_\infty\le 1$. Since $v_n=-\operatorname{div} z_n\to -\operatorname{div} z$ in $\mathcal{D}'(F)$, we have $v=-\operatorname{div} z$. The proofs of the facts $[z,\nu^F]=g$ and $(z,Du)=|Du|$ follow the same arguments as the corresponding proofs in Theorem 5.3, and we shall omit the details. We conclude that $\mathcal{A}'_g$ is closed.

Since $\mathcal{A}'_g\subseteq\partial Q$ and both are maximal monotone, we conclude that $\mathcal{A}'_g=\partial Q$. This proves (a).

The proof of (b) follows along the same lines as the proof of Step 2 in Theorem 5.3.

Step 2. (i) $\Longleftrightarrow$ (ii). Note that, as before, we may replace the condition "$\forall w\in BV(F)$" by "$\forall w\in BV(F)$ such that $w\ge 0$ or $w\le 0$."

Suppose that (6.1) has a solution z. Let $w\in BV(F)$. Multiplying (6.1) by w and integrating by parts, we obtain (6.2).

Assume now that (6.2) holds for any $w\in BV(F)$. Multiplying (6.2) by $(1-\epsilon)$, we deduce that

$$0\le \int_F |Dw| + (1-\epsilon)\sum_{i=1}^k \int_{\partial C_i} w - (1-\epsilon)\sum_{j=k+1}^m \int_{\partial C_j} w \qquad \forall w\in BV(F).$$

Using Step 1(b), we deduce that $u\equiv 0$ is a solution of (6.8) with $f=0$, and $g\equiv 1-\epsilon$ on ∂C_j and $g\equiv -(1-\epsilon)$ on ∂C_i for all $\epsilon>0$. Then by Step 2(a), we know that there exists a solution $\xi_\epsilon\in L^\infty(F,R^2)$ such that $\|\xi_\epsilon\|_\infty\le 1$, $-\operatorname{div}\xi_\epsilon=0$, $[\xi_\epsilon,\nu^F]=g$. Letting $\epsilon\to 0$, we find a vector field z satisfying (6.1).

The equivalence between (ii) and (iii) can be proved in the same manner as the equivalence between (b) and (c) in Theorem 5.3 was, and we shall omit the details.

Step 3. (iii)$\Rightarrow$(iv). Let $X:=E_1\setminus\bigcup_{j=k+1}^m C_j$. Using (iii) we have

$$\sum_{j=k+1}^m \mathcal{H}^1(\partial^* X\cap\partial C_j) - \sum_{i=1}^k \mathcal{H}^1(\partial^* X\cap\partial C_i)\le P(X,F). \tag{6.20}$$

Using Lemma 2.2, we have

$$P(E_1)=P\left(X\cup\bigcup_{j=k+1}^m C_j\right)=P(X)+\sum_{j=k+1}^m P(C_j)-2\mathcal{H}^1\left(\partial^* X\cap\left(\bigcup_{j=k+1}^m \partial C_j\right)\right).$$

Then, using (6.20), we have

$$\begin{aligned}
P(E_1) &= P(X) + \sum_{j=k+1}^{m} P(C_j) - 2 \sum_{j=k+1}^{m} \mathcal{H}^1(\partial^* X \cap \partial C_j) \\
&= P(X,F) + \sum_{j=k+1}^{m} \mathcal{H}^1(\partial^* X \cap \partial C_j) + \sum_{i=1}^{k} \mathcal{H}^1(\partial^* X \cap \partial C_i) \\
&\quad + \sum_{j=k+1}^{m} P(C_j) - 2 \sum_{j=k+1}^{m} \mathcal{H}^1(\partial^* X \cap \partial C_j) \\
&= P(X,F) + \sum_{i=1}^{k} \mathcal{H}^1(\partial^* X \cap \partial C_i) + \sum_{j=k+1}^{m} P(C_j) - \sum_{j=k+1}^{m} \mathcal{H}^1(\partial^* X \cap \partial C_j) \\
&\geq \sum_{j=k+1}^{m} P(C_j).
\end{aligned}$$

The proof for the set E_2 is analogous.

Step 4. (iv)⇒(iii). Let $X \subseteq F$ be a set of finite perimeter. Let E_1 be a minimizer of (6.4) and set $D := \bigcup_{j=k+1}^{m} C_j$. Using (6.5) and the minimality of E_1, we have

$$\sum_{j=k+1}^{m} P(C_j) = P(E_1) \leq P\left(X \cup D\right). \tag{6.21}$$

Using Lemma 2.2 and (6.21), we have

$$\begin{aligned}
P\left(X \cup D\right) &= P(X) + P(D) - 2\mathcal{H}^1(\partial D \cap \partial^* X) \\
&\leq P(X) + P(X \cup D) - 2\mathcal{H}^1(\partial D \cap \partial^* X).
\end{aligned}$$

Hence

$$\begin{aligned}
2 \sum_{j=k+1}^{m} \mathcal{H}^1(\partial^* X \cap \partial C_j) \leq P(X) = P(X,F) &+ \sum_{i=1}^{k} \mathcal{H}^1(\partial^* X \cap \partial C_i) \\
&+ \sum_{j=k+1}^{m} \mathcal{H}^1(\partial^* X \cap \partial C_j).
\end{aligned}$$

We then have

$$\sum_{j=k+1}^{m} \mathcal{H}^1(\partial^* X \cap \partial C_j) \leq P(X,F) + \sum_{i=1}^{k} \mathcal{H}^1(\partial^* X \cap \partial C_i).$$

The other inequality follows by considering the set E_2 and using condition (6.6) instead of (6.4). □

7. Examples of solutions of the eigenvalue problem (1.1). Let us give an example of how, by pasting the solutions of problems (5.5) and (6.1), we can construct solutions of the eigenvalue problem (1.1).

Let C_i, $i = 1, \dots, m$, $1 \le k \le m$, be a family of convex sets of class $\mathcal{C}^{1,1}$ satisfying the conditions in section 5. For each $i \in \{1, \dots, m\}$ let us consider $C_{i1}, C_{i2}, \dots, C_{im_i}$ open bounded sets with boundary of class $\mathcal{C}^{1,1}$ with the following properties:

- $\overline{C_{ij}} \subset C_i$ for any $j \in \{1, \dots, m_i\}$;
- $\overline{C_{ij}} \cap \overline{C_{ij'}} = \emptyset$ for any $j, j' \in \{1, \dots, m_i\}$, $j \neq j'$.

For $i \in \{1, \dots, m\}$ we define

$$F_i := C_i \setminus \bigcup_{j=1}^{m_i} \overline{C_{ij}}, \qquad J_i := \frac{\sum_{j=0}^{k_i} P(C_{ij}) - \sum_{j=k_i+1}^{m_i} P(C_{ij})}{|F_i|},$$

where $k_i \in \{1, \dots, m_i\}$ are given. Assume that

(a) $\operatorname{ess\,inf}_{\partial C_{ij}} \kappa_{\partial C_{ij}} \ge J_i$, $i \in \{1, \dots, m\}$, $\quad j \in \{k_i + 1, \dots, m_i\}$;

(b) $F_i \cup (\bigcup_{j=k_i+1}^{m_i} \overline{C_{ij}})$ satisfies the $\frac{1}{J_i}$-ball condition for any $i \in \{1, \dots, m\}$;

(c) $\operatorname{dist}(\partial C_{ij}, \partial C_{ij'}) > \frac{2}{J_i}$, $i \in \{1, \dots, m\}$, $(j, j') \in \{0, \dots, k_i\}^2 \cup \{k_i+1, \dots, m_i\}^2$, $j \neq j'$, where we have denoted $C_{i0} = C_i$;

(d)

$$\operatorname*{ess\,sup}_{\partial C_{ij}} \kappa_{\partial C_{ij}} \le \frac{P(C_{ij})}{|C_{ij}|} =: J_{ij}, \qquad i \in \{1, \dots, m\},\ j \in \{1, \dots, m_i\}.$$

Notice that $J_i > 0$, since (b) implies $\operatorname{ess\,sup}_{\partial C_{i0}} \kappa_{\partial C_{i0}} \le J_i$, and also

$$\operatorname*{ess\,inf}_{\partial C_{ij}} \kappa_{\partial C_{ij}} \ge -J_i, \qquad j \in \{1, \dots, k_i\}.$$

Using Theorems 5.10 and 6.1, together with [12, Theorem 4], we have the existence of vector fields $\xi_{\text{ext}} \in L^\infty(R^2 \setminus \bigcup_{i=1}^m C_i)$, $\xi_i \in L^\infty(F_i)$, $\xi_{ij} \in L^\infty(C_{ij})$, such that $\|\xi_{\text{ext}}\|_\infty \le 1$, $\|\xi_i\|_\infty \le 1$, $\|\xi_{ij}\|_\infty \le 1$, $i = 1, \dots, m$, $j = 1, \dots, m_i$, satisfying

$$\text{(7.1)} \qquad \begin{cases} -\operatorname{div} \xi_{\text{ext}} = 0 & \text{on } R^2 \setminus \bigcup_{i=1}^m C_i, \\ [\xi_{\text{ext}}, \nu^{R^2 \setminus \cup_{i=1}^m C_i}] = -1 & \mathcal{H}^1\text{-a.e. on } \partial C_i,\ i \in \{1, \dots, k\}, \\ [\xi_{\text{ext}}, \nu^{R^2 \setminus \cup_{i=1}^m C_i}] = 1 & \mathcal{H}^1\text{-a.e. on } \partial C_j,\ j \in \{k+1, \dots, m\}, \end{cases}$$

$$\text{(7.2)} \qquad \begin{cases} -\operatorname{div} \xi_i = J_i & \text{on } F_i, \\ [\xi_i, \nu^{F_i}] = -1 & \mathcal{H}^1\text{-a.e. on } \partial C_{ij},\ j \in \{0, \dots, k_i\}, \\ [\xi_i, \nu^{F_i}] = 1 & \mathcal{H}^1\text{-a.e. on } \partial C_{ij},\ j \in \{k_i + 1, \dots, m_i\}, \end{cases} \qquad i \in \{1, \dots, m\},$$

$$\text{(7.3)} \qquad \begin{cases} -\operatorname{div} \xi_{ij} = \dfrac{P(C_{ij})}{|C_{ij}|} & \text{on } C_{ij}, \\ [\xi_{ij}, \nu^{C_{ij}}] = -1 & \mathcal{H}^1\text{-a.e. on } \partial C_{ij}, \end{cases} \qquad i \in \{1, \dots, m\},\ j \in \{1, \dots, m_i\}.$$

Now, we may paste together these vector fields to define $\xi \in L^\infty(R^2, R^2)$, $\|\xi\|_\infty \le 1$, by

$$\xi := \begin{cases} \xi_{\text{ext}} & \text{on } R^2 \setminus \bigcup_{i=1}^m C_i, \\ -\xi_i & \text{on } F_i,\ i = 1, \dots, k, \\ \xi_i & \text{on } F_i,\ i = k+1, \dots, m, \\ \xi_{ij} & \text{on } C_{ij},\ i = 1, \dots, k,\ j = 1, \dots, k_i, \\ -\xi_{ij} & \text{on } C_{ij},\ i = 1, \dots, k,\ j = k_i + 1, \dots, m_i, \\ -\xi_{ij} & \text{on } C_{ij},\ i = k+1, \dots, m,\ j = 1, \dots, k_i, \\ \xi_{ij} & \text{on } C_{ij},\ i = k+1, \dots, m,\ j = k_i + 1, \dots, m_i, \end{cases}$$

satisfying

$$-\operatorname{div}\xi = \begin{cases} 0 & \text{on } R^2 \setminus \bigcup_{i=1}^m C_i, \\ -J_i & \text{on } F_i,\ i=1,\dots,k, \\ J_i & \text{on } F_i,\ i=k+1,\dots,m, \\ J_{ij} & \text{on } C_{ij},\ i=1,\dots,k,\ j=1,\dots,k_i, \\ -J_{ij} & \text{on } C_{ij},\ i=1,\dots,k,\ j=k_i+1,\dots,m_i, \\ -J_{ij} & \text{on } C_{ij},\ i=k+1,\dots,m,\ j=1,\dots,k_i, \\ J_{ij} & \text{on } C_{ij},\ i=k+1,\dots,m,\ j=k_i+1,\dots,m_i. \end{cases}$$

Thus, if we define

$$\begin{aligned} u := &-\sum_{i=1}^k J_i\chi_{F_i} + \sum_{i=k+1}^m J_i\chi_{F_i} + \sum_{i=1}^k\sum_{j=1}^{k_i} J_{ij}\chi_{C_{ij}} - \sum_{i=1}^k\sum_{j=k_i+1}^{m_i} J_{ij}\chi_{C_{ij}} \\ &- \sum_{i=k+1}^m\sum_{j=1}^{k_i} J_{ij}\chi_{C_{ij}} + \sum_{i=k+1}^m\sum_{j=k_i+1}^{m_i} J_{ij}\chi_{C_{ij}}, \end{aligned}$$

then u is a solution of (1.1). Therefore, by pasting solutions of problems like (7.1), (7.2), (7.3), we may construct solutions of (1.1).

8. Some explicit solutions of the denoising problem. The previous results allow us to explicitly compute the minimum of the denoising problem (1.8) for some data $f \in L^2(R^2)$. Let us recall that a vector field $z \in X_2(R^2)$ with $\|z\|_\infty \le 1$ satisfying

$$-\operatorname{div} z = F \in L^2(R^2)$$

exists if and only if [31, 12]

$$\|F\|_* := \sup\left\{\left|\int_{R^2} Fv\ dx\right| : v \in BV(R^2),\ \int_{R^2}|Dv| \le 1\right\} \le 1.$$

PROPOSITION 8.1. *Let $u_i \in BV(R^2)$, $u_i \ge 0$, be such that $u_i \wedge u_j = 0$, $i,j \in \{1,\dots,m\}$, $i \ne j$. Assume that u_i and $\sum_{i=1}^m u_i$ are solutions of the eigenvalue problem (1.1), $i \in \{1,\dots,m\}$. Let $b_i \in R$, $i=1,\dots,m$, and $f := \sum_{i=1}^m b_i u_i$. Also let $\lambda > 0$. Then the solution u of the variational problem (1.8) is $u := \sum_{i=1}^m \operatorname{sign}(b_i)(|b_i|-\lambda)^+ u_i$.*

Observe that if (*) $\sum_{i=1}^m u_i$ is a solution of (1.1), then (**) $\|\sum_{i=1}^m u_i\|_* \le 1$. Notice that, using (8.2) below, it is easy to prove that both conditions (*) and (**) are, indeed, equivalent.

Proof. Under our assumptions we have $u_i \in BV(R^2) \subset L^2(R^2)$, $i=1,\dots,m$, and hence $f \in L^2(R^2)$. Recall that a function $u \in BV(R^2)$ is the solution of (1.8) if and only if u is the solution of

$$u - \lambda\operatorname{div}\left(\frac{Du}{|Du|}\right) = f. \tag{8.1}$$

Observe that since each u_i is a solution of (1.1), multiplying (1.1) by u_i and integrating by parts, we obtain

$$\int_{R^2} u_i^2\,dx = \int_{R^2}|Du_i|. \tag{8.2}$$

Let us prove that $u = \sum_{i=1}^m \operatorname{sign}(b_i)(|b_i| - \lambda)^+ u_i$ is the solution of (8.1). Let $I_\lambda := \{i \in \{1, \dots, m\} : |b_i| \geq \lambda\}$, $H_\lambda := \{i \in \{1, \dots, m\} : |b_i| < \lambda\}$. Since, in this case,

$$f - u = \lambda \sum_{i \in I_\lambda} \operatorname{sign}(b_i) u_i + \sum_{i \in H_\lambda} b_i u_i,$$

to prove that u is a solution of (8.1) we have to construct a vector field $\xi \in L^\infty(R^2; R^2)$ with $\|\xi\|_\infty \leq 1$, such that

$$-\operatorname{div} \xi = \sum_{i \in I_\lambda} \operatorname{sign}(b_i) u_i + \sum_{i \in H_\lambda} \frac{b_i}{\lambda} u_i \tag{8.3}$$

and $(\xi, Du) = |Du|$. Let $F \in L^2(R^2)$ denote the right-hand side of (8.3), and let $F^+ := \sup(F, 0)$, $F^- := \sup(-F, 0)$. Let us prove that $\|F\|_* \leq 1$. In order to prove this, we let $v \in BV(R^2)$. Since

$$\int_{R^2} Fv \, dx \leq \int_{R^2} (F^+ v^+ + F^- v^-) \, dx$$

and $\int_{R^2} |Dv| = \int_{R^2} |Dv^+| + \int_{R^2} |Dv^-|$, the inequality $\int_{R^2} Fv \, dx \leq \int_{R^2} |Dv|$ follows if we prove that

$$\int_{R^2} F^+ v^+ \, dx \leq \int_{R^2} |Dv^+| \quad \text{and} \quad \int_{R^2} F^- v^- \, dx \leq \int_{R^2} |Dv^-|.$$

Thus, without loss of generality, we may assume that $F \geq 0$ (i.e., all b_i appearing in the definition of F are nonnegative) and $v \geq 0$. Then, using that $\frac{b_i}{\lambda} \leq 1$ for any $i \in H_\lambda$, we have that

$$0 \leq F \leq \sum_{i=1}^m u_i.$$

Since, by assumption, $\| \sum_{i=1}^m u_i \|_* \leq 1$, we have

$$\int_{R^2} Fv \, dx \leq \int_{R^2} \sum_{i=1}^m u_i v \, dx \leq \int_{R^2} |Dv|.$$

Therefore $\|F\|_* \leq 1$. Thus, there is a vector field $\xi \in L^\infty(R^2; R^2)$ such that $\|\xi\|_\infty \leq 1$, satisfying (8.3).

As $(|b_i| - \lambda)^+ = 0$ for all $i \in H_\lambda$, we have

$$\int_{R^2} |Du| = \sum_{i \in I_\lambda} (|b_i| - \lambda) \int_{R^2} |Du_i|.$$

Since $u_i \wedge u_j = 0$ for any $i, j \in \{1, \dots, m\}$, $i \neq j$, then $Fu = \sum_{i \in I_\lambda} (|b_i| - \lambda) u_i^2$, and we have

$$\int_{R^2} (\xi, Du) = -\int_{R^2} \operatorname{div} \xi \, u \, dx = \int_{R^2} Fu \, dx = \sum_{i \in I_\lambda} (|b_i| - \lambda) \int_{R^2} u_i^2 \, dx;$$

applying (8.2) we obtain

$$\int_{R^2} (\xi, Du) = \sum_{i \in I_\lambda} (|b_i| - \lambda) \int_{R^2} |Du_i| \, dx = \int_{R^2} |Du|,$$

which in turn implies that $(\xi, Du) = |Du|$, since $\|\xi\|_\infty \leq 1$. □

Acknowledgment. We would like to thank the anonymous reviewers for pointing out the necessity to assume strict convexity in Lemma 5.8.

REFERENCES

[1] F. Alter, V. Caselles, and A. Chambolle, *Evolution of convex sets in the plane by the minimizing total variation flow*, Interfaces Free Bound., to appear.
[2] F. Alter, V. Caselles, and A. Chambolle, *A Characterization of Convex Calibrable Sets in R^N*, Preprint, 2003.
[3] L. Alvarez, Y. Gousseau, and J. M. Morel, *The size of objects in natural images*, Adv. in Imaging and Electron Phys., 111 (1999), pp. 167–242.
[4] L. Ambrosio, *Corso introduttivo alla teoria geometrica della misura ed alle supefici minime*, Scuola Normale Superiore, Pisa, 1997.
[5] L. Ambrosio, V. Caselles, S. Masnou, and J.-M. Morel, *Connected components of sets of finite perimeter and applications to image processing*, European J. Appl. Math., 3 (2001), pp. 39–92.
[6] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, Oxford Math. Monogr., Oxford University Press, New York, 2000.
[7] L. Ambrosio and E. Paolini, *Partial regularity for quasiminimizers of perimeter*, Ricerche Mat., 48 (1998), pp. 167–186.
[8] F. Andreu, C. Ballester, V. Caselles, and J. M. Mazón, *Minimizing total variational flow*, Differential Integral Equations, 4 (2001), pp. 321–360.
[9] F. Andreu, V. Caselles, J. I. Diaz, and J. M. Mazón, *Qualitative properties of the total variation flow*, J. Funct. Anal., 188 (2002), pp. 516–547.
[10] F. Andreu, V. Caselles, and J. M. Mazón, *Parabolic Quasilinear Equations Minimizing Linear Growth Functionals*, Progr. Math. 223, Birkhäuser Verlag, Basel, 2004.
[11] G. Anzellotti, *Pairings between measures and bounded functions and compensated compactness*, Ann. Mat. Pura Appl. (4), 135 (1983), pp. 293–318.
[12] G. Bellettini, V. Caselles, and M. Novaga, *The total variation flow in R^N*, J. Differential Equations, 184 (2002), pp. 475–525.
[13] G. Bellettini, M. Novaga, and M. Paolini, *Characterization of facet-breaking for nonsmooth mean curvature flow in the convex case*, Interfaces Free Bound., 3 (2001), pp. 415–446.
[14] G. Bellettini, M. Novaga, and M. Paolini, *On a crystalline variational problem* I. *First variation and global L^∞ regularity*, Arch. Ration. Mech. Anal., 157 (2001), pp. 165–191.
[15] G. Bellettini, M. Novaga, and M. Paolini, *On a crystalline variational problem.* II. *BV regularity and structure of minimizers on facets*, Arch. Ration. Mech. Anal., 157 (2001), pp. 193–217.
[16] H. Brezis, *Operateurs Maximaux Monotones*, North-Holland, Amsterdam, 1973.
[17] A. Chambolle and P. L. Lions, *Image recovery via total variation minimization and related problems*, Numer. Math., 76 (1997), pp. 167–188.
[18] J. T. Chen, *On the existence of capillary free surfaces in the absence of gravity*, Pacific J. Math., 88 (1980), pp. 323–361.
[19] M. G. Crandall and T. M. Liggett, *Generation of semigroups of nonlinear transformations on general Banach spaces*, Amer. J. Math., 93 (1971), pp. 265–298.
[20] R. De Vore and B. J. Lucier, *Fast wavelet techniques for near optimal image compression*, in IEEE Military Communications Conference Record, San Diego, Oct. 11–14, IEEE, Piscataway, NJ, 1992, pp. 1129–1135.
[21] D. Donoho, *Denoising via soft-thresholding*, IEEE Trans. Inform. Theory, 41 (1995), pp. 613–627.
[22] D. Donoho, *Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition*, Appl. Comput. Harmon. Anal., 2 (1995), pp. 101–126.
[23] D. Donoho, I. Johnstone, G. Kerkyacharian, and D. Picard, *Wavelet shrinkage: Asymptopia?*, J. Roy. Statist. Soc. Ser. B, 57 (1995), pp. 301–369.
[24] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, Stud. Adv. Math., CRC Press, Boca Raton, FL, 1992.
[25] R. Finn, *A subsidiary variational problem and existence criteria for capillary surfaces*, J. Reine Angew. Math., 353 (1984), pp. 196–214.
[26] R. Finn, *Equilibrium Capillary Surfaces*, Springer-Verlag, New York, 1986.
[27] E. Giusti, *On the equation of surfaces of prescribed mean curvature. Existence and uniqueness without boundary conditions*, Invent. Math., 46 (1978), pp. 111–137.

[28] E. Giusti, *Boundary value problems for non-parametric surfaces of prescribed mean curvature*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 3 (1976), pp. 501–548.
[29] Y. Gousseau and J. M. Morel, *Are natural images of bounded variation?*, SIAM J. Math. Anal., 33 (2001), pp. 634–648.
[30] U. Massari, *Frontiere orientate di curvatura media assegnata in L^p*, Rend. Sem. Mat. Univ. Padova, 53 (1975), pp. 37–52.
[31] Y. Meyer, *Oscillating patterns in image processing and nonlinear evolution equations*, The Fifteenth Dean Jacqueline B. Lewis Memorial Lectures, Univ. Lecture Ser. 22, AMS, Providence, RI, 2001.
[32] M. Miranda, *Un principio di massimo forte per le frontiere minimali e una sua applicazione alla risoluzione del problema al contorno per l'equazione delle superfici di area minima*, Rend. Sem. Mat. Univ. Padova, 45 (1971), pp. 355–366.
[33] L. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D, 60 (1992), pp. 259–268.
[34] L. A. Santaló, *Integral Geometry and Geometric Probability*, Encyclopedia Math. Appl. 1, Addison-Wesley, Reading, MA, London, Amsterdam, 1976.
[35] G. Steidl, J. Weickert, T. Brox, P. Mrázek, and M. Welk, *On the equivalence of soft wavelet shrinkage, total variation diffusion, total variation regularization, and SIDEs*, SIAM J. Numer. Anal., 42 (2004), pp. 686–713.
[36] W. P. Ziemer, *Weakly Differentiable Functions*, Springer-Verlag, Ann Harbor, MI, 1989.

SIAM J. MATH. ANAL.
Vol. 36, No. 4, pp. 1130–1141

ON THE INVISCID LIMIT FOR TWO-DIMENSIONAL INCOMPRESSIBLE FLOW WITH NAVIER FRICTION CONDITION*

M. C. LOPES FILHO†, H. J. NUSSENZVEIG LOPES†, AND G. PLANAS‡

Abstract. In [*Nonlinearity*, 11 (1998), pp. 1625–1636], Clopeau, Mikelić, and Robert studied the inviscid limit of the two-dimensional incompressible Navier–Stokes equations in a bounded domain subject to Navier friction–type boundary conditions. They proved that the inviscid limit satisfies the incompressible Euler equations, and their result ultimately includes flows generated by bounded initial vorticities. Our purpose in this article is to adapt and, to some extent, simplify their argument in order to include pth power integrable initial vorticities, with $p > 2$.

Key words. Navier–Stokes, boundary layers, vanishing viscosity

AMS subject classifications. 35Q30, 76D05, 76D10

DOI. 10.1137/S0036141003432341

1. Introduction. In a recent paper [1], Clopeau, Mikelić, and Robert studied the inviscid limit of solutions of the two-dimensional (2D) incompressible Navier–Stokes equations in a bounded domain with Navier friction–type boundary conditions. They proved that the inviscid limit is a weak solution of the Euler equations, and their results include flows generated by bounded initial vorticities. The purpose of the present work is to extend their argument in order to include flows with initial vorticities in L^p, $p > 2$. Technically this work involves many of the same tools that were used in [1] and relies in an essential manner on the smooth data result of that work.

The main motivation for studying the vanishing viscosity limit for incompressible 2D flow is the problem of boundary layers. This motivation, together with the issue of the physical meaning of the Navier friction condition, was well explored in the introduction to [1]. We will not repeat that discussion here, referring the reader to that article and the references therein for this part of our introduction. We would like to mention that in a pair of recent papers, Jäger and Mikelić rigorously justified the Navier friction condition as a homogenization of the no-slip condition on a rough boundary [5, 6]. Furthermore, 2D boundary layers have been a very active field of inquiry recently, so in addition to [1, 5, 6] we also refer the reader to [3, 14] for other recent developments. Beyond these issues, there is additional background which is specifically related to irregular flows which we must address here.

Existence of weak solutions to the incompressible 2D Euler equations has been established for rather singular initial data, more precisely, initial velocities in L^2_{loc} such that the corresponding vorticity lies in $\mathcal{BM}^+_c + L^1_c$, i.e., nonnegative bounded measures with compact support plus an arbitrary compactly supported integrable function. This result is due primarily to Delort [2]; we refer the reader also to [15]. In both of these papers the weak solutions are obtained by compactness arguments in which the initial data is mollified and the equations are subsequently exactly solved with this

*Received by the editors July 21, 2003; accepted for publication (in revised form) March 26, 2004; published electronically January 27, 2005.

http://www.siam.org/journals/sima/36-4/43234.html

†Departamento de Matemática, IMECC-UNICAMP, Caixa Postal 6065, Campinas, SP 13083-970, Brazil (mlopes@ime.unicamp.br, hlopes@ime.unicamp.br). The research of the first author was supported in part by CNPq grant 300.962/91-6, and that of the second author in part by CNPq grant 300.158/93-9.

‡Departamento de Matemática, ICMC-USP, Caixa Postal 668, São Carlos, SP 13560-970, Brazil (gplanas@icmc.usp.br). This author was supported by FAPESP, grant 01/14455-2.

smooth data. Uniqueness has only been established if the initial vorticity is bounded or nearly so [16, 17, 18], so that the issue of selection principles for singular flows is wide open. It makes sense in this case to investigate whether other approximation schemes also yield weak solutions. For example, this has been established for certain numerical schemes; see [10, 13]. It is natural, from a physical point of view, to investigate the vanishing viscosity limit as well. It is possible to adapt Delort's arguments to study the inviscid limit in the absence of boundaries, and this has been done for full plane flow; see [11]. The problem of studying the existence of viscosity solutions in domains with boundary runs into the classical problem of boundary layers if one supplements the viscous approximations with the no-slip boundary condition. The work of Clopeau, Mikelić, and Robert shows that the boundary layer arising from the inviscid limit with Navier friction condition can be treated, while retaining some physical meaning. In fact, the Navier friction condition still allows for vorticity production at the boundary, but in a controlled fashion. It is therefore natural to investigate the existence of viscosity solutions by considering viscous approximations satisfying the Navier friction condition, searching for critical regularity on the initial data that guarantees the existence of such solutions. This is the main point behind the present work.

The remainder of this article is divided into five sections: the next section contains the basic notation and setup of the problem; the third section investigates approximation of initial data that satisfy the Navier friction condition; the fourth section contains the a priori estimate on the L^p-norm of vorticity which is the heart of this work; the fifth section contains a well-posedness result for the viscous approximations with L^p initial vorticity; the last section contains the passage to the inviscid limit and conclusions.

2. Preliminaries. Let $\Omega \subseteq \mathbb{R}^2$ denote a bounded simply connected domain with smooth boundary. Our point of departure is the incompressible Navier–Stokes equations in Ω. We are interested in the initial-boundary value problem where the velocity satisfies the Navier friction condition with friction coefficient $\alpha = \alpha(x) \in C^2(\partial\Omega)$, $\alpha \geq 0$. More precisely, the initial-boundary value problem is given by

$$\begin{cases} u_t + u \cdot \nabla u = -\nabla p + \nu \Delta u & \text{in } \Omega \times (0,T), \\ \operatorname{div} u = 0 & \text{in } \Omega \times [0,T), \\ u \cdot \mathbf{n} = 0 & \text{on } \partial\Omega \times [0,T), \\ 2(Du)_S \mathbf{n} \cdot \tau + \alpha u \cdot \tau = 0 & \text{on } \partial\Omega \times (0,T), \\ u(x,0) = u_0(x) & \text{in } \Omega, \end{cases} \tag{1}$$

where $\nu > 0$ is the viscosity, $\mathbf{n}$ and τ are the unit outwards normal and counterclockwise tangent vectors to $\partial\Omega$, respectively; u is the fluid velocity; p is the scalar pressure; and $(Du)_S$ is the symmetric part of the Jacobian matrix of u, i.e. $(Du)_S = \frac{1}{2}(Du + (Du)^t)$.

The well-posedness of this initial-boundary value problem was established by Clopeau, Mikelić, and Robert in [1]. More precisely, given a divergence-free initial velocity field $u_0 \in H^2(\Omega)$, tangent to the boundary, and satisfying the Navier friction condition $2(Du)_S \mathbf{n} \cdot \tau + \alpha u \cdot \tau = 0$ on $\partial\Omega$ in the trace sense, they showed that there exists a unique weak solution $u^\nu \in L^2((0,T); H^1(\Omega)) \cap L^\infty((0,T); L^2(\Omega))$ satisfying

$$\begin{aligned} &\frac{d}{dt}\int_\Omega \varphi u^\nu + \int_\Omega \varphi \cdot (u^\nu \cdot \nabla) u^\nu dx \\ &+ 2\nu \int_\Omega (D\varphi)_S : (Du^\nu)_S dx + \nu \int_{\partial\Omega} \alpha (\varphi \cdot \tau)(u^\nu \cdot \tau) dS = 0 \end{aligned} \tag{2}$$

for every divergence-free test vector field $\varphi \in H^1(\Omega)$, tangent to $\partial\Omega$. Here the matrix product $A : B$ means $\sum_{i,j} A_{ij}B_{ij}$ and is called the trace product.

We note that the initial condition is not included in this weak formulation. In fact, Clopeau, Mikelić, and Robert also showed that $u^\nu_t \in L^2((0,T); H^1(\Omega))$, from which it follows by integration that $u^\nu \in C([0,T]; H^1(\Omega))$. Therefore the initial condition $u^\nu(\cdot, 0) = u_0$ can be meaningfully imposed. Furthermore, if one assumes that the initial vorticity curl u_0 is bounded, then $u^\nu \in C([0,T); H^2(\Omega))$.

The Navier friction condition can be formulated in terms of vorticity. In order to do so, a calculus identity was established in [1], which we reproduce in the lemma below.

LEMMA 1. *Let $v \in H^2(\Omega)$ be a vector field which is tangent to $\partial\Omega$. Then*

$$(Dv)_S \mathbf{n} \cdot \tau - \frac{\omega}{2} + \kappa(v \cdot \tau) = 0 \quad \text{on } \partial\Omega,$$

where $\omega = \operatorname{curl} v$ and κ is the curvature of $\partial\Omega$.

One of the main difficulties in addressing the classical vanishing viscosity limit in domains with boundary resides in writing useful boundary conditions for the vorticity formulation of the 2D Navier–Stokes equations. It is through the use of the vorticity formulation that one finds higher order estimates for velocity that are independent of viscosity. The inviscid limit for 2D Navier–Stokes with friction condition is more tractable than the classical problem precisely because the friction boundary condition translates into a useful boundary condition for vorticity. We introduce $\omega_0 = \operatorname{curl} u_0$, the initial vorticity, and $\omega^\nu = \operatorname{curl} u^\nu$, the time-dependent vorticity associated to the weak solution u^ν of (1) with initial data u_0. For each fixed time, the velocity u^ν can be recovered from vorticity by means of the Biot–Savart law. We make this explicit by writing

$$u^\nu = K_\Omega(\omega^\nu),$$

where K_Ω is an integral operator of order -1, with kernel given by $\nabla^\perp G_\Omega$, where G_Ω is the Green's function for the Dirichlet Laplacian in Ω. Using Lemma 1 above, it is a standard calculation to show that ω^ν, u^ν satisfies, in a weak sense, the following parabolic initial-boundary value problem, which is the vorticity formulation of (1):

$$\begin{cases} \omega^\nu_t + u^\nu \cdot \nabla\omega^\nu = \nu\Delta\omega^\nu & \text{in } \Omega \times (0,T), \\ u^\nu = K_\Omega[\omega^\nu] & \text{in } \Omega \times [0,T), \\ \omega^\nu = (2\kappa - \alpha)u^\nu \cdot \tau & \text{on } \partial\Omega \times [0,T) \\ \omega^\nu(\cdot, 0) = \omega_0 & \text{on } \Omega \times \{t = 0\}. \end{cases} \tag{3}$$

3. Approximating nonsmooth initial data. The problem we wish to address in this article is the inviscid limit for (1) with initial velocity $u_0 = K_\Omega[\omega_0]$, and $\omega_0 \in L^p(\Omega)$ for some $p > 2$. We must first discuss this initial-boundary value problem for fixed viscosity, as this initial condition does not satisfy the conditions for the well-posedness mentioned in the previous section. It can be easily seen that this initial velocity u_0 is divergence free, tangent to the boundary, and it belongs to $W^{1,p}(\Omega)$ (by elliptic regularity; see [9]). This means that there is not enough regularity to impose the Navier friction condition on the initial data, so that this initial-boundary value problem is subject to an initial layer.

DEFINITION 1. *We will call a function $\omega \in H^1(\Omega) \cap L^\infty(\Omega)$ compatible if the associated velocity $u = K_\Omega[\omega] \in H^2(\Omega)$ satisfies the Navier condition $\omega = (2\kappa - \alpha)u \cdot \tau$ on the boundary in the trace sense.*

The first issue we need to address is how to approximate an arbitrary function in $L^p(\Omega)$ by compatible functions. This issue was addressed by Clopeau, Mikelić, and Robert for $\omega \in L^\infty(\Omega)$, using a fixed point argument. We will state and prove an extension of their result that applies to functions $\omega \in L^p(\Omega)$, $p > 1$. The proof is a reasonably straightforward adaptation of their argument, which we include for the sake of completeness.

LEMMA 2. *Let $\omega \in L^p(\Omega)$ for some $p > 1$. Then there exists a sequence $\{\omega_n\}$ of compatible functions which converges to ω strongly in L^p.*

Proof. Recall the notation introduced in the proof of Lemma 4.2 of [1]. For $x \in \overline{\Omega}$, let $d = d(x)$ be the distance of x to $\partial\Omega$ and let $U_n \equiv \{x \in \overline{\Omega} : d(x) < 1/n\}$. Let $r = r(x)$ denote the orthogonal projection of U_n onto $\partial\Omega$, defined for n sufficiently large. Let ζ_n be a smooth cutoff for a neighborhood of $\Omega \setminus U_n$, so that $\zeta_n \equiv 0$ in U_{n+1} and $\zeta_n \equiv 1$ outside U_n. Let η_n be a standard Friedrichs mollifier. As in Lemma 4.2 we extend ω to vanish outside of Ω. First, assume that $p < 2$ and let $\widehat{p} = p/(2-p)$. For any $G \in L^{\widehat{p}}(\partial\Omega)$ set

$$\beta \equiv \zeta_n(x)\eta_n * \omega(x) + (1 - \zeta_n(x))e^{-nd(x)}G(r(x)). \tag{4}$$

As $G(r(\cdot))$ appears multiplied by a function which vanishes outside U_n we may assume that $G(r(\cdot))$ vanishes outside U_n. Thus (the extended) $G(r(\cdot))$ is defined on all of Ω. Observe also that, by construction, $\beta|_{\partial\Omega} = G$.

Let

$$v = K_\Omega[\beta]$$

and introduce

$$\Psi(G) = (2\kappa - \alpha)v \cdot \tau.$$

We note that Ψ maps $L^{\widehat{p}}(\partial\Omega)$ into itself. To see this, we begin by observing that $\beta \in L^p(\Omega)$. This follows since $G \in L^{\widehat{p}}(\partial\Omega)$, which implies, by a simple change of variables, that $G(r(\cdot)) \in L^{\widehat{p}}(\Omega)$. Since $\widehat{p} > p$, because $p > 1$, we obtain $\beta \in L^p(\Omega)$. Therefore $v \in W^{1,p}(\Omega)$, so that $v \cdot \tau \in W^{1-1/p,p}(\partial\Omega)$. We conclude using the Sobolev imbedding $W^{1-1/p,p}(\partial\Omega) \subset L^{\widehat{p}}(\partial\Omega)$.

Next we show that Ψ is a contraction mapping if n is sufficiently large. Let $G_1, G_2 \in L^{\widehat{p}}(\partial\Omega)$. Then

$$\|\Psi(G_1) - \Psi(G_2)\|_{L^{\widehat{p}}(\partial\Omega)} \le \|2\kappa - \alpha\|_{L^\infty}\|v_1 - v_2\|_{L^{\widehat{p}}(\partial\Omega)} \le C_p\|\beta_1 - \beta_2\|_{L^p(\Omega)}$$

$$\le C_p\|G_1(r(\cdot)) - G_2(r(\cdot))\|_{L^p(U_n)} \le C_p\frac{1}{n^{1/p}}\|G_1 - G_2\|_{L^{\widehat{p}}(\partial\Omega)}.$$

Therefore, for n sufficiently large, Ψ has a unique fixed point, which we denote by $G^n \in L^{\widehat{p}}(\partial\Omega)$. We denote the corresponding β by ω_n, so that

$$\omega_n \equiv \zeta_n\eta_n * \omega + (1 - \zeta_n)e^{-nd}G^n \circ r.$$

We need to verify that ω_n is compatible. We begin by observing that the fact that G^n is a fixed point for Ψ implies that ω_n satisfies the Navier friction condition. Next we show the required regularity for ω_n. A standard bootstrap argument on identity (4), involving Sobolev imbeddings and elliptic regularity, gains $1 - 1/\widehat{p}$ derivatives on ω_n at each step. Indeed, $G^n \in L^{\widehat{p}}(\partial\Omega)$ implies that $G^n \circ r \in L^{\widehat{p}}(\Omega)$

since $\partial\Omega$ is smooth. Hence $\omega_n \in L^{\widehat{p}}(\Omega)$ because $\zeta_n \eta_n * \omega$ is smooth and compactly supported in the interior of Ω. From this it follows that $v_n = K_\Omega[\omega_n] \in W^{1,\widehat{p}}(\Omega)$, so that $v_n|_{\partial\Omega} \in W^{1-1/\widehat{p},\widehat{p}}(\partial\Omega)$. Thus $\Psi(G_n) = (2\kappa - \alpha) v_n \cdot \tau \in W^{1-1/\widehat{p},\widehat{p}}(\partial\Omega)$. But G_n is a fixed point for Ψ, so we conclude that $G_n \in W^{1-1/\widehat{p},\widehat{p}}(\partial\Omega)$. From this last step it follows that $\omega_n \in W^{1-1/\widehat{p},\widehat{p}}(\Omega)$. We may repeat this argument and gain $1 - 1/\widehat{p}$ derivatives at each step. After a finite number of steps we reach $\omega_n \in H^1 \cap L^\infty$.

Finally, we argue that ω_n converges strongly to ω in L^p. Since the first term on the right-hand side of (4) clearly converges strongly to ω in L^p, all we need to show is that the remaining term converges to zero in L^p. First note that

$$\|(1-\zeta_n)e^{-nd(x)}G^n(r(x))\|_{L^p(\Omega)} \le \|G^n(r(x))\|_{L^p(U_n)} \le \frac{C}{n^{1/p}}\|G^n\|_{L^p(\partial\Omega)}$$

$$\le o(1)\|G^n\|_{L^{\widehat{p}}(\partial\Omega)}.$$

Now we estimate $\|G^n\|_{L^{\widehat{p}}(\partial\Omega)}$:

$$\|G^n\|_{L^{\widehat{p}}(\partial\Omega)} \le \|2\kappa - \alpha\|_{L^\infty}\|K_\Omega[\omega_n]\|_{L^{\widehat{p}}(\partial\Omega)}$$

$$\le C(p,\alpha,\kappa)\|\omega_n\|_{L^p(\Omega)} \le C(\|\omega\|_{L^p(\Omega)} + \|G^n(r(x))\|_{L^p(U_n)})$$

$$\le C_p\|\omega\|_{L^p(\Omega)} + \frac{1}{2}\|G^n\|_{L^{\widehat{p}}(\partial\Omega)}$$

for n sufficiently large, which implies the required bound.

For $p = 2$ one repeats the argument above with an arbitrary $\widehat{p}$, and for $p > 2$ one just takes $\widehat{p} = \infty$. □

Remark 1. The result presented is actually more general than what we require. It applies to initial vorticities in L^p, $p > 1$, when we are only going to use it for $p > 2$. It is worth remarking that it is only for the cases $1 < p \le 2$ that we needed to use a fixed point argument in $L^{\widehat{p}}$. We could have written an argument that works for the case $p > 2$ using the fixed point argument in L^∞, as was done in [1], and the proof would really be a very minor adaptation of the proof in [1], not deserving repetition even for the sake of completeness. One of the points of the present work is to clarify the criticality of this problem. This is the main reason to present the approximation result in this generality. The way it is formulated implies that this approximation issue is not part of the $p > 2$ limitation.

Remark 2. There is no asymptotic description of the structure of the boundary layer for the present problem that would be the adaptation of Prandtl's description of the classical boundary layer. We point out that the small viscosity regime does not appear to be physically meaningful under Navier friction conditions, so that there has been no compelling reason to obtain such a description. However, an account of the structure of the boundary layer under Navier friction conditions would certainly be of mathematical interest. In the absence of such an account, the proof above gives at least a clue as to the nature of this boundary layer, embodied in the structure of the correction term. One key issue in the classical boundary layer is that such a correction term would have, at best, a uniform L^1 estimate, leading to a boundary vortex sheet perturbation in the limit. This is apparent in the explicitly computable flow generated by an impulsively started plate, known as the Rayleigh problem; see, for example, [12]. This vortex sheet at the boundary is present in the inviscid limit even for smooth initial vorticities. Now, vortex sheet–type regularity is critical for passing to the weak limit in approximations of the incompressible 2D Euler equations; see [2]. In some sense, it

is this fact that is the heart of the difficulty in the classical boundary layer problem. The correction term in the proof above suggests that the boundary layer associated to the Navier friction condition would correspond to uniformly bounded vorticity near the boundary for $p > 2$, and $L^{\widehat{p}}$ vorticity near the boundary for $1 < p \leq 2$, so that one would expect criticality only at $p = 1$.

Remark 3. The argument presented breaks down when $p = 1$, mainly because elliptic regularity breaks down, so that one cannot guarantee that Ψ maps L^1 to itself.

4. A priori estimate on vorticity. The purpose of this section is to derive an a priori estimate for vorticity on solutions of (1). We begin with a compatible initial vorticity ω_0, as defined in the previous section. We use $u_0 = K_\Omega[\omega_0]$ as initial data. The well-posedness of the initial-boundary value problem (1) for such initial data was established in [1], as previously mentioned. Let $u = u(x,t)$ be the unique weak solution of (1) with data u_0. The vector field u belongs to $C([0,\infty); H^2(\Omega))$ and satisfies the weak formulation (2) of the Navier–Stokes equation with Navier friction condition. The vorticity $\omega = \operatorname{curl} u$ satisfies the parabolic equation (3) in a weak sense.

LEMMA 3. *Fix $p > 2$. There exists a constant $C > 0$, depending only on p, Ω, and the friction coefficient α such that the vorticity satisfies*

$$\|\omega(\cdot,t)\|_{L^p} \leq C(\|\omega_0\|_{L^p} + \|u_0\|_{L^2}).$$

Proof. The proof involves applying a maximum principle to two auxiliary problems. First observe that $u \cdot \tau \in L^\infty(\partial\Omega \times (0,T))$ since $u \in C([0,T); H^2(\Omega))$. Set

$$\Lambda = \|(2\kappa - \alpha) u \cdot \tau\|_{L^\infty(\partial\Omega \times (0,T))}.$$

Consider the initial-boundary value problem for the Fokker–Planck equation

$$\begin{cases} \widetilde{\omega}_t - \nu\Delta\widetilde{\omega} + u \cdot \nabla\widetilde{\omega} = 0 & \text{in } \Omega \times (0,T), \\ \widetilde{\omega}(\cdot,0) = |\omega_0| & \text{in } \Omega, \\ \widetilde{\omega} = \Lambda & \text{on } \partial\Omega \times (0,T). \end{cases} \tag{5}$$

This problem has a unique weak solution $\widetilde{\omega} \in L^2((0,T); H^1(\Omega))$ by Theorems 6.1 and 6.2 in [7]. Then $\omega_1 = \omega - \widetilde{\omega}$ is a weak solution for the following initial-boundary value problem:

$$\begin{cases} (\omega_1)_t - \nu\Delta\omega_1 + u \cdot \nabla\omega_1 = 0 & \text{in } \Omega \times (0,T), \\ \omega_1(\cdot,0) = \omega_0 - |\omega_0| & \text{in } \Omega, \\ \omega_1 = (2\kappa - \alpha) u \cdot \tau - \Lambda & \text{on } \partial\Omega \times (0,T). \end{cases} \tag{6}$$

The coefficients of the Fokker–Planck operator $\partial_t - \nu\Delta + u \cdot \nabla$ are such that the maximum principle for weak solutions, given in Corollary 6.26 of [7], is valid. Therefore, as $\omega_1 \leq 0$ on the parabolic boundary $\partial\Omega \times (0,T) \cup \Omega \times \{t = 0\}$, it follows that $\omega_1 \leq 0$ a.e. in $\Omega \times [0,T)$. Analogously, we prove that $\omega_2 = -\omega - \widetilde{\omega}$ is nonpositive. We thus obtain

$$|\omega| \leq \widetilde{\omega} \text{ a.e. in } \Omega \times [0,T). \tag{7}$$

Moreover, as ω_0 is compatible, it is bounded. Hence Corollary 6.26 of [7] may also be applied to (5), yielding that $\widetilde{\omega} \in L^\infty((0,T) \times \Omega)$.

Next we obtain an estimate for $\widetilde{\omega}$. Let $\widehat{\omega} = \widetilde{\omega} - \Lambda$. This is a solution of the following problem:

$$\begin{cases} \widehat{\omega}_t - \nu\Delta\widehat{\omega} + u\cdot\nabla\widehat{\omega} = 0 & \text{in } \Omega\times(0,T), \\ \widehat{\omega}(\cdot,0) = |\omega_0| - \Lambda & \text{in } \Omega, \\ \widehat{\omega} = 0 & \text{on } \partial\Omega\times(0,T). \end{cases} \tag{8}$$

We formally multiply (8) by $\widehat{\omega}|\widehat{\omega}|^{p-2}$, where $p>2$, we integrate by parts and use the incompressibility of the flow u to obtain

$$\frac{1}{p}\frac{d}{dt}\int_\Omega |\widehat{\omega}|^p + (p-1)\nu\int_\Omega ||\nabla\widehat{\omega}||\widehat{\omega}|^{(p-2)/2}|^2 dx = 0. \tag{9}$$

Then

$$\|\widehat{\omega}(\cdot,t)\|_{L^p(\Omega)} \le \|\widehat{\omega}(\cdot,0)\|_{L^p(\Omega)} \le \|\omega_0\|_{L^p(\Omega)} + \Lambda|\Omega|^{1/p}.$$

Therefore,

$$\|\widetilde{\omega}\|_{L^p(\Omega)} \le \|\widehat{\omega}\|_{L^p(\Omega)} + \Lambda|\Omega|^{1/p} \le \|\omega_0\|_{L^p(\Omega)} + 2\Lambda|\Omega|^{1/p}. \tag{10}$$

This formal calculation can be made rigorous by using the weak formulation of (8) given in [7]. We begin by observing that $\widehat{\omega}_t \in L^2((0,T);H^{-1}(\Omega))$ and $\widehat{\omega}\in L^2((0,T);H^1_0(\Omega))\cap L^\infty((0,T)\times\Omega)$. This implies that $\widehat{\omega}|\widehat{\omega}|^{p-2}\in L^2((0,T);H^1_0(\Omega))$. Therefore we can multiply (8) by $\widehat{\omega}|\widehat{\omega}|^{p-2}$ if we understand the product with $\widehat{\omega}_t$ and with $\Delta\widehat{\omega}$ as duality pairings. Finally, in order to justify (9) one still needs to approximate $\widehat{\omega}$ by suitable smooth functions and pass to the limit in each term of the weak formulation so as to obtain

$$\frac{1}{p}\frac{d}{dt}\int_\Omega |\widehat{\omega}|^p = \langle\widehat{\omega}_t, \widehat{\omega}|\widehat{\omega}|^{p-2}\rangle$$

$$= \nu\langle\Delta\widehat{\omega}, \widehat{\omega}|\widehat{\omega}|^{p-2}\rangle = -(p-1)\nu\int_\Omega ||\nabla\widehat{\omega}||\widehat{\omega}|^{(p-2)/2}|^2 dx.$$

This can be easily accomplished using mollification in time together with the Dirichlet heat semigroup for Ω, thus generating a family of smooth functions $\widehat{\omega}_\varepsilon$ such that $\partial_t\widehat{\omega}_\varepsilon\to\widehat{\omega}_t$ strongly in $L^2((0,T);H^{-1}(\Omega))$, while $\widehat{\omega}_\varepsilon|\widehat{\omega}_\varepsilon|^{p-2}\rightharpoonup\widehat{\omega}|\widehat{\omega}|^{p-2}$ weakly in $L^2((0,T);H^1_0(\Omega))$ and $\widehat{\omega}_\varepsilon$ is uniformly bounded in $\Omega\times(0,T)$.

Given (10) we now turn to the estimate of Λ. Using Sobolev imbedding and interpolating between $W^{1,p}$ and L^2, we find

$$\|u(\cdot,t)\cdot\tau\|_{L^\infty(\partial\Omega)} \le C\|u(\cdot,t)\|_{C(\bar\Omega)} \le C\|u(\cdot,t)\|^\theta_{L^2(\Omega)}\|u(\cdot,t)\|^{1-\theta}_{W^{1,p}(\Omega)}$$

$$\le C\|u(\cdot,t)\|^\theta_{L^2(\Omega)}\|\omega(\cdot,t)\|^{1-\theta}_{L^p(\Omega)},$$

where $\theta = (p-2)/(2p-2)$.

Let ε be an arbitrary positive number. We now use Young's inequality together with the fact that κ and α are bounded to conclude that

$$\Lambda \le C_\varepsilon\|u\|_{L^\infty((0,T);L^2(\Omega)} + \varepsilon\|\omega\|_{L^\infty((0,T);L^p(\Omega))} \tag{11}$$

for some $C_\varepsilon>0$. Taking ε small enough, from (7)–(11) we obtain

$$\|\omega\|_{L^\infty(0,T;L^p(\Omega))} \le C(\|\omega_0\|_{L^p(\Omega)} + \|u\|_{L^\infty(0,T;L^2(\Omega))}) \tag{12}$$

for any $p > 2$, where $C = C(p, \Omega, \|\kappa\|_{L^\infty(\partial\Omega)}, \|\alpha\|_{L^\infty(\partial\Omega)})$. Finally, a standard energy estimate, such as the one carried out in [1] (see estimate (2.16)), yields $\|u\|_{L^\infty(0,T;L^2(\Omega))} \leq \|u_0\|_{L^2(\Omega)}$, thereby concluding the proof. ∎

Remark 4. This lemma is the heart of this article. Note that the restriction $p > 2$ comes into the proof above because of the need to produce a uniform bound on the velocity at the boundary. It would be interesting to know if this is a physically meaningful restriction. This would mean that the problem of controlling the generation of vorticity by the interaction of incompressible flow with a "Navier condition" boundary is critical at $p = 2$. However, this criticality at $p = 2$ seems unlikely. The limitation on the integrability of vorticity in the proof above appears to reflect a limitation on the maximum principle technique employed rather than an essential feature of this problem. In contrast, the exponent $p = 1$ found to be critical in the proof of Lemma 2 seems much more essential and is already known to be critical in terms of passage to weak limits on the nonlinearity of the incompressible 2D Euler and Navier–Stokes equations.

Remark 5. The natural way to extend this vorticity estimate to $p \leq 2$ would be to derive an L^p energy estimate on the vorticity equation. Multiplying the vorticity equation (3) by $p\omega|\omega|^{p-2}$, integrating in space, and performing the usual integration by parts yields

$$\frac{d}{dt}\int_\Omega |\omega|^p dx = -\nu p(p-1)\int_\Omega |\omega|^{p-2}|\nabla\omega|^2 dx + \nu p\int_{\partial\Omega} |\omega|^{p-2}\omega\nabla\omega\cdot\mathbf{n}dS.$$

We note that the boundary term is the flux of $|\omega|^p$ through the boundary, over which we have no control. One special case for which this simple estimate does provide an improvement over Lemma 3 is the case of $\alpha = 2\kappa$, because then the troublesome boundary term vanishes. This corresponds to the so-called free boundary condition $\omega = 0$ on $\partial\Omega$. It is a well-known fact that one can handle the inviscid asymptotics in this case, as one is imposing that the boundary does not generate vorticity and thus there are no boundary layers. For details, see [8] and the special case of time-dependent domain in [4].

5. Well-posedness for the viscous problem. In this section we observe that the initial-boundary value problem for the Navier–Stokes equations with friction-type boundary condition is well-posed even if the initial vorticity is not compatible. This was already done in [1] for bounded initial vorticity.

We begin by showing that if u^ν is a weak solution of (1), then u^ν satisfies an integrated version of relation (2).

LEMMA 4. *Let $u^\nu \in L^2((0,T);H^1(\Omega)) \cap L^\infty((0,T);L^2(\Omega))$ be a weak solution of (1). Then for any test vector field $\varphi \in C_c^\infty([0,T)\times\overline{\Omega})$, divergence free and tangent to $\partial\Omega$, we have*

$$\begin{aligned}\int_0^T\int_\Omega u^\nu\varphi_t + u^\nu(u^\nu\cdot\nabla)\varphi dxdt + \int_\Omega u_0\varphi(\cdot,0)dx \\ = 2\nu\int_0^T\int_\Omega (D\varphi)_S : (Du)_S dxdt + \nu\int_0^T\int_{\partial\Omega}\alpha(\varphi\cdot\tau)(u^\nu\cdot\tau)dSdt.\end{aligned} \tag{13}$$

Proof. Let φ be a test vector field. For each $s \in [0,T)$, define

$$g(t,s) \equiv \int_\Omega u^\nu(x,t)\varphi(x,s)dx.$$

Then by (2) we have

$$\frac{\partial g}{\partial t} = -\int_\Omega \varphi \cdot (u^\nu \cdot \nabla) u^\nu dx$$

$$- 2\nu \int_\Omega (D\varphi)_S : (Du^\nu)_S dx - \nu \int_{\partial\Omega} \alpha(\varphi \cdot \tau)(u^\nu \cdot \tau) dS$$

$$= \int_\Omega u^\nu (u^\nu \cdot \nabla)\varphi dx - 2\nu \int_\Omega (D\varphi)_S : (Du^\nu)_S dx - \nu \int_{\partial\Omega} \alpha(\varphi \cdot \tau)(u^\nu \cdot \tau) dS,$$

using integration by parts and the fact that u^ν is divergence free. On the other hand, we also have

$$\frac{\partial g}{\partial s} = \int_\Omega u^\nu(x,t)\varphi_s(x,s) dx.$$

Therefore, it follows that

$$\frac{d}{dt}(g(t,t)) = \int_\Omega u^\nu(x,t)\varphi_t(x,t) dx$$

$$+ \int_\Omega u^\nu (u^\nu \cdot \nabla)\varphi dx - 2\nu \int_\Omega (D\varphi)_S : (Du^\nu)_S dx - \nu \int_{\partial\Omega} \alpha(\varphi \cdot \tau)(u^\nu \cdot \tau) dS.$$

Integrating this last identity in time and identifying the initial data yields the desired result. □

We now state and prove the main result in this section.

PROPOSITION 1. *Let $\omega_0 \in L^p(\Omega)$ for some $p > 2$ and $u_0 = K_\Omega[\omega_0]$. Fix $\nu > 0$. Then there exists a unique vector field $u^\nu = u^\nu(x,t) \in C([0,T); L^2(\Omega)) \cap L^2((0,T); H^1(\Omega))$ satisfying the weak formulation* (2) *of the* 2*D incompressible Navier–Stokes system* (1) *with initial data u_0. Moreover, the associated vorticity $\omega^\nu = \operatorname{curl} u^\nu$ satisfies the estimate*

$$\|\omega^\nu(\cdot,t)\|_{L^p(\Omega)} \le C$$

a.e. in time, with constant $C > 0$ independent of viscosity.

Proof. Let $\omega_{0,n}$ be a sequence of compatible functions approximating ω_0 strongly in L^p, as constructed in Lemma 2, and let $u_{0,n} = K_\Omega[\omega_{0,n}]$. Let u_n^ν be the weak solution of system (1) given by the well-posedness result of [1], and let $\omega_n^\nu = \operatorname{curl} u_n^\nu$ be the corresponding vorticity. We begin by observing that Lemma 3 gives the uniform estimate

$$\|\omega_n^\nu\|_{L^\infty((0,T);L^p(\Omega))} \le C(\|\omega_0\|_{L^p(\Omega)} + \|u_0\|_{L^2(\Omega)}) \tag{14}$$

for some $C > 0$. By the Poincaré and Calderón–Zygmund inequalities, it follows that

$$\|u_n^\nu\|_{L^\infty((0,T);W^{1,p}(\Omega))} \le C(\|\omega_0\|_{L^p(\Omega)} + \|u_0\|_{L^2(\Omega)}). \tag{15}$$

Let $\varphi \in C_c^\infty((0,T) \times \overline{\Omega})$ be a divergence-free test vector field which is tangent to the boundary of Ω. We compute the time-derivative of u_n^ν in the sense of distributions. We have, using (13),

$$\langle \varphi, \partial_t u_n^\nu \rangle = -\int_0^T \int_\Omega (\partial_t \varphi) u_n^\nu$$

$$= \int_0^T \int_\Omega u_n^\nu (u_n^\nu \cdot \nabla)\varphi - 2\nu (D\varphi)_S : (Du_n^\nu)_S dx dt - \nu \int_0^T \int_{\partial\Omega} \alpha(\varphi \cdot \tau)(u_n^\nu \cdot \tau) dS.$$

Recall that $p > 2$, so that (15) implies that $\|u_n^\nu\|_{L^\infty((0,T)\times\Omega)} \leq C$ for some constant $C > 0$ depending only on the initial data. Similarly, $\|Du_n^\nu\|_{L^2((0,T)\times\Omega)}$ is bounded uniformly by a positive constant depending only on initial data. We use these facts to estimate $\partial_t u_n^\nu$. Let φ be a test vector field, which we first assume to be divergence free as above. We have

$$|\langle \varphi, \partial_t u_n^\nu \rangle| \leq \big(\|u_n^\nu\|_{L^\infty((0,T)\times\Omega)} \|u_n^\nu\|_{L^2((0,T)\times\Omega)} + 2\nu \|Du_n^\nu\|_{L^2((0,T)\times\Omega)}$$

$$+ \nu C \|\alpha u_n^\nu\|_{L^2((0,T);H^1(\Omega))} \big) \|\varphi\|_{L^2((0,T);H^1(\Omega))} \leq C \|\varphi\|_{L^2((0,T);H^1(\Omega))},$$

where we have used the continuity of the trace operator from $H^1(\Omega)$ onto $L^2(\partial\Omega)$ to estimate the boundary term. Now, if the test vector field φ is not divergence free, we use standard properties of the Leray projector P to obtain the estimate

$$\|P\varphi\|_{L^2((0,T);H^1(\Omega))} \leq C \|\varphi\|_{L^2((0,T);H^1(\Omega))},$$

and we repeat the argument above with $P\varphi$ in place of φ. Note that

$$\langle \varphi, \partial_t u_n^\nu \rangle = \langle P\varphi, \partial_t u_n^\nu \rangle$$

as $\partial_t u_n^\nu$ is divergence free and tangent to the boundary. By duality this implies the estimate

$$\|\partial_t u_n^\nu\|_{L^2((0,T);H^{-1}(\Omega))} \leq C, \tag{16}$$

with $C > 0$ depending only on the initial data. Thus u_n^ν is equicontinuous from $(0,T)$ to $H^{-1}(\Omega)$, and we can use the Aubin–Lions lemma to obtain a subsequence, which we will not relabel, converging strongly in $C([0,T]; L^2(\Omega))$. Without loss of generality this subsequence also converges weakly in $L^2((0,T); H^1(\Omega))$ to a limit u^ν. It is now easy to see that we can pass to the limit in each term in the weak formulation (2) of the Navier–Stokes equations, thereby concluding the proof of existence for the initial-boundary value problem (1). Furthermore, from the estimate on vorticity (14) it follows that

$$\|\omega^\nu(\cdot, t)\|_{L^p(\Omega)} \leq C$$

a.e. in time, for some constant $C > 0$ depending only on the initial data.

The uniqueness portion of this result is standard and may be obtained by adapting the classical argument using an energy estimate on the difference of two solutions with the same data. We conclude that there exists at most one weak solution $u \in C([0,T]; L^2(\Omega)) \cap L^2((0,T); H^1(\Omega))$ of (1). □

Remark 6. The proof above can be adapted for $1 < p \leq 2$, assuming of course that Lemma 3 could be proved in that case. The main steps in this adaptation would be the following:

- Substitute the L^∞ estimate on u_n^ν by an $L^\infty((0,T); L^{p^*})$ estimate, with p^* either the critical Sobolev exponent if $p < 2$ or an arbitrary number $1 < q < \infty$ if $p = 2$.
- Use the fact that $\sqrt{\nu}\|u_n^\nu\|_{L^2((0,T);H^1(\Omega))}$ is bounded uniformly in n and ν by the L^2-norm of the initial velocity. This is a consequence of standard energy estimates for the Navier–Stokes equations.

6. Inviscid limit and conclusions. Let $\omega_0 \in L^p(\Omega)$ for some $p > 2$ and let $u_0 = K_\Omega[\omega_0]$. In this last section we show that the sequence of solutions of the Navier–Stokes equations with initial velocity u_0 and with Navier friction conditions possesses a converging subsequence to a solution of the Euler equations with the same initial velocity as viscosity vanishes. The proof is very similar to the existence part of the proof of Proposition 1.

THEOREM 1. *Let $u^\nu = u^\nu(x,t)$ be the solution of (1) such that $u^\nu(\cdot,0) = u_0$. Then there exists a sequence $\nu_k \to 0$ such that $u^{\nu_k} \to u$ strongly in $C([0,T];L^2(\Omega))$ as $k \to \infty$ and u is a weak solution of the incompressible 2D Euler equations in the sense that*

$$\int_0^T \int_\Omega u\varphi_t + u(u\cdot\nabla)\varphi dxdt + \int_\Omega u_0\varphi(\cdot,0)dx = 0 \tag{17}$$

for any test vector field $\varphi \in C_c^\infty([0,T)\times\overline{\Omega})$ which is divergence free and tangent to the boundary.

Proof. We recall from the proof of Proposition 1 that the following uniform estimates hold for u^ν:

$$\|u^\nu\|_{L^\infty((0,T);W^{1,p}(\Omega))} \le C$$

and

$$\|\partial_t u^\nu\|_{L^2((0,T);H^{-1}(\Omega))} \le C,$$

where $C > 0$ depends only on the initial velocity u_0 and initial vorticity ω_0 and is independent of viscosity (see the proof of (15) and (16)). From these estimates it is possible to extract a subsequence u^{ν_k} which converges strongly in $C([0,T];L^2(\Omega))$ and weakly in $L^2((0,T);H^1(\Omega))$. It is easy to see that these modes of convergence are sufficient to pass to the limit in each term of (13) and guarantee that the limit function u satisfies the identity

$$\int_0^T \int_\Omega u\varphi_t + u(u\cdot\nabla)\varphi dxdt + \int_\Omega u_0\varphi(\cdot,0)dx = 0$$

for any test vector field $\varphi \in C_c^\infty([0,T)\times\overline{\Omega})$ which is divergence free and tangent to $\partial\Omega$. This is precisely the standard formulation of a weak solution of the Euler equations, and hence we conclude the proof. ∎

Remark 7. We note that relation (13) is the natural identity to try to pass to the limit in order to obtain (17). Furthermore, all the derivatives which appear in (13) either are applied to the test function or have a factor of ν, so that, in the limit $\nu \to 0$, there are no derivatives applied to the limit flow. This is relevant since solutions of Euler equations are less regular than solutions of Navier–Stokes equations.

We conclude this article with a few final observations. First, we call attention once more to the fact that the authors are not convinced of the criticality of $p = 2$, so the critical p remains an open problem. Second, as mentioned in section 3, there is no asymptotic description in the fluid mechanics literature of the boundary layer associated with the Navier friction condition, something which would clarify the issues raised here, irrespective of physical relevance. Finally, an interesting question which we have not explored is whether the viscosity weak solution of the incompressible 2D Euler equations obtained above conserves the L^p-norm of vorticity. Conservation

of the L^p-norm of vorticity holds both for weak solutions in the full plane, as a consequence of DiPerna–Lions theory (see [9]), and for strong solutions, as one can ascertain directly from the vorticity equation. In the viscous approximation, vorticity can be generated at the boundary, so that the question is whether this possibility disappears in the vanishing viscosity regime.

Acknowledgments. The authors would like to thank D. Iftimie and M. O. Souza for enlightening discussions. We would also like to thank the referees for several comments which improved the presentation and for pointing out the references [5, 6].

REFERENCES

[1] T. Clopeau, A. Mikelić, and R. Robert, *On the vanishing viscosity limit for the* 2*D incompressible Navier-Stokes equations with the friction type boundary conditions*, Nonlinearity, 11 (1998), pp. 1625–1636.

[2] J.-M. Delort, *Existence de nappes de tourbillon en dimension deux*, J. Amer. Math. Soc., 4 (1991), pp. 553–586.

[3] B. Desjardins and E. Grenier, *Linear instability implies nonlinear instability for various types of viscous boundary layers*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 20 (2003), pp. 87–106.

[4] C. He and L. Hsiao, *Two-dimensional Euler equations in a time dependent domain*, J. Differential Equations, 163 (2000), pp. 265–291.

[5] W. Jäger and A. Mikelić, *On the roughness-induced effective boundary conditions for an incompressible viscous flow*, J. Differential Equations, 170 (2001), pp. 96–122.

[6] W. Jäger and A. Mikelić, *Couette flows over a rough boundary and drag reduction*, Comm. Math. Phys., 232 (2003), pp. 429–455.

[7] G. Lieberman, *Second Order Parabolic Differential Equations*, World Scientific, River Edge, NJ, 1996.

[8] J. L. Lions, *Quelques Méthodes de Resolution des Problémes aux Limites non Linéaires*, Dunod, Gauthier-Villars, Paris, 1969.

[9] P. L. Lions, *Mathematical Topics in Fluid Mechanics. Vol.* I*, Incompressible Models*, Oxford Lecture Ser. Math. Appl. 3, Clarendon Press, Oxford, 1996.

[10] J.-G. Liu and Z. Xin, *Convergence of vortex methods for weak solutions to the* 2*-D Euler equations with vortex sheet data*, Comm. Pure Appl. Math., 48 (1995), pp. 611–628.

[11] A. J. Majda, *Remarks on weak solutions for vortex sheets with a distinguished sign*, Indiana Univ. Math. J., 42 (1993), pp. 921–939.

[12] R. L. Panton, *Incompressible Flow*, John Wiley and Sons, New York, 1984.

[13] S. Schochet, *The point-vortex method for periodic weak solutions of the* 2*-D Euler equations*, Comm. Pure Appl. Math., 49 (1996), pp. 911–965.

[14] R. Temam and X. Wang, *Boundary layers associated with incompressible Navier-Stokes equations: The noncharacteristic boundary case*, J. Differential Equations, 179 (2002), pp. 647–686.

[15] I. Vecchi and S. Wu, *On* L^1*-vorticity for* 2*-D incompressible flow*, Manuscripta Math., 78 (1993), pp. 403–412.

[16] M. Vishik, *Incompressible flows of an ideal fluid with vorticity in borderline spaces of Besov type*, Ann. Sci. École Norm. Sup. (4), 32 (1999), pp. 769–812.

[17] V. I. Yudovich, *Non-stationary flows of an ideal incompressible fluid*, Ž. Vyčisl. Mat. i Mat. Fiz., 3 (1963), pp. 1032–1066 (in Russian).

[18] V. I. Yudovich, *Uniqueness theorem for the basic nonstationary problem in the dynamics of an ideal incompressible fluid*, Math. Res. Lett., 2 (1995), pp. 27–38.

SIAM J. MATH. ANAL.
Vol. 36, No. 4, pp. 1142–1158

THE CONVEX SCATTERING SUPPORT IN A BACKGROUND MEDIUM*

STEVEN KUSIAK† AND JOHN SYLVESTER‡

Abstract. We discuss inverse problems for the Helmholtz equation at fixed energy, specifically the inverse source problem and the inverse scattering problem from a medium or an obstacle. In [S. Kusiak and J. Sylvester, *Comm. Pure Appl. Math.*, 56 (2003), pp. 1525–1548], we introduced the *convex scattering support* of a far field, a set which will be a subset of the convex hull of the support of any source or scattering inhomogeneity which can produce it.

We extend these results and modify the methods to locate a source within a known inhomogeneous background medium, or a deviation from that medium, using observations of a single far field. We also describe some numerical examples that illustrate the robustness of the method.

Key words. inverse scattering, Helmholtz equation, partial differential equations

AMS subject classifications. 81U40, 74J25, 65N21

DOI. 10.1137/S0036141003433577

1. Introduction.

We study an inverse problem for the Helmholtz equation at fixed energy. Our aim is to deduce the location of the source or scatterer from observations of scattered waves made at a distance, which are called *far fields*. Typically, one has access to several far fields. For the inverse medium problem, the index of refraction is uniquely determined by the full scattering kernel, i.e., the observed scattered field for every possible incident wave. In special cases [9, 5, 6], substantial information about the support of the scatterer has been obtained from the scattered field of a few, or even only one, incident wave.

In [8], we showed that, in a homogeneous background medium, we could associate the *convex scattering support* with a single far field. This set is the smallest convex set which supports a source that can produce that far field. We also produced a test, *the circular Paley–Wiener theorem*, for computing the convex scattering support in two dimensions. In [10], we introduced a different numerical method, called the *range test*, for computing this support in a two-dimensional homogeneous medium.

Our work was motivated by the linear sampling method of Colton and Kirsch (see [2]). They first developed a *Picard test*, which determines whether a far field belongs to the range of the (compact) scattering operator, as a tool for inverse scattering. This method, and the subsequent factorization method of Kirsch [7], differ from what we present here in that they require much more data (the full scattering map) and compute much more (the exact support of the scatterer).

In section 2 of this paper, we introduce the necessary scattering formalism and restate the circular Paley–Wiener theorem as a *Picard test*. This restatement, though less explicit, generalizes directly to inhomogeneous media and higher dimensions.

In section 3 we produce this general test. The general test tells us if a far field could have been produced by a source or a scatterer located within a specific domain,

*Received by the editors August 21, 2003; accepted for publication (in revised form) March 5, 2004; published electronically January 27, 2005. This work was supported by NSF grant DMS-0099838.

http://www.siam.org/journals/sima/36-4/43357.html

†Department of Applied Mathematics, University of Washington, Seattle, WA 98195 (kusiak@amath.washington.edu).

‡Department of Mathematics, University of Washington, Seattle, WA 98195 (sylvest@math.washington.edu). The research of this author was supported by ONR grant N00014-93-1-0295.

but not whether the true source was located there. Section 4 addresses this issue by extending the concept of convex scattering support to inhomogeneous media. The conclusion is that we can locate a smallest convex set, which must be contained in the convex hull of the support of any source which radiates that far field. Conversely, we produce a source, supported in any neighborhood of the convex scattering support which does radiate that far field.

Section 5 discusses the relationship between the support of a scatterer, rather than a source, and the convex scattering support. In this case we show that the convex scattering support provides a lower bound for the convex hull of the scatterer. Unlike the source case, we don't expect this lower bound to be optimal.

Section 6 contains a description of an explicit algorithm and some numerical results. Maybe the most important observation in this section is that the practical implementation of the algorithm is much simpler and more robust than the theorem guarantees.

2. Far fields in a homogeneous medium. We model the time harmonic wave radiated by a source in a homogeneous medium as a solution to the inhomogeneous Helmholtz equation:

$$(\Delta + k^2)u(x) = f(x), \quad x \in \mathbb{R}^n. \tag{2.1}$$

Equation (2.1) has a unique *outgoing solution*, $u = G_0^+ f$, which can be computed by the limiting absorption principle (see, e.g., [11, p. 147]).

$$\begin{aligned} G_0^+ f &= \lim_{\epsilon \downarrow 0} (\Delta + (k - i\epsilon)^2)^{-1} f \\ &= -\lim_{\epsilon \downarrow 0} \int_{\mathbb{R}^n} \frac{e^{i\langle x,\xi\rangle} \widehat{f}(\xi)}{|\xi|^2 - (k - i\epsilon)^2} d\xi. \end{aligned} \tag{2.2}$$

The limiting absorption principle chooses the unique solution u of (2.1) which extends to be holomorphic in $\{\mathrm{Im}(k) \le 0\}$ and is continuous up to the boundary. According to (one of the many theorems called) the Paley–Wiener theorem, this solution is the Fourier transform of the unique solution $\tilde{u}$ of the wave equation which is zero in the past. That is,

$$u(k, x) = \int_0^\infty e^{-ikt} \tilde{u}(x, t) dt.$$

We call a function in the range of G_0^+ outgoing. We shall refer to a function as incoming if $v = G_0^- f$, where

$$G_0^- f = \lim_{\epsilon \downarrow 0} (\Delta + (k + i\epsilon)^2)^{-1} f.$$

Alternatively, $u = G_0^+ f$ may be characterized as the unique solution of (2.1) satisfying the Sommerfeld radiation condition:

$$\lim_{r \to \infty} r^{\frac{n-1}{2}} (\partial_r u - iku) = 0, \quad r = |x|.$$

Inverting the Fourier transform in (2.2), we may also represent $u = G_0^+ f$ (cf. [4]) as

$$(G_0^+ f)(x) := -\frac{i}{4} \left(\frac{k}{2\pi} \right)^{\frac{n-2}{2}} \int_{\mathbb{R}^n} |x - y|^{\frac{2-n}{2}} H^{(1)}_{\frac{n-2}{2}}(k|x - y|) f(y) dy. \tag{2.3}$$

Here, $H^{(1)}_{(n-2)/2}$ is the Hankel function of the first kind. The representation of G_0^- uses the other Hankel function; i.e., its kernel is the complex conjugate of the kernel of G_0^+.

The simplest estimate for the solution operators $G_0^\pm$ is on the weighted L^2 spaces, $\mathrm{H}^s_\delta(\mathbb{R}^n)$. For $\delta = 0$, $\mathrm{H}^s_0(\mathbb{R}^n)$ is the Sobolev space, $\mathrm{H}^s(\mathbb{R}^n)$. For $\delta > 0$,

$$\|f\|_{s,\delta} = \|(1+|x|^2)^{\delta/2} f\|_{s,0}.$$

PROPOSITION 2.1. *The operators $G_0^\pm$ are bounded as maps between the weighted L^2 spaces*

$$G_0^\pm : \mathrm{H}^s_\delta(\mathbb{R}^n) \longrightarrow \mathrm{H}^{s+2}_{-\delta}(\mathbb{R}^n) \tag{2.4}$$

for any real s and any $\delta > \frac{1}{2}$. Moreover, G_0^- is the Hilbert space adjoint of G_0^+ on L^2 (i.e., $s = 0$ in (2.4)). That is,

$$G_0^{+\,*} = G_0^-.$$

Proof. The estimate was first proved in [1]. Once we have it, it is a simple matter to interchange the order of integration in the L^2 pairing to check that the two operators are adjoints. □

The far field describes the asymptotics of u as $|x| \to \infty$. Stationary phase applied to (2.2) or Hankel function asymptotics applied to (2.3) yields

$$u(x) \sim \frac{e^{ik|x|}}{|x|^{(n-1)/2}} C_{n,k} \int_{\mathbb{R}^n} e^{-ik\langle \Theta, y\rangle} f(y)dy, \quad |x| \to \infty, \tag{2.5}$$

where $\Theta = \frac{x}{|x|}$ is a unit vector on the $n-1$-dimensional sphere S^{n-1} and

$$C_{n,k} = \frac{-i}{\sqrt{8\pi}} \left(\frac{k}{2\pi}\right)^{\frac{n-2}{2}} e^{-i(n-1)\pi/4}.$$

Hence, given a source f we define the far field, $u_\infty = F_0 f$, by

$$(F_0 f)(\Theta) = \int_{\mathbb{R}^n} e^{-ik\langle \Theta, y\rangle} f(y)dy \tag{2.6}$$

$$= \hat{f}(k\Theta). \tag{2.7}$$

The mapping properties of F_0^+ are important for us.

PROPOSITION 2.2. *F_0^+ is a compact linear map*

$$F_0^+ : \mathrm{H}^s_\delta(\mathbb{R}^n) \longrightarrow L^2(S^{n-1}).$$

Its adjoint with respect to the distributional (not the Hilbert space) pairing

$$F_0^{+\,\dagger} : L^2(S^{n-1}) \longrightarrow \mathrm{H}^{-s}_{-\delta}(\mathbb{R}^n)$$

is the Herglotz operator

$$(F_0^{+\,\dagger}\alpha)(x) = (\mathcal{H}f)(x) \tag{2.8}$$

$$= \int_{S^{n-1}} e^{ik\langle \Theta, x\rangle} \alpha(\Theta) dS(\Theta).$$

Remark 2.3. Functions in the range of the Herglotz operator are usually referred to as incident fields. They are the $\mathrm{H}^s_{-\delta}$ solutions of the homogeneous (or free) Helmholtz equation for all real s and any $\delta > \frac{1}{2}$. The Herglotz operator represents these free solutions as superpositions of plane waves.

Proof. The boundedness of F_0^+ follows easily from the representation (2.7). We need only note that if $f \in \mathrm{H}^s_\delta$, then $\hat{f} \in \mathrm{H}^\delta_s$. As $\delta > \frac{1}{2}$, the restriction map from $\mathrm{H}^\delta_s(\mathbb{R}^n)$ to L^2 of the codimension one sphere, S^{n-1}, is a compact operator.

The boundedness of $F_0^\dagger$ follows from the boundedness of F_0. The equality (2.8) can be seen by using formula (2.6), pairing with an L^2 far field, and interchanging the order of integration. Nevertheless, we give a proof that relies more on scattering.

Let $u = G_0^+ f$ and $v = \mathcal{H}\alpha$. Stationary phase shows that v has the asymptotics

$$v(x) \sim C_{n,k} \frac{e^{ik|x|}}{|x|^{(n-1)/2}} \alpha(\theta) + C_{n,k} \frac{e^{-ik|x|}}{|x|^{(n-1)/2}} \alpha(-\theta), \quad |x| \to \infty. \tag{2.9}$$

Applying Green's theorem on the ball of radius R gives

$$\int_{\partial B_R} v \frac{\partial u}{\partial \nu} - \frac{\partial v}{\partial \nu} u = \int_{B_R} v(\Delta + k^2)u - (\Delta + k^2)vu. \tag{2.10}$$

Letting $R \to \infty$ and making use of (2.5) and (2.9) allows us to evaluate the left-hand side of (2.10). Recalling that v is a free solution of the Helmholtz equation removes the second term from the left-hand side so that

$$\int_{S^{n-1}} \alpha(\theta)(F_0 f)(\theta) dS(\theta) = \int_{B_R} vf$$
$$\langle \alpha, F_0 f \rangle = \langle \mathcal{H}\alpha, f \rangle. \qquad \square$$

Proposition 2.2 has a more general statement which will prove convenient in the next section.

THEOREM 2.4. *Suppose that $s_1 + s_2 > -2$, $\delta > \frac{1}{2}$, and that u and v satisfy*

$$u \in \mathrm{H}^{s_1+2}_{-\delta}(\mathbb{R}^n) \quad \textit{and} \quad (\Delta + k^2)u \in \mathrm{H}^{s_1}_\delta(\mathbb{R}^n), \tag{2.11}$$

$$v \in \mathrm{H}^{s_2+2}_{-\delta}(\mathbb{R}^n) \quad \textit{and} \quad (\Delta + k^2)v \in \mathrm{H}^{s_2}_\delta(\mathbb{R}^n); \tag{2.12}$$

then as $|x| \to \infty$,

$$u(x) \sim \frac{e^{ik|x|}}{|x|^{(n-1)/2}} u^+_\infty(\theta) + \frac{e^{-ik|x|}}{|x|^{(n-1)/2}} u^-_\infty(\theta),$$
$$v(x) \sim \frac{e^{ik|x|}}{|x|^{(n-1)/2}} v^+_\infty(\theta) + \frac{e^{-ik|x|}}{|x|^{(n-1)/2}} v^-_\infty(\theta), \tag{2.13}$$

and

$$\langle u^+_\infty, v^-_\infty \rangle - \langle u^-_\infty, v^+_\infty \rangle = \langle u, (\Delta + k^2)v \rangle - \langle (\Delta + k^2)u, v \rangle. \tag{2.14}$$

Remark 2.5. When we write "$\sim$" meaning "asymptotic to," we mean classical asymptotics only in the case that u happens to be smooth near infinity. This is always the case if $(\Delta + k^2)u$ has compact support. In the more general setting, the space of functions that satisfy (2.11) (or (2.12)) form a Hilbert space, and we are asserting that the mappings

$$u \mapsto u^\pm_\infty$$

extend by continuity as mappings from that Hilbert space into $L^2(S^{n-1})$.

Proof. Suppose first that u and v are compactly supported and smooth. Then every such u (and v) is a linear combination of an outgoing function and a solution of the homogeneous Helmholtz equation, i.e., an outgoing wave plus a Herglotz function. Each of these has the asymptotics asserted in (2.5) and in (2.9), and therefore their sum had these asymptotics as well. Now apply Green's formula as in (2.10) to obtain (2.14).

Finally, notice that the left-hand side of (2.14) is a continuous bilinear functional with respect to L^2 convergence and the right-hand side is continuous when $u_n \to u$ and $v_n \to v$ in the topologies of (2.11). □

DEFINITION 2.6. *We shall refer to u^+_∞ as the outgoing far field of u and u^-_∞ as its incoming far field. An outgoing function has zero incoming far field, i.e., $u^-_\infty = 0$. A Herglotz function has outgoing and incoming far fields related by the antipodal map* (2.9)*. Intuitively, one can see this by thinking about a spherical incoming wave passing through the origin to become an outgoing wave.*

In [8], we began to study the *scattering support* of a far field. The first step is to ask whether a far field could have been produced by a source which is a distribution supported in a closed set. We recall that the restricting of a distribution to an open set means restricting it to act on the subspace $C_0^\infty(\Omega)$ of $C_0^\infty(\mathbb{R}^n)$. The support of a distribution is the closed set defined below.

DEFINITION 2.7. *A point x belongs to the support of a distribution f if there exists no open neighborhood, O_x, such that $f|_{O_x} = 0$.*

Distributions supported on a closed set form natural subspaces of $\mathrm{H}^s_\delta(\mathbb{R}^n)$.

DEFINITION 2.8. *$\mathrm{H}^s_0(\Omega)$ is the closed subspace of $\mathrm{H}^s_\delta(\mathbb{R}^n)$ consisting of those distributions which are supported in $\overline{\Omega}$.*

The definition is independent of δ as long as Ω is bounded.

We point out that this is different from $H^s(\Omega)$, which denotes the restrictions of distributions to a bounded open set, and is not a subspace of any $\mathrm{H}^s_\delta(\mathbb{R}^n)$. In fact, $\mathrm{H}^{-s}(\Omega)$ is the natural dual to $\mathrm{H}^s_0(\Omega)$ for all real s. For bounded open sets Ω and positive s, our definition of $\mathrm{H}^s_0(\Omega)$ coincides with the common definition, i.e., the closure of $C_0^\infty(\Omega)$ in the H^s norm. In [8], we proved the following theorem.

THEOREM 2.9. *Let $\alpha \in L^2(S^1)$ represent a far field. There exists $f \in \mathrm{H}^s_0(B_R)$ such that*

$$F_0^+ f = \alpha = \sum_{n=-\infty}^{\infty} \alpha_n e^{in\theta} \tag{2.15}$$

if and only if

$$\sum_{n=-\infty}^{\infty} \left| \frac{\alpha_n n^s}{\sigma_n(R)} \right|^2 < \infty, \tag{2.16}$$

where

$$\sigma_n(R) = \left(\int_0^R |J_n(kr)|^2 r\,dr \right)^{\frac{1}{2}}.$$

In section 3 of this paper, we will prove a generalization of this result to variable index of refraction, higher dimensions, and more general domains. We close this section with a restatement of this theorem which anticipates the generalization to follow.

We don't give the proof here, as it will follow as a corollary of the more general Theorem 3.6 in section 3.

THEOREM 2.10. *Let $\alpha \in L^2(S^{n-1})$ represent a far field. Let $F_0^+|_{\mathrm{H}_0^s(\Omega)}$ represent the restriction of the compact operator F_0^+ to $\mathrm{H}_0^s(\Omega)$ and let*

$$F_0^+|_{\mathrm{H}_0^s(\Omega)} = \sum \sigma_n \psi_n \otimes \overline{\phi_n}$$

be its singular value decomposition. Then

$$\alpha \in \mathrm{Range}(F_0^+|_{\mathrm{H}_0^s(\Omega)})$$

if and only if

$$\sum \left| \frac{(\alpha, \psi_n)}{\sigma_n} \right|^2 < \infty. \tag{2.17}$$

To facilitate the comparison of Theorems 2.9 and 2.10, we describe two examples with Ω equal to $B_R \in \mathbb{R}^2$, the ball of radius R centered at the origin. We can separate variables in this case, representing the operator F_0^+ in terms of complex exponentials, Bessel functions, and the characteristic function of the ball, χ_{B_R}.

$$F_0^+|_{L^2(B_R)} = \sum_{n=-\infty}^{\infty} \mathrm{e}^{in(\theta - \frac{\pi}{2})} \otimes \chi_{B_R} \mathrm{e}^{-in(\phi - \frac{\pi}{2})} J_n(kr). \tag{2.18}$$

Because we are in L^2, its Hilbert space adjoint is

$$(F_0^+|_{L^2(B_R)})^* = \sum_{n=-\infty}^{\infty} \chi_{B_R} \mathrm{e}^{in(\phi - \frac{\pi}{2})} J_n(kr) \otimes \mathrm{e}^{-in(\theta - \frac{\pi}{2})}$$

and its singular values are the eigenvalues of $(F_0^+|_{L^2(B_R)} F_0^+|^*_{L^2(B_R)})$

$$\sigma_n^2 = 4\pi^2 \int_0^R J_n^2(ks) s ds,$$

so we see that (2.16) and (2.17) agree in the case $s = 0$.

Next, we consider $F_0^+|_{\mathrm{H}_0^1(B_R)}$. The operator itself has the same representation as in (2.18); we are just considering it on a smaller subspace. A bit of a calculation shows that the Hilbert space adjoint is now

$$(F_0^+|_{\mathrm{H}_0^1(B_R)})^* = \sum_{n=-\infty}^{\infty} \chi_{B_R} \mathrm{e}^{in(\phi - \frac{\pi}{2})} \left(J_n(kr) - \left(\frac{r}{R}\right)^{|n|} J_n(kR) \right) \otimes \mathrm{e}^{-in(\theta - \frac{\pi}{2})}$$

with singular values (we'll call them $\tilde{\sigma}_n$'s)

$$\tilde{\sigma}_n^2 = 4\pi^2 \left[\int_0^R J_n^2(ks) s ds - \frac{J_n(kR)}{R^n} \int_0^R J_n(ks) s^{n+1} ds \right]. \tag{2.19}$$

Now, the large n asymptotics of the Bessel function is

$$J_n(kr) \sim \frac{1}{\sqrt{\pi n}} \left(\frac{ekr}{2n} \right)^n \quad \text{for } n \gg kr.$$

Integrating with respect to r then gives

$$\sigma_n^2 \sim \frac{8\pi}{(ek)^2}\left(\frac{ekR}{2n}\right)^{2n+2}\left(1+O\left(\frac{1}{n}\right)\right), \quad n\to\infty,$$

while the leading order asymptotics of the two terms in (2.19) cancel, so that

$$\tilde{\sigma}_n^2 = O\left(\frac{\sigma_n^2}{n^2}\right),$$

which agrees with (2.16) in the case $s=1$.

3. Far fields in an inhomogeneous medium. If the medium is inhomogeneous, (2.1) is replaced by

$$(\Delta + k^2 n(x))u(x) = f(x), \quad x\in\mathbb{R}^n. \tag{3.1}$$

The coefficient $n(x)$ is the index of refraction and is the square of the reciprocal of the wave speed at x. We will assume that n has positive imaginary part, that $n-1$ is compactly supported, and that $n\in L^p(\mathbb{R}^n)$ for $p>\max(n-2,n/2)$. The unique continuation principle holds for this is the class of n's.

It will be convenient to rewrite (3.1) as

$$(\Delta + k^2 - q(x))u = f$$

with $q=k^2(n-1)$. Because q is not necessarily smooth, we must restrict the regularity of the Sobolev spaces. We want to allow single and double layer potentials as sources, so that the application of our results to active scattering will include scattering from an obstacle. We will treat $f\in \mathrm{H}_\delta^{s-2}(\mathbb{R}^n)$ with $0\le s\le n/p$ and $\delta>\frac{1}{2}$.

We will standardize the notation used in this section. We will use η and σ to denote unrestricted real numbers. The symbols δ, p, and s will always satisfy the inequalities

$$\begin{gathered}\delta > 1/2,\\ p>\max(2,n/2),\\ 0\le s\le n/p.\end{gathered} \tag{3.2}$$

Our next theorem asserts the existence of the analogues of G_0 and F_0.

THEOREM 3.1. *Let $q\in L^p(\mathbb{R}^n)$ and have compact support. Let $f\in \mathrm{H}_\delta^{s-2}(\mathbb{R}^n)$ with p, s, δ satisfying* (3.2). *Then there exists a unique outgoing (resp., incoming) solution of*

$$(\Delta + k^2 - q(x))u = f$$

which has the asymptotic behavior

$$u(x)\sim \frac{e^{ik|x|}}{|x|^{(n-1)/2}}u_\infty^\pm(\Theta), \quad |x|\to\infty.$$

Moreover, the unique solution u is computed by the operator

$$\begin{aligned} u(x) &= (G_q^\pm f)(x)\\ &:= (I-G_0^\pm q)^{-1}G_0^\pm f \qquad (3.3)\\ &= G_0^\pm(I-qG_0^\pm)^{-1}f \qquad (3.4)\end{aligned}$$

and

$$\begin{aligned} u_\infty^\pm(\Theta) &= (F_q^\pm f)(\Theta) \\ &:= F_0^\pm (I - qG_0^\pm)^{-1} f. \end{aligned} \tag{3.5}$$

Additionally, both $G_q^\pm$ and $F_q^\pm$ are compact operators:

$$G_q^\pm : \mathrm{H}_\delta^{s-2}(\mathbb{R}^n) \to \mathrm{H}_{-\delta}^{s-n/p}(\mathbb{R}^n), \tag{3.6}$$

$$F_q^\pm : \mathrm{H}_\delta^{s-2}(\mathbb{R}^n) \to L^2(S^{n-1}). \tag{3.7}$$

The proof requires that we show that $(I - G_0^\pm q)$ is invertible. Let M_q denote the operator of multiplication by q. Then we have the following lemma.

LEMMA 3.2. *Let q be a compactly supported function on $\mathbb{R}^n$. For any real $p \geq 2$, any real δ and η, and any $0 \leq s \leq \frac{n}{p}$,*

$$M_q : \mathrm{H}_{\eta_1}^s(\mathbb{R}^n) \to \mathrm{H}_{\eta_2}^{s-\frac{n}{p}}(\mathbb{R}^n) \tag{3.8}$$

is bounded.

Proof. According to Hölder's inequality,

$$\begin{aligned} \|qu\|_{L^2(\mathbb{R}^n)} &\leq \|q\|_{L^p(\mathbb{R}^n)} \|u\|_{L^{\frac{2p}{p-2}}(\mathbb{R}^n)} \\ &\leq \|q\|_{L^p(\mathbb{R}^n)} \|u\|_{\mathrm{H}^{\frac{n}{p}}(\mathbb{R}^n)}, \end{aligned}$$

with the second line a consequence of the Sobolev inequality. Thus

$$M_q : \mathrm{H}_0^{\frac{n}{p}}(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$$

is bounded. Duality implies that

$$M_q : L^2(\mathbb{R}^n) \to \mathrm{H}_0^{-\frac{n}{p}}(\mathbb{R}^n)$$

is also bounded. Interpolation then gives (3.8) in the case that $\delta = \eta = 0$. However, because q is compactly supported,

$$\left\| (1 + |x|^2)^{\frac{\eta - \delta}{2}} q \right\|_{L^p(\mathbb{R}^n)} \leq C \|q\|_{L^p(\mathbb{R}^n)},$$

which implies (3.8) for any δ and any η. ∎

As a consequence, we have the following corollary.

COROLLARY 3.3. *Let $p \geq 2$, $0 \leq s \leq \frac{n}{p}$, and $\delta > \frac{1}{2}$. Then*

$$G_0^+ q : \mathrm{H}_\eta^s(\mathbb{R}^n) \to \mathrm{H}_{-\delta}^{s+2-\frac{n}{p}}(\mathbb{R}^n), \tag{3.9}$$

and

$$qG_0^+ : \mathrm{H}_\delta^{s-2}(\mathbb{R}^n) \to \mathrm{H}_\eta^{s-\frac{n}{p}}(\mathbb{R}^n)$$

is bounded. If, in addition, $p > \frac{n}{2}$, then

$$G_0^\pm q : \mathrm{H}_{-\delta}^s(\mathbb{R}^n) \to \mathrm{H}_{-\delta}^s(\mathbb{R}^n)$$

are compact.

Proof. The first statement is a direct consequence of Proposition 2.1 and Lemma 3.2, while the second follows from the compact embedding of $\mathrm{H}^{s_1}_{\eta}$ in $\mathrm{H}^{s_2}_{\delta}$ whenever $\eta > \delta$ and $s_1 > s_2$. ∎

COROLLARY 3.4. *Let $q \in L^p(\mathbb{R}^n)$ with compact support and let p, s, δ satisfy* (3.2). *Then $(I - G_0^+ q)^{-1}$ exists as a bounded linear operator from $\mathrm{H}^s_{-\delta}(\mathbb{R}^n)$ to $\mathrm{H}^s_{-\delta}(\mathbb{R}^n)$.*

Proof. Corollary 3.3 implies that $(I - G_0^+ q)$ is Fredholm, so we only need to show uniqueness. Suppose that $u \in \mathrm{H}^s_{-\delta}(\mathbb{R}^n)$ satisfies

$$(I - G_0^+ q)u = 0. \tag{3.10}$$

Repeated application of (3.9) shows us that $u \in \mathrm{H}^2_{-\delta}(\mathbb{R}^n)$ and satisfies

$$(\Delta + k^2)u = qu.$$

Since u is outgoing, $\overline{u}$ is incoming and satisfies

$$(\Delta + k^2)\overline{u} = \overline{qu}$$

so that we may apply (2.14) to obtain the identities

$$2ik \left\langle \overline{u^+_\infty}, u_\infty \right\rangle = \langle \overline{qu}, u\rangle - \langle \overline{u}, qu\rangle,$$
$$2ik\|u^+_\infty\|^2_{L^2} = 2i \int \operatorname{Im} q|u|^2,$$

which implies

$$\|u^+_\infty\|^2_{L^2} \le 0.$$

Thus, u is an outgoing function with no far field. Rellich's lemma [3] implies that u vanishes outside the support of q, and unique continuation then implies that $u = 0$ everywhere. ∎

COROLLARY 3.5. *Let $q \in L^p(\mathbb{R}^n)$ with compact support and let p, s, δ satisfy* (3.2). *Then $(I - qG_0^+)^{-1}$ exists as a bounded linear operator from $\mathrm{H}^{s-2}_{\delta}(\mathbb{R}^n)$ to $\mathrm{H}^{s-2}_{\delta}(\mathbb{R}^n)$.*

Proof. $(I - qG_0^+)^{-1}$ is also Fredholm, so only uniqueness need be checked. Suppose $f = qG_0^+ f$. Then $u = G_0^+ f$ satisfies (3.10), and therefore it must be zero. However, $f = (\Delta + k^2)u$, so f must also vanish. ∎

Proof of Theorem 3.1. We have shown that formulas (3.3), (3.4), and (3.5) make sense. Both (3.6) and (3.7) follow from the previous corollaries and the mapping properties of F_0 and G_0. To verify that (3.3) and (3.4) are equal, we start with the factorization

$$(I - G_0^\pm q)G_0^\pm = G_0^\pm(I - qG_0^\pm).$$

Now, both $(I - G_0^\pm q)$ and $(I - qG_0^\pm)$ are invertible, so that

$$(I - G_0^\pm q)^{-1}(I - G_0^\pm q)G_0^\pm(I - qG_0^\pm)^{-1} = (I - G_0^\pm q)^{-1}G_0^\pm(I - qG_0^\pm)(I - qG_0^\pm)^{-1},$$

which implies

$$G_0^\pm(I - qG_0^\pm)^{-1} = (I - G_0^\pm q)^{-1}G_0^\pm. \qquad ∎$$

With Theorem 3.1 in place, we may extend the Picard test (2.17) to the inhomogeneous equation.

THEOREM 3.6. *Let $\alpha \in L^2(S^{n-1})$ represent a far field and let $F_q^+|_{\mathrm{H}_0^s(\Omega)}$ represent the restriction of the compact operator F_q^+ to $\mathrm{H}_0^s(\Omega)$. Then*

$$\alpha \in \mathrm{Range}(F_q^+|_{\mathrm{H}_0^s(\Omega)})$$

if and only if

$$\sum \left| \frac{(\alpha_n, \psi_n)}{\sigma_n} \right|^2 < \infty,$$

where

$$F_q^+|_{\mathrm{H}_0^s(\Omega)} = \sum \sigma_n \psi_n \otimes \overline{\phi_n}$$

is the singular value decomposition of $F_q^+|_{\mathrm{H}_0^s(\Omega)}$.

This theorem tells us that we can look for a source in a known inhomogeneous background by simply replacing F_0 by F_q in the convergence test given in Theorem 2.10. Of course, to apply it we must numerically or analytically compute the singular value decomposition of $F_q^+|_{\mathrm{H}_0^s(\Omega)}$.

4. The convex scattering support. We are ready to use Theorem 3.6 to locate the support of a source in an inhomogeneous medium. In Theorem 5.2 of section 5, we will locate the region where a medium differs from a known background by applying this test. In both cases our data will be a single far field.

As we pointed out in [8], a single far field is not enough information to uniquely determine the support of a source. For example, if ϕ has compact support, then $f_\phi = (\Delta + k^2)\phi$ will always have zero far field. We can always add f_ϕ to a source to produce a new one with bigger support which produces the same far field. Thus we cannot associate with a far field a set which contains the support of any source which produces it.

However, we can determine a unique smallest convex set which must be a subset of the convex hull of the support of any source which produces that far field. We refer to this set as the *convex scattering support* of a far field. We will show below that the convex scattering support of any nonzero far field is a nonempty closed set, and that there always exists an L^2 source, supported in an arbitrarily small neighborhood of the convex scattering support, which will reproduce the far field.

We begin with the definition.

DEFINITION 4.1. *The convex scattering support of the far field u_∞, with respect to the background q, is*

$$\mathrm{cS}_k\mathrm{supp}_q u_\infty = \bigcap_{\substack{F_q f = u_\infty \\ f \in \mathrm{H}_\delta^s(\mathbb{R}^n)}} \mathrm{ch}(\mathrm{supp}\, f). \tag{4.1}$$

Here, $\mathrm{ch}(\mathrm{supp}\, f)$ *denotes the convex hull of the support of f.*

We must take $s > -2$ because F_q is only defined for such sources. The next lemma asserts that the $\mathrm{cS}_k\mathrm{supp}_q u_\infty$ doesn't depend on s for $-2 < s \leq 0$. It doesn't depend on s at all if q is smooth. We will use the notation $N_\epsilon(\Omega)$ to denote an open epsilon neighborhood of a set Ω.

LEMMA 4.2. *For any $f \in \mathrm{H}_0^s(\Omega)$ and any $\varepsilon > 0$, there exists $\tilde{f} \in L^2(N_\varepsilon(\Omega))$ such that*

$$F_q^+ \tilde{f} = F_q^+ f.$$

Proof. Let $u = G_q^+ f$ and let $\phi \in C^\infty(\mathbb{R}^n)$ satisfy

$$\phi = \begin{cases} 1, & x \in \mathbb{R}^n \backslash N_\varepsilon(\Omega), \\ 0, & x \in N_\varepsilon(\Omega), \end{cases}$$

and set

$$\tilde{f} = (\Delta + k^2 - q(x))\phi u.$$

Now, $G_q^+ f = u$ outside $N_\varepsilon(\Omega)$ and therefore has the same far field. Note that ϕu is supported outside $\operatorname{supp} f$ so that u is H^2 there, and thus $\tilde{f} \in L^2$. ∎

THEOREM 4.3. *For any far field $\alpha \in L^2(S^{n-1})$ with a compactly supported source, and any $\varepsilon > 0$, there exists an L^2 source f_ε such that $G_q^+ f_\varepsilon = \alpha$ and*

$$\operatorname{ch}(\operatorname{supp} f_\epsilon) \subset N_\epsilon(\mathrm{cS}_k \operatorname{supp}_q \alpha).$$

We shall need two lemmas for the proof.

LEMMA 4.4. *Suppose $\operatorname{supp} f_1 \subset \Omega_1$, $\operatorname{supp} f_2 \subset \Omega_2$, and that $\mathbb{R}^n \setminus (\Omega_1 \bigcup \Omega_2)$ is connected and contains a neighborhood of ∞. If*

$$F_q^+ f_1 = F_q^+ f_2 = \alpha,$$

then, for any $\varepsilon > 0$, there exists an $f_3 \in C^\infty(\mathbb{R}^n)$ with

$$\operatorname{supp} f_3 \subset N_\varepsilon(\Omega_1 \cap \Omega_2)$$

and

$$F_q^+ f_3 = \alpha.$$

Proof. According to Rellich's lemma and unique continuation [3], $u_1 = G_q^+ f_1$ and $u_2 = G_q^+ f_2$ agree on the $\mathbb{R}^n \setminus (\Omega_1 \bigcup \Omega_2)$.

Let $\phi \in C^\infty(\mathbb{R}^n)$ satisfy

$$\phi = \begin{cases} 1, & x \in \mathbb{R}^n \backslash N_\varepsilon(\Omega_1 \cap \Omega_2), \\ 0, & x \in N_{\frac{\varepsilon}{2}}(\Omega_1 \cap \Omega_2); \end{cases}$$

then

$$v = \begin{cases} \phi u_1, & x \in \mathbb{R}^n \backslash \Omega_1, \\ \phi u_2, & x \in \mathbb{R}^n \backslash \Omega_2, \\ 0, & x \in \Omega_1 \cap \Omega_2 \end{cases}$$

is a well-defined C^∞ function and $v = u_1 = u_2$ outside a compact set, so that

$$f_3 = (\Delta + k^2 - q(x))v$$

must also have far field α. ∎

LEMMA 4.5. *For any $\epsilon > 0$ and any far field α with a compactly supported source, there exists an integer N and a sequence of sources f_n such that*

$$N_\epsilon(\mathrm{cS}_k\mathrm{supp}_q\alpha) \supset \bigcap_{n=1}^{N} \mathrm{ch\,supp}(f_n). \tag{4.2}$$

Proof. Let B denote the complement of $\mathrm{cS}_k\mathrm{supp}_q\alpha$, let B_ϵ denote the complement of $N_\epsilon(\mathrm{cS}_k\mathrm{supp}_q\alpha)$, and let A_f denote the complement of $\mathrm{ch\,supp}(f)$. The A_f's are open and B_ϵ is closed. Taking complements in the definition (4.1) tells us that

$$B = \bigcup_{F_q^+ f = \alpha} A_f.$$

We will prove the theorem by showing that we may choose f_n such that

$$B_\epsilon \subset \bigcup_{n=1}^{N} A_{f_n}.$$

Let f_1 be a compactly supported source which radiates α. Now $B_\epsilon \backslash A_{f_1}$ is compact and the A_f's provide an open cover of that compact set, so a finite subcover exists. Numbering that finite subcover A_{f_2} through A_{f_N} establishes (4.2) and proves the theorem. □

Proof of Theorem 4.3. Lemma 4.5 implies that $N_\epsilon(\mathrm{cS}_k\mathrm{supp}_q\alpha)$ is contained in the intersection of finitely many sources. We may take Ω_1 and Ω_2 in Lemma 4.4 to be the convex hulls of the supports of two of the sources, so that the hypothesis that $\mathbb{R}^n \backslash (\Omega_1 \bigcup \Omega_2)$ is connected is automatic. Thus we can produce a source supported on a neighborhood of the intersection of the convex hulls of the supports of any two sources, and we complete the proof by induction. □

5. Active sensing: Finding the support of a scatterer. The convex scattering support of a far field which was not radiated by a source, but rather scattered by an inhomogeneity in a homogeneous medium, detects the deviation of the index of refraction from that of the homogeneous background medium. That is, when we illuminate the medium with an incoming wave, the inhomogeneity becomes a secondary, or induced, source, and our test can be applied to locate that source.

Similarly, we can apply the Picard test in an inhomogeneous background if we wish to locate the deviation of the index of refraction from that known background. In both cases, we apply the test to the deviation of the measured outgoing far field from the outgoing far field that we should have measured if no deviation were present. If the background is homogeneous, the test is applied to the *scattered wave*, the outgoing far field minus the antipodal map of the incoming field, which is the outgoing far field of the free solution with the same incoming far field. In the case of an inhomogeneous background, we subtract the wave scattered by the background.

In order to be mathematically precise we need to recall the scattering operator. We may formulate the scattering problem as

$$(\Delta + k^2 - q(x))u = 0,$$
$$u_\infty^-(\Theta) = \beta(\Theta),$$

where $\beta \in L^2(S^{n-1})$ parameterizes the incoming far field (recall (2.13)). It is customary to seek the total wave u as the sum of an incident wave and a scattered wave:

$$\begin{aligned} u &= u_{inc} + u_{sc} \\ &= \mathcal{H}\beta + u_{sc}. \end{aligned}$$

In our notation, the incident wave is just the Herglotz operator (2.8) acting on β. Because $\mathcal{H}\beta$ has incoming far field equal to β, the scattered wave is outgoing and satisfies

$$\begin{aligned} (\Delta + k^2 - q(x))u_{sc} &= q\mathcal{H}\beta, \\ (u_{sc})^-_\infty &= 0. \end{aligned} \tag{5.1}$$

Thus,

$$u_{sc} = G^+_q q\mathcal{H}\beta$$

and has far field

$$(u_{sc})^+_\infty = F^+_q q\mathcal{H}\beta.$$

The scattering operator, which maps the incoming far field to the scattered far field, is given by

$$\begin{aligned} S_q &= F^+_q q\mathcal{H} \\ &= F^+_0 q(I - G^+_0 q)^{-1}\mathcal{H}. \end{aligned}$$

If we can measure the far field $S_q\beta$ for a single incoming wave β, we may apply the Picard test to $S_q\beta$ to find what must be a subset of the convex hull of the support of the *induced source*

$$\begin{aligned} f &= q(I - G^+_0 q)^{-1}\mathcal{H}\beta \\ &= qu. \end{aligned}$$

Because the unique continuation principle guarantees that u cannot vanish on an open set, we can be certain that we are truly estimating the support of q, i.e., we have the following lemma.

LEMMA 5.1.

$$\operatorname{supp} qu = \operatorname{supp} q.$$

Suppose now that we are looking to locate not the support of q, but the support of $q - q_{bg}$. That is, we want to find the places where the medium deviates from the known inhomogeneous background q_{bg}. We rewrite (5.1) as

$$\begin{aligned} (\Delta + k^2 - q_{bg})u_{sc} &= q_{bg}\mathcal{H}\beta + (q - q_{bg})u, \\ (u_{sc})^-_\infty &= 0, \end{aligned}$$

which shows that

$$u_{sc} = G^+_{q_{bg}}\mathcal{H}\beta + G^+_{q_{bg}}(q - q_{bg})u$$

with outgoing far field

$$S_q\beta = S_{q_{bg}}\beta + F^+_{q_{bg}}(q - q_{bg})u.$$

Thus, if we apply the Picard test to the far field

$$S_q\beta - S_{q_{bg}}\beta,$$

we obtain an estimate of the support of the induced source $(q-q_{bg})u$, which, according to Lemma 5.1, is a lower bound on the support of $q-q_{bg}$. We state this as a corollary of Theorem 3.6.

THEOREM 5.2. *For every incident field α,*

$$\mathrm{cS}_k\mathrm{supp}_q(S^+_{q_{bg}} - S^+_q)\alpha \subset \mathrm{ch}\,\mathrm{supp}(q - q_{bg}),$$

i.e., if $q = q_{bg}$ in $\mathbb{R}^n \setminus \Omega$,

$$(S^+_{q_{bg}} - S^+_q)\alpha \in \mathrm{Range}(F^+_{q_{bg}}|_{L^2(\Omega)})$$

for every incident field α.

6. Computing the convex scattering support. In the previous section, we have shown how to unambiguously associate a closed convex set, the convex scattering support, with a far field; we showed that any source which produces that far field must contain that set in the convex hull of its support, and that there always exists a source supported in any neighborhood of the convex scattering support which radiates that far field.

Theorem 3.6 tests whether that set is contained in a test region Ω. In this section we describe a simple algorithm to make use of Theorem 3.6 to find the convex scattering support and show a numerical result for a homogeneous background in two dimensions to illustrate how the method works. We do not intend to suggest that what we present below is careful numerical study. It is meant to be illustrative. We do, however, view it as strong evidence that this provides a stable numerical method.

ALGORITHM. We will choose as test domain, $\Omega = B_R(c)$, the ball of radius R with center c.

1. *Choose the center $c = 0$ and find the smallest R such that $B_R(0)$ contains the scattering support.* For a homogeneous background medium this is easily seen by simply plotting the modulus of the Fourier coefficients of the far field and looking for the place they become effectively zero (i.e., uniformly small). In the plot on the bottom left in Figure 1, the modulus of the Fourier coefficients are effectively zero for $|n|$ a little bigger than 50. The wavenumber in this example is $k = 50$, so the radius R of the circle about zero is one ($\frac{50}{50}$).
   This succeeds because the decomposition of $F^+_0|_{L^2(B_c(R))}$ is exactly

$$F^+_0|_{L^2(B_0(R))} = \sum_{n=-\infty}^{\infty} \mathrm{e}^{in(\theta-\frac{\pi}{2})} \otimes \chi_{B_0(R)}(r) J_n(kr) \mathrm{e}^{-in(\phi-\frac{\pi}{2})}$$

   (here $\chi_{B_0(R)}(r)$ is the characteristic function of the ball) so that its singular

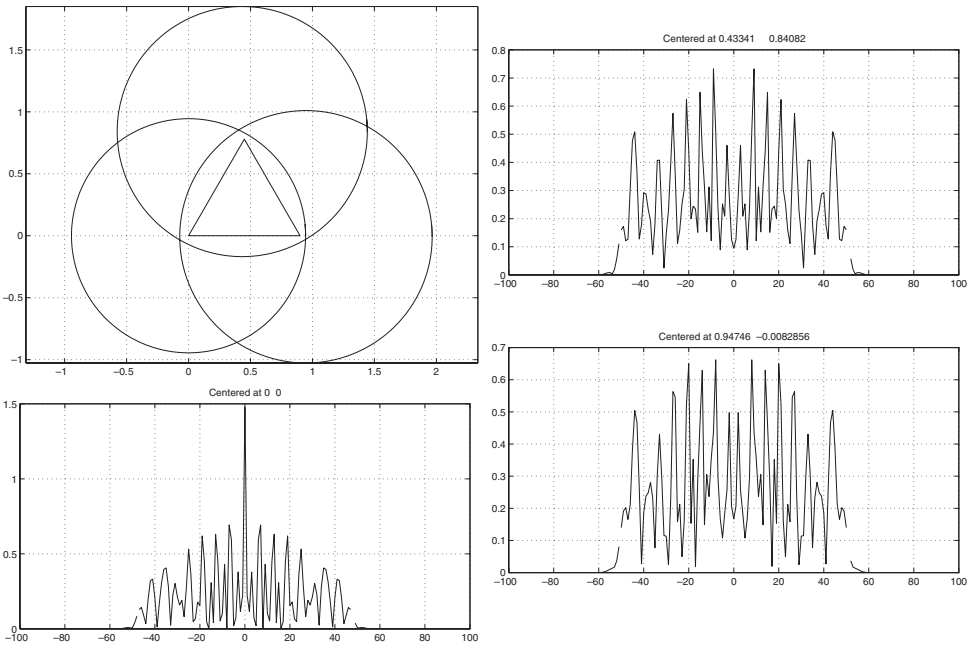

FIG. 1. *Estimating the convex scattering support of a triangular source.*

values are

$$\sigma_n = \left(\int_0^R J_n^2(ks) s ds \right)^{\frac{1}{2}} \sim \begin{cases} (R^2 - n^2)^{\frac{1}{4}} & \text{for } n < kR, \\ \dfrac{1}{\sqrt{n}} \left(\dfrac{eR}{2n} \right)^n & \text{for } n > kR, \end{cases} \tag{6.1}$$

which means that the σ_n are uniformly large for $n < kR$ and decay rapidly to zero as soon as $n > kR$.

2. *Choose another center and repeat.* In the homogeneous case, we compute the far field of the translated source instead of translating the test region. We use the formula

$$F_0^+ f(x - c) = e^{ik|c| \cos(\theta - \theta_c)} F_0^+ f \tag{6.2}$$

and then apply the test from the previous step to the new far field given by the left-hand side of (6.2). The plots on the left-hand sides of Figures 1 and 2 are the results of translating the far fields to the centers indicated in the figures.

We don't suggest a specific algorithm for choosing the centers here. In the first example, we chose a new center to be on the intersection of previous circles. In the second, we needed to choose centers far from the line source to see that it was flat. One expensive alternative is just to grid space and then choose centers at each grid point.

3. Our estimate of the convex scattering support is the intersection of these balls as in the large plots at the top left of Figures 1 and 2).

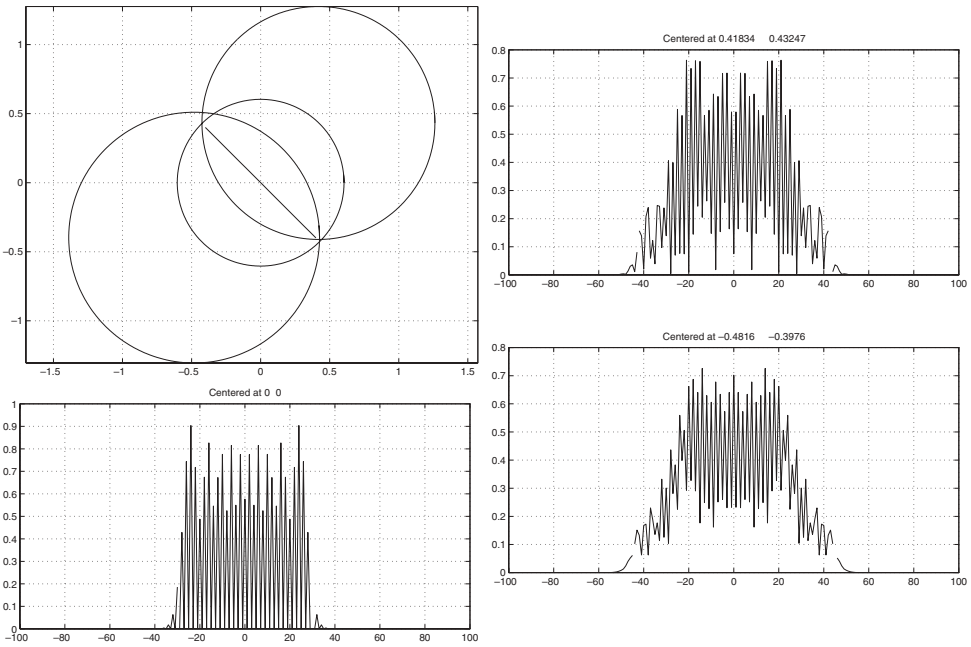

FIG. 2. *Estimating the convex scattering support of a line source.*

7. Conclusions. We have shown that the notion of the convex scattering support extends to scattering in an inhomogeneous background. Theoretically, the notion is actually quite general, relying only on unique continuation, and we expect it to hold in a very general mathematical setting. The observation of a far field can easily be replaced by the observation of a set of Cauchy data on all or part of the boundary of a region.

From a practical point of view, it is the *threshold behavior* of the σ_n's in (6.1) that is the most encouraging and intriguing. Equation (6.1) tells us not only that the Fourier coefficients of a far field produced in the ball of radius R will go to zero rapidly when n becomes greater than kR, but that they can, in general,[1] be expected to be uniformly large for n even slightly less than kR. This means that we need only look for this transition to zero, which is much less sensitive to noise than any sort of convergence or ratio test.

This threshold is intimately associated with wave propagation, and not merely a consequence of unique continuation. While the convex scattering support can easily be defined for Laplace's equation (the Helmholtz equation with $\omega = 0$), the corresponding σ_n's exhibit no such behavior.

From our point of view, one very relevant question is whether this thresholding behavior occurs for other test domains and for Helmholtz equations with inhomogeneous backgrounds. If it does, we can expect these tests to be robust in the presence of noise as well.

[1] Theorem 3.6 itself guarantees that we can always artificially choose examples where the first 1000 Fourier coefficients are zero, and only after that do the hypotheses of the theorem hold. Nevertheless, we expect that, for a broad class of sources, this won't be the case.

REFERENCES

[1] S. Agmon, *Spectral properties of Schrödinger operators and scattering theory*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 2 (1975), pp. 151–218.

[2] D. Colton, J. Coyle, and P. Monk, *Recent developments in inverse acoustic scattering theory*, SIAM Rev., 42 (2000), pp. 369–414.

[3] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, Springer-Verlag, Berlin, 1998.

[4] Y. V. Egorov and M. A. Shubin, eds., *Partial Differential Equations* I, Encyclopaedia Math. Sci. 30, Springer-Verlag, Berlin, 1991.

[5] M. Ikehata, *On reconstruction in the inverse conductivity problem with one measurement*, Inverse Problems, 16 (2000), pp. 785–793.

[6] M. Ikehata, *A regularized extraction formula in the enclosure method*, Inverse Problems, 18 (2002), pp. 435–440.

[7] A. Kirsch, *Factorization of the far-field operator for the inhomogeneous medium case and an application in inverse scattering theory*, Inverse Problems, 15 (1999), pp. 413–429.

[8] S. Kusiak and J. Sylvester, *The scattering support*, Comm. Pure Appl. Math., 56 (2003), pp. 1525–1548.

[9] D. R. Luke and R. Potthast, *The no response test—a sampling method for inverse scattering problems*, SIAM J. Appl. Math., 63 (2003), pp. 1292–1312.

[10] R. Potthast, J. Sylvester, and S. Kusiak, *A "range test" for determining scatterers with unknown physical properties*, Inverse Problems, 19 (2003), pp. 533–547.

[11] M. E. Taylor, *Partial Differential Equations* II: *Qualitative Studies of Linear Equations*, Appl. Math. Sci. 116, Springer-Verlag, New York, 1996.

SIAM J. MATH. ANAL.
Vol. 36, No. 4, pp. 1159–1176

ON PRESSURELESS GASES DRIVEN BY A STRONG INHOMOGENEOUS MAGNETIC FIELD*

ISABELLE GALLAGHER† AND LAURE SAINT-RAYMOND‡

Abstract. We are interested in the life span and the asymptotic behavior of the solutions to a system governing the motion of a pressureless gas that is submitted to a strong, inhomogeneous magnetic field $\varepsilon^{-1}B(x)$ of variable amplitude but fixed direction; this is a first step in the direction of the study of rotating Euler equations. This leads to the study of a multidimensional Burgers-type system on the velocity field u_ε, penalized by a rotating term $\varepsilon^{-1}u_\varepsilon \wedge B(x)$. We prove that the unique, smooth solution of this Burgers system exists on a uniform time interval $[0,T]$. We also prove that the phase of oscillation of u_ε is an order one perturbation of the phase obtained in the case of a pure rotation (with no nonlinear transport term), $\varepsilon^{-1}B(x)t$. Finally, going back to the pressureless gas system, we obtain the asymptotics of the density as ε goes to zero.

Key words. rotating pressureless gas, asymptotic behavior, oscillations

AMS subject classifications. Primary, 35B40; Secondary, 76U05, 76W05

DOI. 10.1137/S0036141003435540

1. Introduction.

The aim of this paper is to study the asymptotic behavior of a fluid submitted to a strong external inhomogeneous magnetic field.

The case when the field is constant has been studied by a number of authors, both for compressible and incompressible models of fluids (see, for instance, [1], [3], or [7] for incompressible fluids, and [4] or [6] for rarefied plasmas). In that case, one cannot only derive the asymptotic average motion (which is given by the weak limit of the velocity field), but one can also describe all the oscillations in the system and possibly their coupling: the filtering techniques used for that rely on explicit computations in Fourier space.

In the case when the magnetic field is inhomogeneous, those methods are not relevant anymore. Weak compactness and compensated compactness arguments nevertheless allow us to determine the average motion (see [6] in the case of a rarefied plasma governed by the Vlasov–Poisson system, and [5] in the case of a viscous incompressible fluid). In order to describe the oscillating component of the motion, one has to understand the interaction between the penalization and the nonlinear term of transport: indeed one expects that the flow modifies substantially the phase of oscillation (which is of course inhomogeneous).

We propose here to analyze this interaction for a simplified model of magnetohydrodynamics, the so-called Euler system of pressureless gas dynamics.

1.1. A simple model for magnetohydrodynamics.

We consider the following system of partial differential equations:

$$\begin{aligned} &\partial_t \rho + \nabla\cdot(\rho u) = 0, \quad x \in \mathbf{R}^3, t > 0, \\ &\partial_t(\rho u) + \nabla\cdot(\rho u \otimes u) = \rho u \wedge B, \quad x \in \mathbf{R}^3, t > 0, \\ &\rho(t=0) \equiv \rho_0, \quad u(t=0) \equiv u_0, \quad x \in \mathbf{R}^3, \end{aligned} \tag{1.1}$$

*Received by the editors October 1, 2003; accepted for publication (in revised form) May 21, 2004; published electronically January 27, 2005.

http://www.siam.org/journals/sima/36-4/43554.html

†Centre de Mathématiques Laurent Schwartz, UMR 7640, Ecole Polytechnique, 91128 Palaiseau, France (Isabelle.Gallagher@math.polytechnique.fr). Current address: Institut de Mathématiques de Jussieu, Université Paris 7, Case 7012, 2 place Jussieu, 75251 Paris Cedex 05, France.

‡Laboratoire J.-L. Lions, UMR 7598, Université Paris VI, 175, rue du Chevaleret, 75013 Paris, France (saintray@ann.jussieu.fr).

where ρ denotes the density of the fluid, u its mean velocity, and B the external magnetic field ($\nabla\cdot B = 0$). The first equation expresses the local conservation of mass, while the second one gives the local conservation of momentum provided that there is no internal force (no pressure). This assumption is relevant only in some particular regimes (corresponding to sticky particles [2]). From a physical point of view, this may seem a strong restriction, but it allows us to perform a first mathematical study of that type of inhomogeneous singular perturbation problem: indeed in this special case a major simplification arises since the equation on the mean velocity can be (at least formally) decoupled from the rest of the system:

$$\partial_t u + (u\cdot\nabla)u = u\wedge B, \quad x\in\mathbf{R}^3, t>0.$$

We then obtain a system of Burgers type that is a prototype of hyperbolic system. A work in progress should extend the present results to more realistic models, in particular to the three-dimensional incompressible Euler system.

In order to further simplify the analysis, we assume that the direction of the field B is constant,

$$B(x) \equiv \frac{1}{\varepsilon} b(x_1,x_2)e_3, \quad (x_1,x_2)\in\mathbf{R}^2 \text{ and } e_3 = {}^t(0,0,1),$$

which allows us to get rid of the geometry of the field lines (for detailed comments on this subject, see, for instance, [5, Remark 1.4]). Any solution to the system (1.1) has then uniform regularity with respect to the variable x_3. To isolate the phenomenon of inhomogeneous oscillations with instantaneous loss of regularity, we restrict therefore our attention to the two-dimensional singular perturbation problem in the plane orthogonal to the magnetic field. We finally have

$$\begin{aligned} &\partial_t\rho + \nabla\cdot(\rho u) = 0, \quad x\in\mathbf{R}^2, t>0,\\ &\partial_t(\rho u) + \nabla\cdot(\rho u\otimes u) = \frac{b}{\varepsilon}\rho u^\perp, \quad x\in\mathbf{R}^2, t>0,\\ &\rho(t=0)\equiv\rho_0, \quad u(t=0)\equiv u_0, \quad x\in\mathbf{R}^2, \end{aligned} \tag{1.2}$$

where $u^\perp$ denotes the vector field with components $(u_2,-u_1)$, and the intensity b of the magnetic field satisfies the following assumptions:

$$b\in C^\infty(\mathbf{R}^2)\cap W^{2,\infty}(\mathbf{R}^2), \tag{H0}$$

$$\inf_{x\in\mathbf{R}^2} b(x) = b_- > 0. \tag{H1}$$

A standard fixed point argument then allows us to prove the local well-posedness of (1.2). The result is the following.

THEOREM 1. *Consider a function b satisfying assumptions* (H0) *and* (H1). *Let ρ_0 be a nonnegative function and let u_0 be a vector field in $H^s(\mathbf{R}^2)$ ($s>2$). Then, for all $\varepsilon>0$, there exist $T_\varepsilon\in\,]0,+\infty]$ and a unique solution of* (1.2), *$(\rho_\varepsilon,u_\varepsilon)\in L^\infty_{loc}([0,T_\varepsilon[,H^s(\mathbf{R}^2))$.*

Note that the lifespan T_ε of the solution depends on ε, and that the lower bound on T_ε coming from the Duhamel formula goes to zero as $\varepsilon\to 0$. The first difficulty in studying the asymptotics $\varepsilon\to 0$ consists then in understanding why the magnetic penalization does not destabilize the system, and in proving that the solution $(\rho_\varepsilon,u_\varepsilon)$ exists on a uniform interval of time.

1.2. Formal analysis. Before stating more precise results on the life span of the solutions and on the asymptotics $\varepsilon \to 0$, we have chosen to give some simple observations about the problem to guide intuition. In this first approach we restrict our attention to the analysis of the equation governing the velocity.

The first step of the formal analysis consists in determining the mean behavior of the velocity field, that is, the weak limit of u_ε. We have

$$u_\varepsilon = \frac{\varepsilon}{b} \left(\partial_t u_\varepsilon + (u_\varepsilon \cdot \nabla) u_\varepsilon\right)^\perp .$$

As b is bounded from below, if we are able to establish convenient a priori bounds on u_ε, this will imply

$$u_\varepsilon \rightharpoonup 0$$

in some weak sense. This means that we expect the velocity to oscillate at high frequency (on vanishing temporal or spatial scales).

Another way to get an idea of the asymptotic behavior of the velocity is to study the simple case when b is constant. The group of oscillations generated by the magnetic penalization is then homogeneous,

$$R\left(\frac{t}{\varepsilon}\right) u = u \cos\left(\frac{bt}{\varepsilon}\right) - u^\perp \sin\left(\frac{bt}{\varepsilon}\right) ,$$

which corresponds to the rotation with frequency $2\pi b/\varepsilon$. As the coefficients are constant, this group is not perturbed by the transport. Classical filtering methods (see namely [7], [8]) can then be applied: setting

$$v_\varepsilon \stackrel{\text{def}}{=} R\left(-\frac{t}{\varepsilon}\right) u_\varepsilon$$

leads to

$$\partial_t v_\varepsilon + Q\left(\frac{t}{\varepsilon}, v_\varepsilon, v_\varepsilon\right) = 0,$$

where $Q\left(\frac{t}{\varepsilon}, ., .\right)$ is a quadratic form with bounded coefficients depending on t/ε. As there is only one oscillation frequency, there is no resonance, which implies that

$$v_\varepsilon \to u_0$$

in some strong sense, provided that convenient a priori bounds on u_ε (and consequently on v_ε) hold. This means that we can describe completely the oscillations and get a strong convergence result. Of course, we get as a corollary that the life span T_ε is uniformly bounded from below, and we even expect that $T_\varepsilon \to +\infty$ as $\varepsilon \to 0$.

The case we consider here is much more complicated. The group of oscillations generated by the magnetic penalization is again very easy to describe,

$$R\left(\frac{t}{\varepsilon}, x\right) u = u \cos\left(\frac{b(x)t}{\varepsilon}\right) - u^\perp \sin\left(\frac{b(x)t}{\varepsilon}\right) ,$$

but it is nonhomogeneous, which entails

- a loss of regularity ($R\left(\frac{t}{\varepsilon}, x\right) u$ blows up in all Sobolev norms $H^s(\mathbf{R}^2)$ for $s > 0$);
- an interaction with the transport operator (with the same definition of $v_\varepsilon = R\left(-\frac{t}{\varepsilon}, x\right) u_\varepsilon$ as previously, we do not expect $\partial_t v_\varepsilon$ to be bounded in any space of distributions).

The purpose behind this model problem is to understand how to overcome these difficulties. The first step is to explain how the phase of oscillations is modified by the flow: note that even a small correction on the phase changes strongly the vector field. Then we have to establish a strong convergence result using a new method: classical energy methods fail because of the lack of regularity on approximate solutions. Here an appropriate rewriting of the system by means of characteristics associated with the flow allows us to understand the underlying structure and to answer both questions: in particular we will see that the spaces which are well adapted for this type of study are constructed on $L^\infty(\mathbf{R}^2)$. In the case of incompressible dynamics, the analysis will be therefore much more difficult since the transport is replaced by a nonlocal pseudodifferential operator.

1.3. Main results. As long as the solution $(\rho_\varepsilon, u_\varepsilon)$ of system (1.2) is smooth, the velocity u_ε satisfies the following equation of Burgers type:

$$\begin{aligned} \partial_t u_\varepsilon + (u_\varepsilon \cdot \nabla) u_\varepsilon + \frac{b}{\varepsilon} u_\varepsilon^\perp = 0, \quad x \in \mathbf{R}^2, t > 0, \\ u_\varepsilon(t=0) = u^0, \quad x \in \mathbf{R}^2. \end{aligned} \tag{1.3}$$

Using refined a priori estimates on this last equation, we can prove that for all $\varepsilon > 0$ it admits a smooth solution on a uniform time $T > 0$. We will prove the following result.

THEOREM 2. *Consider a function b satisfying assumptions* (H0) *and* (H1). *Let ρ_0 be a nonnegative function in $W^{s-1,\infty}(\mathbf{R}^2)$, and let u_0 be a vector field in $W^{s,\infty}(\mathbf{R}^2)$ ($s \geq 1$). Then there exists $T^* \in]0, +\infty]$ such that, for all $T < T^*$ and all $\varepsilon \leq \varepsilon_T$, there is a unique $(\rho_\varepsilon, u_\varepsilon) \in L^\infty([0,T], W^{s-1,\infty}(\mathbf{R}^2) \times W^{s,\infty}(\mathbf{R}^2))$ solution of* (1.2) *(which is nevertheless not uniformly bounded in $L^\infty([0,T], W^{s-1,\infty}(\mathbf{R}^2) \times W^{s,\infty}(\mathbf{R}^2))$ for $s > 0$).*

Remark 1. The proof of Theorem 2 shows actually that the supremum lifetime T_ε^* of the solution $(\rho_\varepsilon, u_\varepsilon)$ (corresponding to the first crossing of characteristics) converges to T^{**} as ε goes to 0 with

$$T^{**} = C \|u_0\|_{L^\infty}^{-1} \|\nabla b\|_{L^\infty}^{-1}.$$

In this framework, it is relevant to consider the asymptotics $\varepsilon \to 0$ on the time interval $[0,T]$. The same type of computations as used previously allows us to prove that the velocity field behaves almost as in the constant case (with slight modifications of the phase of oscillations).

THEOREM 3. *Consider a function b satisfying assumptions* (H0) *and* (H1). *Let u_0 be a vector field in $W^{s,\infty}(\mathbf{R}^2)$ ($s \geq 1$). For all $T \leq T^*$ as in Theorem* 2 *and all $\varepsilon \leq \varepsilon_T$, denote by u_ε the solution of* (1.3) *in $L^\infty([0,T], W^{s,\infty}(\mathbf{R}^2))$. Then*

$$u_\varepsilon(t,x) - \big(u_0(x) \cos\theta_\varepsilon(t,x) - u_0^\perp(x) \sin\theta_\varepsilon(t,x)\big)$$

converges strongly to 0 in $L^\infty([0,T] \times \mathbf{R}^2)$, where the phase θ_ε is defined by the equation

$$\begin{aligned} \theta_\varepsilon(t,x) = \frac{b(x)t}{\varepsilon} - t u_0(x) \cdot \nabla \log b(x) \sin\theta_\varepsilon(t,x) \\ + \ t u_0^\perp(x) \cdot \nabla \log b(x) \cos\theta_\varepsilon(t,x). \end{aligned} \tag{1.4}$$

Rewriting the equation on the density ρ_ε with a transport term and a penalization term (coming from the divergence of u_ε which is of order $1/\varepsilon$)

$$\begin{aligned} \partial_t \rho_\varepsilon + (u_\varepsilon \cdot \nabla)\rho_\varepsilon + \rho_\varepsilon \nabla \cdot u_\varepsilon = 0, \quad x \in \mathbf{R}^2, t > 0, \\ \rho_\varepsilon(t=0) = \rho^0, \quad x \in \mathbf{R}^2, \end{aligned} \tag{1.5}$$

we can then determine the global asymptotics of the Euler system of pressureless gases (1.2).

THEOREM 4. *Consider a function b satisfying assumptions* (H0) *and* (H1). *Let* ρ_0 *be a nonnegative function in* $W^{s-1,\infty}(\mathbf{R}^2)$, *and let* u_0 *be a vector field in* $W^{s,\infty}(\mathbf{R}^2)$ *(*$s \geq 1$*). For all* $T \leq T^*$ *as in Theorem* 2 *and all* $\varepsilon \leq \varepsilon_T$, *denote by* $(\rho_\varepsilon, u_\varepsilon)$ *the solution of* (1.2) *in* $L^\infty([0,T], W^{s-1,\infty}(\mathbf{R}^2) \times W^{s,\infty}(\mathbf{R}^2))$. *Then*

$$\rho_\varepsilon(t,x) - \rho_0(x)\left(1 + tu_0 \cdot \nabla \log(x) \cos\theta_\varepsilon(t,x) - tu_0^\perp \cdot \nabla \log b(x) \sin\theta_\varepsilon(t,x)\right)$$

converges strongly to 0 *in* $L^\infty([0,T] \times \mathbf{R}^2)$, *where the phase* θ_ε *is defined as in equation* (1.4).

Remark 2. It should be noted that assumption (H1) is not merely a technical artifact which could be removed with some work. In fact all the uniform estimates in this paper blow up if the field B is allowed to have a level curve where $B = 0$—which could occur in many physical cases—and new mathematical difficulties arise in that case.

Let us comment a little on the proof of those theorems and give the structure of the paper.

It is quite clear that energy methods will not enable us to have a good control on the asymptotics of $(\rho_\varepsilon, u_\varepsilon)$, since as soon as we want a control on derivatives of u_ε, unbounded terms will appear. So the most appropriate way to study system (1.2) is to rewrite it using the characteristics of the flow and to study those characteristics precisely.

Section 2 is therefore devoted to rewriting system (1.2) in characteristic form, and in the derivation of a few a priori estimates.

In order to establish the existence of a solution $(\rho_\varepsilon, u_\varepsilon)$ to system (1.2) on a uniform time interval $[0,T]$, it is enough to see that the solution is well-defined (and smooth) as long as the flow generates a diffeomorphism $X_\varepsilon(t,.)$,

$$\frac{dX_\varepsilon}{dt}(t,x) = u_\varepsilon(t, X_\varepsilon(t,x)),$$

and hence to prove that the characteristics cannot cross before time T. The precise estimates on DX_ε leading to Theorem 2 are performed in section 3; they use in a crucial way some results of nonstationary phase type.

The asymptotic behavior of $u_\varepsilon(t, X_\varepsilon(t,.))$ and $\rho_\varepsilon(t, X_\varepsilon(t,.))$ is then simply obtained from the explicit approximation of the characteristics X_ε, using Taylor expansions for the various fields. In order to establish the convergence results stated in Theorems 3 and 4, the main difficulty lies therefore in getting a precise description of the inverse characteristics $X_\varepsilon^{-1}(t,.)$, which is done in section 4.

2. Appropriate formulation of the system. As pointed out in the introduction, energy estimates do not seem to be the right angle of attack for our problem. We shall therefore in this short section present a new formulation of system (1.2), by means of characteristics (subsection 2.1). In that way some a priori estimates can be deduced immediately (see subsection 2.2).

To simplify notation, from now on we shall drop the index ε in u_ε and simply write u (and similarly for any other ε-dependent function).

2.1. Trajectories associated with the flow. Let us write system (1.2) in the following form:

$$\begin{aligned} &\frac{dX}{dt} = u(t,X), \quad X_{|t=0} = x, \\ &\frac{d}{dt}(\rho(t,X)) + \rho\nabla\cdot u(t,X) = 0, \quad \rho_{|t=0} = \rho_0, \\ &\frac{d}{dt}(u(t,X)) + \frac{b(X)}{\varepsilon}u^\perp(t,X) = 0, \quad u_{|t=0} = u_0. \end{aligned} \tag{2.1}$$

As seen in Theorem 1, there is a solution to system (1.2) for a time depending on ε, and as long as the trajectories do not intersect we can write in particular

$$u(t,X(t,x)) = u_0(x)\cos\left(\frac{\phi(t,x)}{\varepsilon}\right) - u_0^\perp(x)\sin\left(\frac{\phi(t,x)}{\varepsilon}\right), \tag{2.2}$$

where we have defined the functions

$$\phi(t,x) = \int_0^t \beta(s,x)\,ds, \quad \beta(t,x) \stackrel{\text{def}}{=} b(X(t,x))$$

(these functions are well-defined as long as the characteristics do not cross each other).

If u is smooth enough, then ρ is uniquely defined by the transport equation it satisfies. So from now on we can concentrate on u (and X). As one of the aims of this article is to prove Theorem 2 (which will be achieved in the next section), we shall from now on call T^ε the largest time before which no characteristics intersect; one of our goals is to prove that T^ε is uniformly bounded from below as ε goes to zero.

In the next subsection we are going to derive from (2.1) and (2.2) some easy a priori estimates for times $0 \le t \le T^\varepsilon$, which will help us prove Theorem 2 in section 3, and Theorems 3 and 4 in section 4.

2.2. A priori estimates. Formula (2.2) immediately enables us to deduce the a priori estimate

$$\|u\|_{L^\infty([0,T^\varepsilon[\times\mathbf{R}^2)} \le 2\|u_0\|_{L^\infty}\,, \tag{2.3}$$

which implies that

$$\left\|\frac{dX}{dt}\right\|_{L^\infty([0,T^\varepsilon[\times\mathbf{R}^2)} \le 2\|u_0\|_{L^\infty}. \tag{2.4}$$

In particular $X - x$ remains bounded in space for all times $0 \le t < T^\varepsilon$, and we have

$$\forall t \in [0,T^\varepsilon[, \quad \|X(t,\cdot) - x\|_{L^\infty(\mathbf{R}^2)} \le 2t\|u_0\|_{L^\infty(\mathbf{R}^2)}. \tag{2.5}$$

Since $\beta(t,x) = b(X(t,x))$, we have

$$\begin{aligned} \|\partial_t\beta\|_{L^\infty([0,T^\varepsilon[\times\mathbf{R}^2)} &\le \|\nabla b\|_{L^\infty(\mathbf{R}^2)}\left\|\frac{dX}{dt}\right\|_{L^\infty([0,T^\varepsilon[\times\mathbf{R}^2)} \\ &\le 2\|\nabla b\|_{L^\infty(\mathbf{R}^2)}\|u_0\|_{L^\infty}, \end{aligned} \tag{2.6}$$

as well as

$$\forall t \in [0,T^\varepsilon[, \quad \forall x \in \mathbf{R}^2, \quad b_- \le \beta(t,x) \le \|b\|_{L^\infty(\mathbf{R}^2)} \tag{2.7}$$

with b_- defined in (H1).

Now we are going to look for an approximation of X: integrating formula (2.2) in time yields

$$X(t,x) = x + u_0(x)\int_0^t \cos\left(\frac{\phi(s,x)}{\varepsilon}\right)\,ds - u_0^\perp(x)\int_0^t \sin\left(\frac{\phi(s,x)}{\varepsilon}\right)\,ds, \tag{2.8}$$

recalling that $\phi(t,x) = \int_0^t b(X(s,x))\,ds$. The following section will be devoted to a precise study of the trajectories X, which will enable us to infer Theorem 2.

3. Study of the trajectories. Formulation (2.1) of system (1.3) shows that the study of the Euler system of pressureless gases with magnetic penalization comes down to a precise analysis of the characteristics, and in particular of their invertibility.

In this section we will establish that the trajectories defined by (2.8) are invertible on a time interval $[0, T^\varepsilon[$ with

$$\lim_{\varepsilon\to 0} T^\varepsilon = T^* > 0,$$

where T^* depends on the magnetic field b and on the initial velocity field u_0. This result is based on an asymptotic expansion of the Jacobian

$$J(t,x) \overset{\text{def}}{=} |\det(DX(t,x))|,$$

which implies that

$$\forall t \in [0,T^*[, \quad \liminf_{\varepsilon\to 0} J(t,x) > 0.$$

The asymptotic expansions of X and DX (subsections 3.2 and 3.4) are obtained using some results of nonstationary phase type and the L^∞-bounds established in subsections 3.1 and 3.3.

3.1. Bounds on $X(t,\cdot)$. The first step of the analysis consists in showing that for any point $x \in \mathbf{R}^2$, the characteristic stemming from x stays in a ball of size $O(\varepsilon)$ around x. This shows that the rotation has a drastic influence over the transport by u.

We have the following proposition.

PROPOSITION 1. *Let $x \in \mathbf{R}^2$ be given, and let $X(\cdot,x)$ be the trajectory starting from x at time 0, defined by (2.8). As long as it is defined, it satisfies*

$$\forall t < \min(T, T^\varepsilon), \quad |X(t,x) - x| \le 4\frac{\varepsilon}{b_-}\|u_0\|_{L^\infty}\left(1 + T\frac{\|\nabla b\|_{L^\infty}\|u_0\|_{L^\infty}}{b_-}\right).$$

Proof of Proposition 1. The proof is an immediate application of the nonstationary phase theorem. As we will be using such arguments many times in the following, let us state and prove the following lemma, which will be invoked systematically in the following sections.

LEMMA 1. *Let T be a given real number, possibly depending on ε. Let F be a function uniformly bounded in $W^{1,\infty}([0,T], L^\infty(\mathbf{R}^2))$, and let β be a positive function, also uniformly bounded in $W^{1,\infty}([0,T], L^\infty(\mathbf{R}^2))$, and bounded from below by b_-. Then for all $t \in [0,T]$ and all $x \in \mathbf{R}^2$, the following bounds hold:*

$$\left|\int_0^t F(s,x)\cos\left(\int_0^s \frac{\beta(s',x)}{\varepsilon}\,ds'\right)\,ds\right| \le \varepsilon\left(\frac{\|F(t,\cdot)\|_{L^\infty(\mathbf{R}^2)}}{b_-} + t\left\|\partial_s\frac{F(s,\cdot)}{\beta(s,\cdot)}\right\|_{L^\infty([0,t]\times\mathbf{R}^2)}\right)$$

and

$$\left|\int_0^t F(s,x)\sin\left(\int_0^s \frac{\beta(s',x)}{\varepsilon}\,ds'\right)\,ds\right| \leq \varepsilon\left(\frac{\|F(t,\cdot)\|_{L^\infty(\mathbf{R}^2)}+\|F(0,\cdot)\|_{L^\infty(\mathbf{R}^2)}}{b_-}+t\left\|\partial_s\frac{F(s,\cdot)}{\beta(s,\cdot)}\right\|_{L^\infty([0,t]\times\mathbf{R}^2)}\right).$$

Proof of Lemma 1. The proof is a simple application of the nonstationary phase theorem: an integration by parts leads to

$$\int_0^t F(s,x)\cos\left(\int_0^s \frac{\beta(s',x)}{\varepsilon}\,ds'\right)\,ds = \varepsilon\frac{F(t,x)}{\beta(t,x)}\sin\left(\int_0^t\frac{\beta(s,x)}{\varepsilon}\,ds\right) - \varepsilon\int_0^t\partial_s\left(\frac{F(s,x)}{\beta(s,x)}\right)\sin\left(\int_0^s\frac{\beta(s',x)}{\varepsilon}\,ds'\right)\,ds,$$

and similarly

$$\int_0^t F(s,x)\sin\left(\int_0^s \frac{\beta(s',x)}{\varepsilon}\,ds'\right)\,ds = \varepsilon\frac{F(0,x)}{\beta(0,x)}-\varepsilon\frac{F(t,x)}{\beta(t,x)}\cos\left(\int_0^t\frac{\beta(s,x)}{\varepsilon}\,ds\right) + \varepsilon\int_0^t\partial_s\left(\frac{F(s,x)}{\beta(s,x)}\right)\cos\left(\int_0^s\frac{\beta(s',x)}{\varepsilon}\,ds'\right)\,ds.$$

The result follows immediately.

Now let us go back to the proof of Proposition 1. Recalling formula (2.8), we simply apply Lemma 1 to the case $F(t,x)=u_0(x)$ to get

$$|X(t,x)-x|\leq 4\varepsilon\frac{\|u_0\|_{L^\infty}}{b_-}+2\varepsilon t\left\|u_0\frac{\partial_s\beta(s,\cdot)}{\beta^2(s,\cdot)}\right\|_{L^\infty([0,t]\times\mathbf{R}^2)}.$$

Estimates (2.6) and (2.7) immediately yield Proposition 1.

3.2. Asymptotics of $X(t,\cdot)$. The same type of computations based on the nonstationary phase theorem allows us actually to obtain an explicit approximation of the characteristic X at any order with respect to ε (in fact we will stop at order 2, but the argument can be pushed as far as wanted if necessary).

LEMMA 2. *For any point $x\in\mathbf{R}^2$ and any time $t\leq\min(T,T^\varepsilon)$, the following approximation of the trajectories defined in* (2.8) *holds:*

$$\left|X(t,x)-x-\varepsilon\frac{u_0(x)}{b(x)}\sin\left(\frac{\phi(t,x)}{\varepsilon}\right)+\varepsilon\frac{u_0^\perp(x)}{b(x)}\left(1-\cos\left(\frac{\phi(t,x)}{\varepsilon}\right)\right)+\varepsilon t v(x)\right|\leq C_T\varepsilon^2,$$

where the drift velocity is given by

$$v(x)=\frac{1}{2b^2(x)}\left((u_0^\perp\cdot\nabla b)u_0(x)-(u_0\cdot\nabla b)u_0^\perp(x)\right),$$

and C_T denotes a constant depending only on T, u_0, and b.

Proof of Lemma 2. Let us write the following expression for $X(t,x)$, obtained from (2.8): we have

$$X(t,x)=x+R^\varepsilon(t,x),$$

with

$$R^\varepsilon(t,x) \stackrel{\text{def}}{=} u_0(x)\int_0^t \cos\left(\frac{\phi(s,x)}{\varepsilon}\right)\,ds - u_0^\perp(x)\int_0^t \sin\left(\frac{\phi(s,x)}{\varepsilon}\right)\,ds \tag{3.1}$$
$$= R_1^\varepsilon(t,x) + R_2^\varepsilon(t,x).$$

We shall compute only the approximation for $R_1^\varepsilon(t,x)$, and we leave $R_2^\varepsilon(t,x)$ to the reader. By an integration by parts we have

$$R_1^\varepsilon(t,x) = \varepsilon\frac{u_0(x)}{\beta(t,x)}\sin\left(\frac{\phi(t,x)}{\varepsilon}\right) - \varepsilon u_0(x)\int_0^t \partial_s\left(\frac{1}{\beta(s,x)}\right)\sin\left(\frac{\phi(s,x)}{\varepsilon}\right)\,ds. \tag{3.2}$$

The first term is easy to approximate: we have, due to Proposition 1,

$$\begin{aligned}\left|\frac{1}{\beta(t,x)} - \frac{1}{b(x)}\right| &\le \frac{\|\nabla b\|_{L^\infty}}{b_-^2}|X(t,x)-x| \\ &\le \frac{4\|\nabla b\|_{L^\infty}\|u_0\|_{L^\infty}}{b_-^3}\left(1+T\frac{\|\nabla b\|_{L^\infty}\|u_0\|_{L^\infty}}{b_-}\right)\end{aligned} \tag{3.3}$$

for all $t \le \min(T, T^\varepsilon)$ and all $x \in \mathbf{R}^2$. So $\beta(t,x)$ can be replaced by $b(x)$ in the first term of R_1^ε in (3.2), up to a remainder $\varepsilon\mathcal{R}^\varepsilon$ with $\|\mathcal{R}^\varepsilon\|_{L^\infty([0,T]\times\mathbf{R}^2)} \le C_T$.

Now we need to approximate the second term. Using the fact that

$$\partial_s\beta(s,x) = (u\cdot\nabla b)(s,X(s,x))$$

with

$$u(s,X(s,x)) = u_0(x)\cos\left(\frac{\phi(s,x)}{\varepsilon}\right) - u_0^\perp(x)\sin\left(\frac{\phi(s,x)}{\varepsilon}\right),$$

we can therefore write

$$-\int_0^t \partial_s\left(\frac{1}{\beta(s,x)}\right)\sin\left(\frac{\phi(s,x)}{\varepsilon}\right)\,ds = \int_0^t u_0(x)\cdot\frac{\nabla b(X(s,x))}{2b^2(X(s,x))}\sin\left(\frac{2\phi(s,x)}{\varepsilon}\right)\,ds$$

$$-\int_0^t u_0^\perp(x)\cdot\frac{\nabla b(X(s,x))}{2b^2(X(s,x))}\left(1-\cos\left(\frac{2\phi(s,x)}{\varepsilon}\right)\right)\,ds. \tag{3.4}$$

Note that similar computations lead to the following formula, which is useful to estimate R_2^ε:

$$\int_0^t \partial_s\left(\frac{1}{\beta(s,x)}\right)\cos\left(\frac{\phi(s,x)}{\varepsilon}\right)\,ds = \int_0^t u_0^\perp(x)\cdot\frac{\nabla b(X(s,x))}{2b^2(X(s,x))}\sin\left(\frac{2\phi(s,x)}{\varepsilon}\right)\,ds$$

$$-\int_0^t u_0(x)\cdot\frac{\nabla b(X(s,x))}{2b^2(X(s,x))}\left(1+\cos\left(\frac{2\phi(s,x)}{\varepsilon}\right)\right)\,ds. \tag{3.5}$$

Both formulas (3.4) and (3.5) show that new harmonics have been created by the coupling in the equation.

Let us go back to the estimate of the right-hand side in (3.4). To estimate the oscillating terms, we use Lemma 1 with

$$F_1(s,x) = u_0(x)\cdot\frac{\nabla b(X(s,x))}{2b^2(X(s,x))} \quad\text{and}\quad F_2(s,x) = u_0^\perp(x)\cdot\frac{\nabla b(X(s,x))}{2b^2(X(s,x))}.$$

We get

$$\Big|\int_0^t F_1(s,x)\sin\left(\frac{2\phi(s,x)}{\varepsilon}\right)\,ds\Big| \leq 2\varepsilon\|u_0\|_{L^\infty}\frac{\|\nabla b\|_{L^\infty}}{b_-^3} + \varepsilon t\|u_0\|_{L^\infty}\left\|\partial_s\frac{\nabla b(X(s,\cdot))}{b^2(X(s,\cdot))}\right\|_{L^\infty([0,t]\times\mathbf{R}^2)},$$

and similarly

$$\Big|\int_0^t F_2(s,x)\cos\left(\frac{2\phi(s,x)}{\varepsilon}\right)\,ds\Big| \leq \varepsilon\|u_0\|_{L^\infty}\frac{\|\nabla b\|_{L^\infty}}{b_-^3} + \varepsilon t\|u_0\|_{L^\infty}\left\|\partial_s\frac{\nabla b(X(s,\cdot))}{b^2(X(s,\cdot))}\right\|_{L^\infty([0,t]\times\mathbf{R}^2)}.$$

By (2.4) and (2.6) we have

$$\left\|\partial_s\frac{\nabla b(X(s,\cdot))}{b^2(X(s,\cdot))}\right\|_{L^\infty([0,t]\times\mathbf{R}^2)} \leq 2\|D^2b\|_{L^\infty}\frac{\|u_0\|_{L^\infty}}{b_-^2} + 4\|\nabla b\|_{L^\infty}^2\frac{\|u_0\|_{L^\infty}}{b_-^3}.$$

Plugging that estimate along with (3.3) into the definition of R_1^ε in (3.2), we finally get

$$R_1^\varepsilon(t,x) = \varepsilon\frac{u_0(x)}{b(x)}\sin\left(\int_0^t\frac{\beta(s,x)}{\varepsilon}\,ds\right) + \varepsilon^2\mathcal{R}^\varepsilon(t,x) - \varepsilon u_0(x)\int_0^t\frac{u_0^\perp(x)\cdot\nabla b(X(s,x))}{2b^2(X(s,x))}ds.$$

Since the trajectories lie in balls of size ε,

$$|\nabla b(X(\tau,x)) - \nabla b(x)| \leq C_T\|D^2b\|_{L^\infty}\varepsilon \tag{3.6}$$

for all $x\in\mathbf{R}^2$ and all $\tau\leq t\leq\min(T^\varepsilon,T)$. So we can approximate $\frac{\nabla b(X(s,x))}{b^2(X(s,x))}$ by $\frac{\nabla b(x)}{b(x)}$ up to a remainder $\varepsilon\mathcal{R}^\varepsilon$, and we have

$$-\varepsilon u_0(x)\int_0^t\frac{u_0^\perp(x)\cdot\nabla b(X(s,x))}{2b^2(X(s,x))}\,ds = -\varepsilon t\frac{u_0^\perp(x)\cdot\nabla b(x)}{2b^2(x)}u_0(x) + \varepsilon^2\mathcal{R}^\varepsilon.$$

The estimate of $R_2^\varepsilon(t,x)$ is similar and left to the reader. This ends the proof of Lemma 2.

3.3. A priori estimates on $DX(t,\cdot)$. A necessary and sufficient condition for $X(t,\cdot)$ to be invertible is that

$$J(t,x)\stackrel{\text{def}}{=}|\det(DX(t,x))|$$

does not cancel. In order to obtain a lower bound on the time T^ε (before which the characteristics do no cross each other), we therefore need to study the behavior of the derivatives $DX(t,\cdot)$. First of all we derive a uniform L^∞-bound which will allow us to neglect some terms in the asymptotic expansion.

LEMMA 3. *Let $x\in\mathbf{R}^2$ be given, and let $X(\cdot,x)$ be the trajectory starting from x at time 0, defined by (2.8). As long as it is defined, it satisfies*

$$\forall t<\min(T,T^\varepsilon),\quad \forall x\in\mathbf{R}^2,\quad \|DX(t,x)\|\leq C_T,$$

where C_T denotes a constant depending only on b, u_0, and T.

Proof of Lemma 3. Differentiating (2.8) leads to

$$DX(t,x) - Id = Du_0(x)\int_0^t \cos\left(\frac{\phi(s,x)}{\varepsilon}\right)\,ds - Du_0^\perp(x)\int_0^t \sin\left(\frac{\phi(s,x)}{\varepsilon}\right)\,ds$$
$$-\frac{1}{\varepsilon}u_0(x)\int_0^t D\phi(s,x)\sin\left(\frac{\phi(s,x)}{\varepsilon}\right)\,ds - \frac{1}{\varepsilon}u_0^\perp(x)\int_0^t D\phi(s,x)\cos\left(\frac{\phi(s,x)}{\varepsilon}\right)\,ds,$$

with

$$D\phi(s,x) \overset{\text{def}}{=} \int_0^s DX(\tau,x)\cdot\nabla b(X(\tau,x))d\tau.$$

Applying the Fubini theorem to both last terms, we can set this identity in a suitable form to get a Gronwall estimate

$$\begin{aligned} DX(t,x) - Id = {} & Du_0(x)\int_0^t \cos\left(\frac{\phi(s,x)}{\varepsilon}\right)\,ds - Du_0^\perp(x)\int_0^t \sin\left(\frac{\phi(s,x)}{\varepsilon}\right)\,ds \\ & - \frac{1}{\varepsilon}u_0(x)\int_0^t (DX(\tau,x)\cdot\nabla)b(X(\tau,x))\int_\tau^t \sin\left(\frac{\phi(s,x)}{\varepsilon}\right)\,ds\,d\tau \\ & - \frac{1}{\varepsilon}u_0^\perp(x)\int_0^t (DX(\tau,x)\cdot\nabla)b(X(\tau,x))\int_\tau^t \cos\left(\frac{\phi(s,x)}{\varepsilon}\right)\,ds\,d\tau. \end{aligned} \tag{3.7}$$

From formula (2.8) we deduce that

$$u_0(x)\int_\tau^t \sin\left(\frac{\phi(s,x)}{\varepsilon}\right)\,ds + u_0^\perp(x)\int_\tau^t \cos\left(\frac{\phi(s,x)}{\varepsilon}\right)\,ds = (X(t,x)-X(\tau,x))^\perp.$$

Plugging this identity back into (3.7) leads to

$$\begin{aligned} DX(t,x) - Id = {} & Du_0(x)\int_0^t \cos\left(\frac{\phi(s,x)}{\varepsilon}\right)\,ds - Du_0^\perp(x)\int_0^t \sin\left(\frac{\phi(s,x)}{\varepsilon}\right)\,ds \\ & - \frac{1}{\varepsilon}\int_0^t (X(t,x)-X(\tau,x))^\perp \otimes (DX(\tau,x)\cdot\nabla)b(X(\tau,x))\,d\tau. \end{aligned} \tag{3.8}$$

As in the proof of Proposition 1, Lemma 1 yields the following estimate: for all $t \le \min(T, T^\varepsilon)$,

$$\left|Du_0(x)\int_0^t \cos\left(\frac{\phi(s,x)}{\varepsilon}\right)\,ds - Du_0^\perp(x)\int_0^t \sin\left(\frac{\phi(s,x)}{\varepsilon}\right)\,ds\right| \le C_T\varepsilon\|\nabla u_0\|_{L^\infty}, \tag{3.9}$$

with

$$C_T = \frac{4}{b_-}\left(1 + T\frac{\|\nabla b\|_{L^\infty}\|u_0\|_{L^\infty}}{b_-}\right).$$

From (3.8) we then deduce an inequality of Gronwall type:

$$\begin{aligned} \|DX(t,\cdot)\|_{L^\infty} \le {} & 1 + C_T\varepsilon\|\nabla u_0\|_{L^\infty} \\ & + \int_0^t \|DX(\tau,\cdot)\|_{L^\infty}\|\nabla b\|_{L^\infty}\left\|\frac{1}{\varepsilon}(X(\tau,x)-X(t,x))\right\|_{L^\infty} d\tau. \end{aligned} \tag{3.10}$$

By Proposition 1,

$$\forall t \leq \min(T, T^\varepsilon), \quad \left\| \frac{1}{\varepsilon}(X(t,\cdot) - X(\tau,\cdot)) \right\|_{L^\infty} \leq 2C_T \|u_0\|_{L^\infty},$$

and hence

$$\|DX(t,\cdot)\|_{L^\infty} \leq (1 + C_T\varepsilon \|\nabla u_0\|_{L^\infty}) \exp\left(2C_T \|\nabla b\|_{L^\infty} \|u_0\|_{L^\infty} t\right),$$

which is the expected estimate, proving Lemma 3.

3.4. Asymptotics of DX. In view of the results established in Lemma 2, we expect actually the derivatives $\partial_i X(t,x)$ to behave asymptotically as

$$\lambda(t,x) + \mu(t,x)\cos\left(\frac{\phi(t,x)}{\varepsilon}\right) + \nu(t,x)\sin\left(\frac{\phi(t,x)}{\varepsilon}\right),$$

where λ, μ, and ν denote some functions which do not depend on ε. Such asymptotics can be justified using the same techniques as in the previous subsection: let us prove the following lemma.

LEMMA 4. *Let $x \in \mathbf{R}^2$ be given, and let $X(\cdot,x)$ be the trajectory starting from x at time 0, defined by* (2.8). *Then, for all $t \leq \min(T^\varepsilon, T)$ and for all $x \in \mathbf{R}^2$,*

$$\left\| DX(t,x) - Id - tu_0 \otimes \nabla \log b \cos\left(\frac{\phi(t,x)}{\varepsilon}\right) + tu_0^\perp \otimes \nabla \log b \sin\left(\frac{\phi(t,x)}{\varepsilon}\right) \right\| \leq C_T\varepsilon,$$

where C_T denotes a constant depending only on b, u_0, and T.

Proof of Lemma 4. Denote by g the function defined on $[0,T] \times \mathbf{R}^2$ by

$$g(t,x) \stackrel{\text{def}}{=} DX(t,x) - Id - tu_0 \otimes \nabla \log b \cos\left(\frac{\phi(t,x)}{\varepsilon}\right) + tu_0^\perp \otimes \nabla \log b \sin\left(\frac{\phi(t,x)}{\varepsilon}\right).$$

In view of (3.10), we expect g to satisfy a Gronwall inequality of the type

$$\|g(t,x)\| \leq \int_0^t \|g(\tau,x)\| \left\| \frac{1}{\varepsilon}(X(t,\cdot) - X(\tau,\cdot)) \right\|_{L^\infty} \|\nabla b\|_{L^\infty} d\tau + C_T\varepsilon \tag{3.11}$$

for all $t \leq \min(T^\varepsilon, T)$, where C_T denotes a constant depending only on T, u_0, and b.

Let us postpone the proof of this inequality for a while and show how it enables us to infer Lemma 4. It is easy to see that

$$g(0,x) = 0.$$

Applying the Gronwall lemma and using Proposition 1 as in (3.10) leads to

$$\forall t \leq \min(T^\varepsilon, T), \quad \forall x \in \mathbf{R}^2, \quad \|g(t,x)\| \leq C_T\varepsilon.$$

Now let us go back to the proof of (3.11). We first compute

$$A(t,x) \stackrel{\text{def}}{=} \int_0^t \left(\frac{1}{\varepsilon}(X(t,x) - X(\tau,x))\right)^\perp \otimes (g(\tau,x) \cdot \nabla b(X(\tau,x)))\, d\tau. \tag{3.12}$$

By Lemma 2,

$$\begin{aligned}\frac{1}{\varepsilon}(X(t,x)-X(\tau,x))^{\perp} = {}& \frac{u_0^{\perp}}{b}(x)\left(\sin\left(\frac{\phi(t,x)}{\varepsilon}\right)-\sin\left(\frac{\phi(\tau,x)}{\varepsilon}\right)\right)\\ & -\frac{u_0}{b}(x)\left(\cos\left(\frac{\phi(t,x)}{\varepsilon}\right)-\cos\left(\frac{\phi(\tau,x)}{\varepsilon}\right)\right)\\ & -\frac{t-\tau}{2b^2(x)}\left((u_0^{\perp}\cdot\nabla b)u_0^{\perp}(x)+(u_0\cdot\nabla b)u_0(x)\right)+\varepsilon\mathcal{R}^{\varepsilon}(t,\tau,x),\end{aligned}$$

where $\mathcal{R}^{\varepsilon}$ is uniformly bounded in $L^{\infty}([0,T]^2\times\mathbf{R}^2)$. Plugging this formula back into the integral (3.12) leads to

$$\int_0^t\left(\frac{1}{\varepsilon}(X(t,x)-X(\tau,x))\right)^{\perp}\otimes(g(\tau,x)\cdot\nabla b(X(\tau,x)))\,d\tau = A_1(t,x)-A_2(t,x), \tag{3.13}$$

with

$$A_1(t,x)\overset{\text{def}}{=}\int_0^t\left(\frac{1}{\varepsilon}(X(t,x)-X(\tau,x))\right)^{\perp}\otimes(DX(\tau,x)\cdot\nabla b)(X(\tau,x))d\tau$$

and

$$\begin{aligned}A_2(t,x)\overset{\text{def}}{=}\int_0^t & \left[\frac{u_0^{\perp}}{b}(x)\left(\sin\left(\frac{\phi(t,x)}{\varepsilon}\right)-\sin\left(\frac{\phi(\tau,x)}{\varepsilon}\right)\right)\right.\\ & -\frac{u_0}{b}(x)\left(\cos\left(\frac{\phi(t,x)}{\varepsilon}\right)-\cos\left(\frac{\phi(\tau,x)}{\varepsilon}\right)\right)\\ & \left.-\frac{1}{2b^2(x)}(t-\tau)\left((u_0^{\perp}\cdot\nabla b)u_0^{\perp}(x)+(u_0\cdot\nabla b)u_0(x)\right)+\varepsilon\mathcal{R}^{\varepsilon}(t,\tau,x)\right]\\ & \otimes\left[\nabla b(X(\tau,x))+(u_0(x)\cdot\nabla b(X(\tau,x)))\nabla\log b(x)\tau\cos\left(\frac{\phi(\tau,x)}{\varepsilon}\right)\right.\\ & \left.-(u_0^{\perp}(x)\cdot\nabla b(X(\tau,x)))\nabla\log b(x)\tau\sin\left(\frac{\phi(\tau,x)}{\varepsilon}\right)\right]d\tau.\end{aligned}$$

From (3.8) and (3.9) we deduce that

$$A_1(t,x)=Id-DX(t,x)$$

up to terms of order ε.

In order to estimate the second term $A_2(t,x)$, we again use a nonstationary phase theorem. We can use (3.6) again, and as $\partial_s\beta$ is uniformly bounded according to (2.6), Lemma 1 shows that

$$\begin{aligned}A_2(t,x)= {}& \left[t\frac{u_0^{\perp}}{b}(x)\sin\left(\frac{\phi(t,x)}{\varepsilon}\right)-t\frac{u_0}{b}(x)\cos\left(\frac{\phi(t,x)}{\varepsilon}\right)-\frac{t^2}{2}v^{\perp}(x)\right]\otimes\nabla b(x)\\ & +\int_0^t\left(\frac{u_0^{\perp}}{b}(x)\sin\left(\frac{\phi(\tau,x)}{\varepsilon}\right)\right)\\ & \qquad\otimes\left((\tau u_0^{\perp}(x)\cdot\nabla b(x))\nabla\log b(x)\sin\left(\frac{\phi(\tau,x)}{\varepsilon}\right)\right)d\tau\\ & +\int_0^t\left(\frac{u_0}{b}(x)\cos\left(\frac{\phi(\tau,x)}{\varepsilon}\right)\right)\\ & \qquad\otimes\left((\tau u_0(x)\cdot\nabla b(x))\nabla\log b(x)\cos\left(\frac{\phi(\tau,x)}{\varepsilon}\right)\right)d\tau+\varepsilon\mathcal{R}^{\varepsilon}(t,x).\end{aligned}$$

Using the identities

$$\cos^2\phi = \frac{1}{2}(1+\cos(2\phi)), \quad \sin^2\phi = \frac{1}{2}(1-\cos(2\phi)),$$

we then obtain that

$$A_2(t,x) = \left(t\frac{u_0^\perp}{b}(x)\sin\left(\frac{\phi(t,x)}{\varepsilon}\right) - t\frac{u_0}{b}(x)\cos\left(\frac{\phi(t,x)}{\varepsilon}\right)\right) \otimes \nabla b(x)$$

up to terms of order ε.

Then (3.13) can be rewritten

$$\int_0^t \left(\frac{1}{\varepsilon}(X(t,x)-X(\tau,x))\right)^\perp \otimes (g(\tau,x)\cdot\nabla b(X(\tau,x)))\,d\tau = -g(t,x)+\varepsilon\mathcal{R}^\varepsilon(t,x),$$

which implies immediately (3.11) and yields Lemma 4 as explained above.

3.5. Existence on a uniform time interval. As an immediate corollary of Lemma 4, we obtain that $X(t,\cdot)$ is a diffeomorphism of $\mathbf{R}^2$ on a uniform time interval. Indeed, DX is invertible as long as

$$\|DX - Id\|_{L^\infty} < 1.$$

COROLLARY 1. *Consider a function b satisfying assumptions* (H0) *and* (H1). *Let (ρ_0,u_0) be, respectively, a nonnegative function of $W^{s-1,\infty}(\mathbf{R}^2)$ and a vector field of $W^{s,\infty}(\mathbf{R}^2)$ ($s\geq 1$). Then, for all $T < \|u_0\|_{L^\infty}^{-1}\|\nabla b\|_{L^\infty}^{-1}$, there exists $\varepsilon_T>0$ such that system* (1.2) *admits a unique solution $(\rho_\varepsilon,u_\varepsilon)\in L^\infty([0,T],W^{s-1,\infty}(\mathbf{R}^2)\times W^{s,\infty}(\mathbf{R}^2))$ for all $\varepsilon\leq\varepsilon_T$.*

Proof of Corollary 1. By Lemma 4, the trajectories defined by (2.8) are continuously differentiable and satisfy, for all $t\leq T$ and all $x\in\mathbf{R}^2$,

$$\left\| DX(t,x) - Id - tu_0\otimes\nabla\log b\cos\left(\frac{\phi(t,x)}{\varepsilon}\right) + tu_0^\perp\otimes\nabla\log b\sin\left(\frac{\phi(t,x)}{\varepsilon}\right)\right\| \leq C_T\varepsilon.$$

This implies in particular the following estimate on the Jacobian $J(t,x) \stackrel{\text{def}}{=} |\det(DX(t,x))|$:

$$\left| J(t,x) - 1 - tu_0\cdot\nabla\log b\cos\left(\frac{\phi(t,x)}{\varepsilon}\right) + tu_0^\perp\cdot\nabla\log b\sin\left(\frac{\phi(t,x)}{\varepsilon}\right)\right| \leq C_T\varepsilon.$$

Then for $T<\|u_0\|_{L^\infty}^{-1}\|\nabla b\|_{L^\infty}^{-1}$, there exists ε_T such that

$$\forall\varepsilon\leq\varepsilon_T,\quad \forall t\in[0,T],\quad \sup_{x\in\mathbf{R}^2}|J(t,x)-1|<1,$$

which means that X is a C^1-diffeomorphism of $\mathbf{R}^2$.

Moreover, from formula (2.8) we can deduce by induction that $X(t,.)$ (and consequently its inverse $X^{-1}(t,\cdot)$) is smooth, its regularity being the same as the regularity of the initial velocity field u_0. Then the vector field u given by

$$u(t,x) = u_0(X^{-1}(t,x))\cos\left(\frac{\phi(t,X^{-1}(t,x))}{\varepsilon}\right) - u_0^\perp(X^{-1}(t,x))\cos\left(\frac{\phi(t,X^{-1}(t,x))}{\varepsilon}\right)$$

belongs to $L^\infty([0,T],W^{s,\infty}(\mathbf{R}^2))$, and it is easy to check that it satisfies system (1.3) in a strong sense.

The density ρ is then obtained as the strong solution of the linear transport equation

$$\partial_t\rho + u\cdot\nabla\rho + \rho\nabla\cdot u = 0$$

whose coefficients belong to $L^\infty([0,T],W^{s-1,\infty}(\mathbf{R}^2))$, with initial data in $W^{s-1,\infty}(\mathbf{R}^2)$. It therefore stays in $L^\infty([0,T],W^{s-1,\infty}(\mathbf{R}^2))$. We emphasize once again that no uniform bound on (ρ,u) is available in $L^\infty([0,T],W^{s-1,\infty}(\mathbf{R}^2)\times W^{s,\infty}(\mathbf{R}^2))$.

Theorem 2 is proved.

Remark 3. The supremum of the life span of the solutions corresponds to a crossing phenomenon, to be compared with the caustic in geometrical optics. Beyond this time, the differential system

$$\left\{\begin{array}{l}\dot X = \xi,\\ \dot\xi = \xi\wedge b\end{array}\right.$$

with initial data $(x,u_0(x))_{x\in\mathbf{R}^2}$ still admits a unique smooth solution, but the application $(X(t,x),\xi(t,x))\mapsto X(t,x)$ is no longer injective, and it cannot be lifted. The hyperbolic system (1.3) no longer has a solution.

4. Study of the asymptotics of u_ε and ρ_ε. Let $T<T^*=\|u_0\|_{L^\infty}^{-1}\|\nabla b\|_{L^\infty}^{-1}$ be fixed. Then, for any $\varepsilon\le\varepsilon_T$ as in Corollary 1, the solution (ρ,u) of system (1.2) with initial data $(\rho_0,u_0)\in W^{s-1,\infty}(\mathbf{R}^2)\times W^{s,\infty}(\mathbf{R}^2)$ belongs to $L^\infty([0,T],W^{s-1,\infty}(\mathbf{R}^2)\times W^{s,\infty}(\mathbf{R}^2))$. Then it makes sense to study their asymptotic behavior as $\varepsilon\to 0$, and the aim of this section is to prove Theorems 3 and 4.

Subsection 4.1 is devoted to the asymptotics of $u(t,X(t,x))$ and $\rho(t,X(t,x))$. Subsection 4.2 consists in inverting the characteristics in order to infer Theorems 3 and 4.

4.1. Asymptotics of $u(t,X)$ and $\rho(t,X)$. From the characteristic formulation of system (1.2) and the asymptotic expansion of $X(t,\cdot)$, we immediately deduce the asymptotic behavior of $u(t,X(t,\cdot))$ and $\rho(t,X(t,\cdot))$.

PROPOSITION 2. *Consider a function b satisfying assumptions* (H0) *and* (H1). *Let u_0 be a vector field in $W^{s,\infty}(\mathbf{R}^2)$ ($s\ge 1$). For all $T<T^*$ and $\varepsilon\le\varepsilon_T$ as in Theorem* 2, *denote by u the solution of* (1.3) *in $L^\infty([0,T],W^{s,\infty}(\mathbf{R}^2))$. Then*

$$u(t,X(t,x)) - \Big(u_0(x)\cos(\widetilde\phi_\varepsilon(t,x)) - u_0^\perp(x)\sin(\widetilde\phi_\varepsilon(t,x))\Big)$$

converges strongly to 0 *in $L^\infty([0,T]\times\mathbf{R}^2)$, at speed $O(\varepsilon)$, where the phase $\widetilde\phi_\varepsilon$ is defined by*

$$\widetilde\phi_\varepsilon(t,x) = \frac{b(x)t}{\varepsilon} - t\left(u_0^\perp(x)\cdot\nabla\right)\log b(x). \tag{4.1}$$

Proof of Proposition 2. Let us first recall that

$$u(t,X(t,x)) = u_0(x)\cos\left(\frac{\phi(t,x)}{\varepsilon}\right) - u_0^\perp(x)\sin\left(\frac{\phi(t,x)}{\varepsilon}\right),$$

where the phase ϕ is given by

$$\phi(t,x) = \int_0^t b(X(s,x))ds.$$

Then in order to establish Proposition 2, we have to approximate the phase. By Lemma 2,

$$b(X(t,x)) = b(x) + \frac{\varepsilon u_0(x)}{b(x)} \sin\left(\frac{\phi(t,x)}{\varepsilon}\right) \cdot \nabla b(x) - \frac{\varepsilon u_0^\perp(x)}{b(x)} \left(1 - \cos\left(\frac{\phi(t,x)}{\varepsilon}\right)\right) \cdot \nabla b(x) + \varepsilon^2 \mathcal{R}^\varepsilon(t,x),$$

noticing that $v \cdot \nabla b = 0$. It follows that

$$\begin{aligned} u(t, X(t,x)) = u_0(x) \cos\Bigg(& \frac{b(x)t}{\varepsilon} + \frac{u_0(x)\cdot\nabla b(x)}{b(x)} \int_0^t \sin\left(\frac{\phi(s,x)}{\varepsilon}\right) ds \\ & - \frac{u_0^\perp(x)\cdot\nabla b(x)}{b(x)} \int_0^t \left(1 - \cos\left(\frac{\phi(s,x)}{\varepsilon}\right)\right) ds + \varepsilon \mathcal{R}^\varepsilon(t,x)\Bigg) \\ - u_0^\perp(x) \sin\Bigg(& \frac{b(x)t}{\varepsilon} + \frac{u_0(x)\cdot\nabla b(x)}{b(x)} \int_0^t \sin\left(\frac{\phi(s,x)}{\varepsilon}\right) ds \\ & - \frac{u_0^\perp(x)\cdot\nabla b(x)}{b(x)} \int_0^t \left(1 - \cos\left(\frac{\phi(s,x)}{\varepsilon}\right)\right) ds + \varepsilon \mathcal{R}^\varepsilon(t,x)\Bigg). \end{aligned}$$

Finally, remembering that due to Lemma 1

$$\left| \int_0^t \sin\left(\frac{\phi(s,x)}{\varepsilon}\right) ds \right| + \left| \int_0^t \cos\left(\frac{\phi(s,x)}{\varepsilon}\right) ds \right| \leq \varepsilon \mathcal{R}^\varepsilon(t,x),$$

with the usual uniform bounds on $\mathcal{R}^\varepsilon$, this yields Proposition 2.

The asymptotic behavior of ρ is obtained in a similar way using the fact that ρ is proportional to the Jacobian $J(t,x) = |\det DX(t,x)|$.

PROPOSITION 3. *Consider a function b satisfying assumptions* (H0) *and* (H1). *Let ρ_0 be a nonnegative function in $W^{s-1,\infty}(\mathbf{R}^2)$, and let u_0 be a vector field in $W^{s,\infty}(\mathbf{R}^2)$ ($s \geq 1$). For all $T < T^*$ and $\varepsilon \leq \varepsilon_T$ as in Theorem* 2, *denote by (ρ, u) the solution of* (1.2) *in $L^\infty([0,T], W^{s-1,\infty}(\mathbf{R}^2))$ and $L^\infty([0,T], W^{s,\infty}(\mathbf{R}^2))$, respectively. Then*

$$\rho(t, X(t,x)) - \rho_0(x) \left(1 + t u_0 \cdot \nabla \log b(x) \sin(\widetilde{\phi}_\varepsilon(t,x)) - t u_0^\perp \cdot \nabla \log b(x) \cos(\widetilde{\phi}_\varepsilon(t,x))\right)$$

converges strongly to 0 *in $L^\infty([0,T] \times \mathbf{R}^2)$, where the phase $\widetilde{\phi}_\varepsilon$ is defined as in* (4.1).

Proof of Proposition 3. As long as the solution of (1.2) is regular, the equation governing ρ can be rewritten

$$\frac{d}{dt}(\log \rho) = \nabla \cdot u,$$

where $\frac{d}{dt}$ denotes as usual the derivative along the trajectories associated with the flow. Of course, the Liouville theorem implies that the equation on the Jacobian of the flow states

$$\frac{d}{dt}(J) = \nabla \cdot u.$$

Then, for all $\varepsilon \leq \varepsilon_T$, all $t \in [0,T]$, and all $x \in \mathbf{R}^2$,

$$\rho(t, X(t,x)) = \rho_0(x) J(t,x),$$

since $J_0(t,x) = \det(Id) = 1$.

From Lemma 4 we then deduce that

$$\left|\rho(t,X(t,x)) - \rho_0(x)\left(1 + tu_0\cdot\nabla\log b(x)\sin\left(\frac{\phi(t,x)}{\varepsilon}\right) - tu_0^\perp\cdot\nabla\log b(x)\left(\frac{\phi(t,x)}{\varepsilon}\right)\right)\right| \leq C_T\varepsilon. \tag{4.2}$$

Plugging back into formula (4.2) the approximation of the phase obtained previously,

$$\frac{\phi(t,x)}{\varepsilon} = \widetilde{\phi}_\varepsilon + \varepsilon\mathcal{R}^\varepsilon(t,x), \tag{4.3}$$

then leads to the expected asymptotics.

4.2. Inversion of the characteristics. In this section we shall prove Theorems 3 and 4. From now on T^* is the time given by Theorem 2, and we will call T any time smaller than T^* (in the following we will also suppose $\varepsilon \leq \varepsilon_T$ as given in Theorem 2).

Let $X^{-1}(t,x)$ be the point at time 0 of the trajectory reaching x at time t. By Proposition 2, we have

$$u(t,x) = u_0(X^{-1}(t,x))\cos\left(\widetilde{\phi}_\varepsilon(t,X^{-1}(t,x))\right) - u_0^\perp(X^{-1}(t,x))\sin\left(\widetilde{\phi}_\varepsilon(t,X^{-1}(t,x))\right) + \varepsilon\mathcal{R}^\varepsilon(t,x)$$

with the usual uniform bounds on $\mathcal{R}^\varepsilon$. That remainder function $\mathcal{R}^\varepsilon(t,x)$ is liable to change from line to line in this subsection.

By Proposition 1 there is a constant C_T (depending on T, u_0, and b), such that

$$\forall x\in\mathbf{R}^2,\quad \forall t\in[0,T],\quad |X^{-1}(t,x) - x| \leq C_T\varepsilon, \tag{4.4}$$

so we can write rather

$$u(t,x) = u_0(x)\cos\left(\widetilde{\phi}_\varepsilon(t,X^{-1}(t,x))\right) - u_0^\perp(x)\sin\left(\widetilde{\phi}_\varepsilon(t,X^{-1}(t,x))\right) + \varepsilon\mathcal{R}^\varepsilon(t,x).$$

By definition of $\widetilde{\phi}_\varepsilon$ in (4.1), we have, using again (4.4),

$$\widetilde{\phi}_\varepsilon(t,X^{-1}(t,x)) = \frac{b(x)t}{\varepsilon} + \frac{t}{\varepsilon}(X^{-1}(t,x)-x)\cdot\nabla b(x) - tu_0^\perp(x)\cdot\nabla\log b(x) + \varepsilon\mathcal{R}^\varepsilon(t,x),$$

and hence, defining

$$\widetilde{\theta}_\varepsilon(t,x) \stackrel{\text{def}}{=} \frac{t}{\varepsilon}(X^{-1}(t,x)-x)\cdot\nabla b(x) - tu_0^\perp(x)\cdot\nabla\log b(x),$$

we have

$$\widetilde{\phi}_\varepsilon(t,X^{-1}(t,x)) = \frac{b(x)t}{\varepsilon} + \widetilde{\theta}_\varepsilon(t,x) + \varepsilon\mathcal{R}^\varepsilon(t,x). \tag{4.5}$$

Now we shall try to make $\widetilde{\theta}_\varepsilon$ more precise. According to Lemma 2 and the approximation for the phase derived in the previous paragraph, we have

$$x - X^{-1}(t,x) = \varepsilon\frac{u_0(x)}{b(x)}\sin\left(\widetilde{\phi}_\varepsilon(t,X^{-1}(t,x)) + \varepsilon\mathcal{R}^\varepsilon(t,x)\right) - \varepsilon\frac{u_0^\perp(x)}{b(x)}\left(1-\cos\left(\widetilde{\phi}_\varepsilon(t,X^{-1}(t,x)) + \varepsilon\mathcal{R}^\varepsilon(t,x)\right)\right) - \varepsilon tv(x) + \varepsilon^2\mathcal{R}^\varepsilon(t,x),$$

where again we have used (4.4). So we obtain, using the fact that $v \cdot \nabla b = 0$,

$$\widetilde{\theta}_\varepsilon(t,x) = -tu_0(x)\cdot\nabla \log b(x) \sin\left(\widetilde{\phi}_\varepsilon(t,X^{-1}(t,x)) + \varepsilon\mathcal{R}^\varepsilon(t,x)\right) + tu_0^\perp(x)\cdot\nabla\log b(x)\cos\left(\widetilde{\phi}_\varepsilon(t,X^{-1}(t,x)) + \varepsilon\mathcal{R}^\varepsilon(t,x)\right) + \varepsilon\mathcal{R}^\varepsilon(t,x),$$

which by (4.5) yields directly the result (1.4), defining $\theta_\varepsilon(t,x) = \widetilde{\theta}_\varepsilon(t,x) + \frac{b(x)t}{\varepsilon}$.

Theorem 3 is proved.

The proof of Theorem 4 is now immediate: we use the formula obtained in Proposition 3 and replace $\rho_\varepsilon(t,X(t,x))$ by $\rho_\varepsilon(t,x)$ using the above formulation of $X^{-1}(t,x)$. The result follows.

REFERENCES

[1] A. Babin, A. Mahalov, and B. Nicolaenko, *Regularity and integrability of 3D Euler and Navier-Stokes equations for rotating fluids*, Asymptot. Anal., 15 (1997), pp. 103–150.

[2] Y. Brenier and E. Grenier, *Sticky particles and scalar conservation laws*, SIAM J. Numer. Anal., 35 (1998), pp. 2317–2328.

[3] J.-Y. Chemin, B. Desjardins, I. Gallagher, and E. Grenier, *Anisotropy and dispersion in rotating fluids*, in Nonlinear Partial Differential Equations and Their Applications, Collège de France Seminar, Stud. Math. Appl. 31, North-Holland, Amsterdam, 2002, pp. 171–192.

[4] E. Frénod and E. Sonnendrücker, *Homogenization of the Vlasov equation and the Vlasov–Poisson system with a strong external magnetic field*, Asymptot. Anal., 18 (1998), pp. 193–213.

[5] I. Gallagher and L. Saint-Raymond, *Weak convergence results for inhomogeneous rotating fluid equations*, C. R. Math. Acad. Sci. Paris, 336 (2003), pp. 401–406.

[6] F. Golse and L. Saint-Raymond, *The Vlasov-Poisson system with strong magnetic field*, J. Math. Pures Appl. (9), 78 (1999), pp. 791–817.

[7] E. Grenier, *Pseudo-differential energy estimates of singular perturbations*, Comm. Pure Appl. Math., 50 (1997), pp. 821–865.

[8] S. Schochet, *Fast singular limits of hyperbolic PDEs*, J. Differential Equations, 114 (1994), pp. 476–512.

SIAM J. MATH. ANAL.
Vol. 36, No. 4, pp. 1177–1199

GLOBAL SOLUTIONS OF NONLINEAR TRANSPORT EQUATIONS FOR CHEMOSENSITIVE MOVEMENT*

HYUNG JU HWANG†‡, KYUNGKEUN KANG†§, AND ANGELA STEVENS†

Abstract. A widespread phenomenon in moving microorganisms and cells is their ability to reorient themselves depending on changes of concentrations of certain chemical signals. In this paper we discuss kinetic models for chemosensitive movement, which also takes into account evaluations of gradient fields of chemical stimuli which subsequently influence the motion of the respective microbiological species. The basic type of model was discussed by Alt [*J. Math. Biol.*, 9 (1980), pp. 147–177], [*J. Reine Angew. Math.*, 322 (1981), pp. 15–41] and by Othmer, Dunbar, and Alt [*J. Math. Biol.*, 26 (1988), pp. 263–298]. Chalub et al. rigorously proved that, in three dimensions, these kinds of kinetic models lead to the classical Keller–Segel model as its drift-diffusion limit when the equation for the chemo-attractant is of elliptic type [*Monatsh. Math.*, 142 (2004), pp. 123–141], [*On the Derivation of Drift-Diffusion Model for Chemotaxis from Kinetic Equations*, ANUM preprint 14/02, Vienna Technical University, 2002]. In [H. Hwang, K. Kang, and A. Stevens, *Drift-diffusion limits of kinetic models for chemotaxis: A generalization*, Discrete Contin. Dyn. Syst. Ser. B., to appear] it was proved that the macroscopic diffusion limit exists in both two and three dimensions also when the equation of the chemo-attractant is of parabolic type. So far in the rigorous derivations, only the density of the chemo-attractant was supposed to influence the motion of the chemosensitive species. Here we show that in the macroscopic limit some types of evaluations of gradient fields of the chemical stimulus result in a change of the classical parabolic Keller–Segel model for chemotaxis. Under suitable structure conditions, global solutions for the kinetic models can be shown.

Key words. chemosensitive movement, sensing of gradient fields, nonlinear transport equations, global solutions, drift-diffusion limit, Keller–Segel model

AMS subject classifications. 35K55, 45K05, 82C70, 92C17

DOI. 10.1137/S0036141003431888

1. Introduction.

The starting point of our considerations is the classical chemotaxis model as discussed by Keller and Segel (see [14] and [15]). This system is of advection-diffusion type and consists of two coupled parabolic equations:

$$\frac{\partial \rho}{\partial t} = \nabla \cdot (D(\rho, S)\nabla\rho - \chi(\rho, S)\rho\nabla S), \tag{1.1}$$

$$\tau\frac{\partial S}{\partial t} = D_0\Delta S + \alpha\rho - \beta S, \quad \alpha, \beta, \tau \geq 0. \tag{1.2}$$

Here $\rho = \rho(x,t)$ denotes the density of chemotactic cells and $S = S(x,t)$ is the density of the chemo-attractant. The cells are attracted by the chemical, and χ denotes their chemotactic sensitivity. The first rigorous derivation of the macroscopic chemotaxis equations from microscopic models, namely, interacting stochastic many particle systems, was given in [21]. In [11] a survey about known results on existence of global solutions and finite time blowup for this type of model was given.

*Received by the editors July 20, 2003; accepted for publication (in revised form) May 14, 2004; published electronically February 3, 2005.
http://www.siam.org/journals/sima/36-4/43188.html
†Max-Planck-Institute for Mathematics in the Sciences, Inselstr. 22 - 26, D-04103 Leipzig, Germany (hwang@mis.mpg.de, kkang@mis.mpg.de, stevens@mis.mpg.de).
‡Current address: Department of Mathematics, Duke University, Durham, NC 27708 (hjhwang@math.duke.edu).
§Current address: Department of Mathematics, University of British Columbia, Vancouver, BC, Canada V6T 1Z2 (kkang@pims.math.ca).

In [3] a kinetic model for (1.1) was discussed coupled with the Poisson equation without decay term

$$-\Delta S = \alpha\rho. \tag{1.3}$$

In [3, p. 3] the following kinetic equation for the oriented cell density $f = f(x, v, t) \geq 0$ was considered:

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f = \int_V (T[S]f' - T^*[S]f)dv', \tag{1.4}$$

where x, v, and t indicate position, velocity, and time, respectively. Here the abbreviations $f' = f(x, v', t)$, $T[S] = T[S](x, v, v', t)$, and $T^*[S] = T[S](x, v', v, t)$ are used. The first term on the right-hand side of (1.4) describes the turning into direction v, and the second term the turning away from v. The cell density ρ fulfills

$$\rho(x, t) = \int_V f(x, v, t)dv,$$

where V is the set of admissible velocities which is assumed to be compact.

Using stochastic models for the motion of bacteria and leukocytes, Alt derived (1.1) from a transport equation similar to (1.4) [1, section 8], [2, section 3]. Later a general formulation of this velocity-jump process was presented and studied in [18, section 3]. In [10] and [19] Othmer and Hillen studied the formal diffusion limit of a transport equation of (1.4) by moment expansions, which generalizes parts of Alt's earlier works [1], [2]. A hyperbolic scaling and its formal limit were discussed in [6].

Based on [19] a rigorous proof of the macroscopic limit was given in [3]. After using diffusive scaling of time and space, the nondimensional form of (1.4) leads to [3, p. 4]

$$\epsilon^2 \frac{\partial f_\epsilon}{\partial t} + \epsilon v \cdot \nabla_x f_\epsilon = -\mathcal{T}_\epsilon[S_\epsilon](f_\epsilon), \quad x \in \mathbb{R}^n,\ v \in V,\ t > 0, \tag{1.5}$$

where

$$\mathcal{T}_\epsilon[Z](g) = \int_V (T^*_\epsilon[Z]g - T_\epsilon[Z]g')dv'.$$

The diffusion limit $\epsilon \to 0$ was studied for initial conditions

$$f_\epsilon(x, v, 0) \equiv f_0(x, v), \quad x \in \mathbb{R}^n,\ v \in V, \tag{1.6}$$

with (1.5) coupled to (1.3) for the chemo-attractant. In [3] it was shown that the coupled nonlinear system (1.5), (1.6), and (1.3) resulted in Keller–Segel-type equations for chemotaxis as its macroscopic drift-diffusion limit under suitable conditions on the turning kernel in three dimensions (compare, e.g., [3, Theorem 5] and [4, Theorem 2]). In [3] and [4] also global solutions were proved for suitable turning kernels for fixed $\epsilon > 0$.

In [12], as an extension of [3], the authors proved that such kinetic models have a macroscopic diffusion limit in both two and three dimensions also when the equation of the chemo-attractant is of parabolic type, i.e., $\tau > 0$, which is the original version of the chemotaxis model. An independent related result was given in [5].

In this article, we consider turning kernels depending not only on S but also on ∇S, as formally discussed, among others, in [22] and [19], i.e.,

$$\epsilon^2 \frac{\partial f_\epsilon}{\partial t} + \epsilon v \cdot \nabla_x f_\epsilon = -\mathcal{T}_\epsilon[S_\epsilon, \nabla S_\epsilon](f_\epsilon), \quad x \in \mathbb{R}^n,\ v \in V,\ t > 0, \tag{1.7}$$

with initial condition (1.6) coupled to

$$\tau \frac{\partial S_\epsilon}{\partial t} = \Delta S_\epsilon + \alpha \rho_\epsilon - \beta S_\epsilon, \quad \tau \geq 0,\ \alpha > 0,\ \beta \geq 0, \tag{1.8}$$

where

$$\rho_\epsilon = \int_V f_\epsilon dv. \tag{1.9}$$

In what follows, for notational convenience, we write $\mathcal{T}_\epsilon[S_\epsilon]$ instead of $\mathcal{T}_\epsilon[S_\epsilon, \nabla S_\epsilon]$, unless any confusion is to be expected. Here we emphasize that the conditions on the turning kernel include also detection of spatial gradients of the chemo-attractant by the chemotactic cells. This behavior results under certain conditions in a macroscopic model which varies from the classical Keller–Segel system by additional higher order terms.

Our main result is that for suitable turning kernels which take into account the effects of gradient measurements of the chemical, global solutions exist also in two dimensions, and thus blowup of the solutions does not happen in finite time (compare Theorems 3.6 and 3.12 for the elliptic and parabolic cases, respectively).

The result is extended to three dimensions under some restrictions on the turning kernels. We also show the existence of a macroscopic diffusion limit of the kinetic model in two and three dimensions. More precisely, under similar assumptions on the turning kernel $T[S]$ as given in [3], we prove that the coupled nonlinear system (1.6), (1.7), and (1.8) converges to Keller–Segel-type equations and their variants for $\epsilon \to 0$ (compare Theorem 4.4). Our main tool is the potential estimate for S. In particular, in case the chemo-attractant equation is of elliptic type, i.e., $\tau = 0$ and in two dimensions, log-type estimates for the chemical S are used to obtain global existence for the kinetic model (similar techniques were used in [13, Lemma 4]).

The plan of this paper is as follows: In section 2, we introduce some notation used and briefly review the derivation of the macroscopic equation as presented in [3] and [12]. In section 3, we prove that the kinetic model (1.7)–(1.9) has a global solution for "suitable" turning kernels. In section 4, we prove the existence of the diffusion limit for a short time interval. In section 5 we give concrete examples on how the specific dependencies of the turning kernel result in different types of macroscopic equations.

2. Preliminaries. We first introduce some notation which will be used throughout this article and recall some of the observations presented in [3].

• By G we denote the Bessel potential, which is the fundamental solution of the differential operator $1 - \Delta$ in $\mathbb{R}^n$ (see [20, pp. 130–132]):

$$G(x) = \frac{1}{4\pi} \int_0^\infty e^{-\pi \frac{|x|^2}{4s} - \frac{s}{4\pi}} s^{\frac{-n+2}{2}} \frac{ds}{s}. \tag{2.1}$$

• By Γ we denote the fundamental solution of the differential operator $\partial_t - \Delta_x + \beta$ in $\mathbb{R}^n \times \mathbb{R}_+$:

$$\Gamma(x,t) = \frac{1}{(4\pi t)^{\frac{n}{2}}} \exp\left(-\frac{|x|^2}{4t} - \beta t\right). \tag{2.2}$$

• By $C = C(\alpha, \beta, \dots)$ we denote a constant depending on the prescribed quantities $\alpha, \beta, \dots$. The domain Ω considered in this article is $\mathbb{R}^n$, $n = 2, 3$.

To make this note self-contained, we review the formal derivation of the macroscopic equation from the kinetic model presented in [3] (compare the details in [3, pp. 5–7]). For simplicity we assume for a moment that $\tau = 1$, $\alpha = 1$, and $\beta = 1$ (other cases can be formally derived in a similar way without any difficulty). Since the integral of $\mathcal{T}_\epsilon[S](f)$ with respect to the velocity vanishes, we obtain the macroscopic conservation equation

$$\frac{\partial \rho_\epsilon}{\partial t} + \nabla \cdot J_\epsilon = 0, \tag{2.3}$$

where $J_\epsilon(x,t) = \epsilon^{-1} \int_V v f_\epsilon(x,v,t) dv$ is the flux density. The turning kernel is assumed to have the following asymptotic expansion: $T_\epsilon[S] = T_0[S] + \epsilon T_1[S] + O(\epsilon^2)$. Then the turning operator can be expanded in a similar way and

$$\mathcal{T}_\epsilon[S](f) = \int_V (T_\epsilon^*[S] f - T_\epsilon[S] f') dv'.$$

By asymptotic expansion of $f_\epsilon = f_0 + \epsilon f_1 + O(\epsilon^2)$ and $S_\epsilon = S_0 + \epsilon S_1 + O(\epsilon^2)$, the equation for the leading order terms can be obtained from (1.7):

$$\mathcal{T}_0[S_0](f_0) = 0, \quad S_0 = \rho_0 * \Gamma, \quad \rho_0 = \int_V f_0 dv. \tag{2.4}$$

Comparing coefficients in (1.7) results in

$$v \cdot \nabla_x f_0 = -\mathcal{T}_0[S_0](f_1) - \mathcal{T}_1[S_0](f_0) - \mathcal{T}_{0S}[S_0, S_1](f_0),$$

where $\mathcal{T}_{0S}[S_0, S_1]$ is part of the turning operator $\mathcal{T}$ and its kernel is the Fréchet derivative of T_0 with respect to S, evaluated at S_0 in the direction S_1. Here, we recall the assumptions on the leading order terms of the turning operator and two useful lemmas presented in [3, (A0), Lemma 1, Lemma 2, pp. 6–7].

Assumption 2.1. There exists a bounded velocity distribution $F(v) > 0$, such that $T_0^*[S]F = T_0[S]F'$ and

$$\int_V vF(v) dv = 0, \qquad \int_V F(v) dv = 1.$$

The turning rate $T_0[S]$ is bounded, and there exists a constant $\gamma = \gamma[S] > 0$ such that $T_0[S]/F \geq \gamma$ for all $(v, v') \in V \times V$, $x \in \mathbb{R}^n$, and $t > 0$.

LEMMA 2.2. *Let $\zeta : \mathbb{R} \to \mathbb{R}$, $g : V \to \mathbb{R}$, and let*

$$\phi_\epsilon^S[S] = \frac{T_\epsilon[S]F' + T_\epsilon^*[S]F}{2}, \qquad \phi_\epsilon^A[S] = \frac{T_\epsilon[S]F' - T_\epsilon^*[S]F}{2}$$

denote, respectively, the symmetric and antisymmetric parts of $T_\epsilon[S]F'$. Then

$$\begin{aligned}\int_V \int_V \mathcal{T}_\epsilon(Fg)\zeta(g) dv = &\frac{1}{2} \int_V \int_V \phi_\epsilon^S[S](g - g')(\zeta(g) - \zeta(g')) dv' dv \\ &+ \frac{1}{2} \int_V \int_V \phi_\epsilon^A[S](g + g')(\zeta(g) - \zeta(g')) dv' dv.\end{aligned}$$

The same holds for $\mathcal{T}_\epsilon[S]$ with analogous definitions of $\phi_\epsilon^S[S]$ and $\phi_\epsilon^A[S]$.

Proof. See Lemma 1 in [3] for the proof. ∎

With $g = f/F$ and $\zeta = \mathrm{id}$ one obtains the following.

LEMMA 2.3. *Let Assumption* 2.1 *hold. Then the entropy equality*

$$\int_V \mathcal{T}_0[S](f)\frac{f}{F}dv = \frac{1}{2}\int_V\int_V \phi_0^S[S]\left(\frac{f}{F} - \frac{f'}{F'}\right)^2 dv'dv \geq 0$$

holds. For $g \in L^2(V; dv/F)$*, the equation* $\mathcal{T}_0[S](f) = g$ *has a unique solution* $f \in L^2(V; dv/F)$ *satisfying* $\int_V f dv = 0$ *if and only if* $\int_V g dv = 0$.

Proof. See Lemma 2 in [3] for the proof. ∎

From the entropy equality, we deduce that

$$f_0(x, v, t) = \rho_0(x, t)F(v).$$

Since $\mathcal{T}_{0S}[S_0, S_1](f_0) = 0$, we obtain

$$\mathcal{T}_0[S](f_1) = -vF \cdot \nabla\rho_0 - \rho_0\mathcal{T}_1[S_0](F).$$

The right-hand side satisfies the solvability condition from Lemma 2.3, and therefore the solution can be written as

$$f_1 = -\kappa(x, v, t) \cdot \nabla\rho_0(x, t) - \Theta(x, v, t)\rho_0(x, t) + \rho_1(x, t)F(v),$$

where $\kappa = \kappa[S_0]$ and $\Theta = \Theta[S_0]$ are the solutions of

$$\mathcal{T}_0[S_0](\kappa) = vF, \qquad \mathcal{T}_0[S_0](\Theta) = \mathcal{T}_1[S_0](F),$$

and ρ_1 is the macroscopic density of f_1, which is a new unknown. By passing to the limit $\epsilon \to 0$ in (2.3), the convection-diffusion equation reads

$$\partial_t\rho_0 - \nabla \cdot (D[S_0]\nabla\rho_0 - \rho_0 H[S_0]) = 0,$$

where

$$D[S_0](x, t) = \int_V v \otimes \kappa[S_0](x, v, t)dv, \qquad H[S_0] = -\int_V v\Theta[S_0](x, v, t)dv,$$

together with

$$\frac{\partial S_0}{\partial t} = \Delta S_0 + \rho_0 - S_0.$$

The specific form of $D[S_0]$ and $H[S_0]$ will depend on the choice of the turning kernels and will be discussed later.

3. Global solution of the kinetic model. In this section we show that solutions of the coupled system (1.6)–(1.9) in two and three dimensions do not blow up in finite time for fixed $\epsilon > 0$ if the turning kernel satisfies a certain structure condition. Without loss of generality we set $\epsilon = 1$ in (1.6) and $\alpha = 1$ in (1.8). We consider two problems, namely, the elliptic and the parabolic equations for the chemo-attractant.

We start with an inequality of Gronwall type in the next lemma. Since it is of the nonstandard form among the Gronwall-type inequalities, we present its proof for clarity, although the proof is similar to that of the usual one.

LEMMA 3.1. *Let a and b be positive constants. Let $y(t)$ and $y'(t)$ be positive and differentiable in t and satisfy*

$$y' \le ay \ln y' + by. \tag{3.1}$$

Then

$$y(t) \le \left[y(0) \exp\left(2b \int_0^t e^{-2as} ds \right) \right]^{\exp(2at)}.$$

Proof. We subtract and add $\ln y$ from the right-hand side of (3.1) to get $y' \le ay \ln y + ay \ln (\ln y)' + by$. Dividing both sides of the above inequality by y, we get $(\ln y)' \le a \ln y + a \ln (\ln y)' + b$. Set $z = \ln y$ to get $z' \le az + a \ln z' + b$. Since we may assume $\ln z' \le (1/2a)z'$ (otherwise, $\ln z' \le C$ and the above inequality reduces to a standard Gronwall inequality), we have $z' \le az + \frac{1}{2} z' + b$. We get $z' \le 2az + 2b$, where $z = \ln y$. Using a standard Gronwall argument, we deduce the lemma. ∎

The structure condition on the turning kernel $T[S]$ is assumed to be as follows.

Assumption 3.2 (structure condition). There exist nonnegative constants $C_i \ge 0$, $i = 1, 2, \ldots, 5$, such that for all $x \in \mathbb{R}^n$, $n = 2, 3$, $v, v' \in V$, $t \in \mathbb{R}^+$, and $S \in W^{1,\infty}(\mathbb{R}^n)$, the turning kernel T satisfies

$$\begin{aligned} 0 \le T_\epsilon[S](x, v, v', t) &\le C_1 + C_2 S(x + \epsilon v, t) + C_3 S(x - \epsilon v', t) \\ &\quad + C_4 |\nabla S(x + \epsilon v, t)| + C_5 |\nabla S(x - \epsilon v', t)|, \end{aligned} \tag{3.2}$$

$$\begin{aligned} |\nabla T_\epsilon[S](x, v, v', t)| &\le C_2 |\nabla S(x + \epsilon v, t)| + C_3 |\nabla S(x - \epsilon v', t)| \\ &\quad + C_4 |\nabla^2 S(x + \epsilon v, t)| + C_5 |\nabla^2 S(x - \epsilon v', t)|. \end{aligned} \tag{3.3}$$

This means that the cells can measure the concentration and the spatial gradient of the chemo-attractant up to a distance ϵ from their position, and this may affect the movement of the cells.

Remark 3.3. The turning kernel, as given above, describes the turning from direction v' into direction v. This means that the actual or "old" direction is evaluated by checking backwards, whereas the evaluation of possible new directions are checked forwards (e.g., by lamelliopodial protrusion). Checking the possible new directions backwards if compared to the actual direction of motion is also possible and could have been taken into account in the following considerations. Nevertheless, it is important to note that a forward evaluation of the actual direction v' causes a technical problem in our approach so far.

We first consider the case that the chemo-attractant equation is of elliptic type.

3.1. Elliptic case: $\tau = 0$. In this part, we consider the elliptic equation for the chemo-attractant S for two cases: $\beta > 0$ and $\beta = 0$. When $\beta > 0$ we may set $\beta = 1$ without loss of generality. So

$$-\Delta S = \rho - \beta S, \quad \beta \in \{0, 1\}, \quad n = 2, 3. \tag{3.4}$$

For $n = 2$ we need some preliminaries and start with elementary properties of the Bessel potential G in two dimensions.

LEMMA 3.4. *Let G be the Bessel potential in $\mathbb{R}^2$. Then $G \in L^p(\mathbb{R}^2)$ for any p with $1 \le p < \infty$ and $\nabla G \in L^p(\mathbb{R}^2)$ for any p with $1 \le p < 2$. Furthermore,*

$$\|G\|_{L^p(\mathbb{R}^2)} \le Cp, \quad 1 \le p < \infty, \tag{3.5}$$

$$\|\nabla G\|_{L^p(\mathbb{R}^2)} \le C \frac{2p}{2 - p}, \quad 1 \le p < 2. \tag{3.6}$$

Proof. For $n = 2$, the Bessel potential is (cf. (2.1))

$$G(x) = \frac{1}{4\pi}\int_0^\infty e^{-\pi\frac{|x|^2}{4s}-\frac{s}{4\pi}}\frac{ds}{s}.$$

Using a change of variables, we have

$$\|G\|_{L^p(\mathbb{R}^2)} \le C\int_0^\infty \frac{e^{-s}}{s}\big\|e^{-\frac{|x|^2}{4s}}\big\|_{L^p(\mathbb{R}^2)}ds \le C\int_0^\infty e^{-s}s^{-1+1/p}ds \le Cp.$$

We thus obtain (3.5). In a similar way we get

$$\|\nabla G\|^p_{L^p(\mathbb{R}^2)} \le C\int_0^\infty \frac{e^{-s}}{s^2}\big\|xe^{-\frac{|x|^2}{4s}}\big\|_{L^p(\mathbb{R}^2)}ds \le C\int_0^\infty e^{-s}s^{-\frac{3}{2}+\frac{1}{p}}ds \le C\frac{2p}{2-p},$$

as long as $1 \le p < 2$. Therefore we deduce (3.6). ∎

The next lemma shows various estimates for the chemo-attractant S.

LEMMA 3.5. *Let S be a solution of* (3.4) *in* $\mathbb{R}^2$. *Then S satisfies the following estimates:*

$$\|S(t)\|_{L^p(\mathbb{R}^2)} + \|\nabla S(t)\|_{L^q(\mathbb{R}^2)} \le C(p,q)\|\rho_0\|_{L^1(\mathbb{R}^2)}, \quad 1 \le p < \infty,\ 1 \le q < 2, \tag{3.7}$$

$$\|\nabla S(t)\|_{L^2(\mathbb{R}^2)} \le C\|\rho_0\|_{L^1(\mathbb{R}^2)}\big[\ln\big(\|\rho(t)\|^2_{L^2(\mathbb{R}^2)}+1\big)\big]^{1/2}. \tag{3.8}$$

Proof. The first estimate (3.7) is an easy consequence of mass conservation, Lemma 3.4, and Young's inequality (see, e.g., [7, pp. 624–625]). Thus it suffices to show the estimate (3.8).

From (3.4) we obtain the Fourier transform $\hat{S}(\xi) = \hat{\rho}(\xi)/(|\xi|^2+1)$, and thus

$$\|\nabla S(t)\|_{L^2(\mathbb{R}^2)} = \|\xi\hat{S}(t)\|_{L^2(\mathbb{R}^2)} = \left\|\frac{|\xi|\hat{\rho}(t)}{|\xi|^2+1}\right\|_{L^2(\mathbb{R}^2)},$$

where Plancherel's equality is used. The above integral can be estimated by splitting $\mathbb{R}^2$ of the ξ-space into two parts:

$$\int_{\mathbb{R}^2}\frac{|\xi\hat{\rho}(t)|^2}{(|\xi|^2+1)^2}d\xi = \int_{|\xi|<R}+\cdots+\int_{|\xi|>R}+\cdots = I_1+I_2,$$

where $R > 0$ will be chosen later. Using Hölder's inequality and Plancherel's equality we have

$$I_1 \le \|\hat{\rho}(t)\|^2_{L^\infty(\mathbb{R}^2)}\int_{|\xi|<R}\frac{|\xi|^2}{(|\xi|^2+1)^2}d\xi \le C\|\rho(t)\|^2_{L^1(\mathbb{R}^2)}\ln(R^2+1),$$

$$I_2 \le \left\|\frac{|\xi|}{|\xi|^2+1}\right\|^2_{L^\infty(|\xi|>R)}\|\hat{\rho}(t)\|^2_{L^2(\mathbb{R}^2)} \le CR^{-2}\|\rho(t)\|^2_{L^2(\mathbb{R}^2)}.$$

Therefore, by choosing $R = \|\rho(t)\|_{L^2(\mathbb{R}^2)}$, we obtain

$$\begin{aligned}\|\nabla S(t)\|_{L^2(\mathbb{R}^2)} &\le C\|\rho(t)\|_{L^1(\mathbb{R}^2)}\{\ln(R^2+1)\}^{1/2} + CR^{-1}\|\rho(t)\|_{L^2(\mathbb{R}^2)}\\ &\le C\left[1+\|\rho(t)\|_{L^1(\mathbb{R}^2)}\big\{\ln\big(\|\rho(t)\|^2_{L^2(\mathbb{R}^2)}+1\big)\big\}^{1/2}\right].\end{aligned}$$

Since $\|\rho\|_{L^1(\mathbb{R}^2)} = \|f_0\|_{L^1(\mathbb{R}^2\times V)}$, we deduce (3.8) and our lemma. ∎

The next theorem shows global existence of solutions for system (1.6)–(1.9) with $\tau = 0$, namely, blowup does not happen in finite time.

THEOREM 3.6. *Suppose the chemo-attractant equation is of elliptic type ($\tau = 0$). Assume that $f_0, \nabla f_0 \in (L^1 \cap L^\infty)(\mathbb{R}^n \times V)$, with $n = 2, 3$.*

1. *Case $n = 2$, $\beta > 0$: Let Assumption* 3.2 *hold. Then there exist global solutions $f, \nabla f \in L^\infty_{\mathrm{loc}}((0,\infty); L^1 \cap L^\infty(\mathbb{R}^2 \times V))$ and $S \in L^\infty_{\mathrm{loc}}((0,\infty); W^{1,p}(\mathbb{R}^2))$ for all $1 \le p \le +\infty$ of the system* (1.6)–(1.9) *with $\epsilon > 0$ fixed but arbitrary.*

2. *Case $n = 2$, $\beta = 0$: Let Assumption* 3.2 *hold with $C_2 = C_3 = C_5 = 0$. Then there exist global solutions $f, \nabla f \in L^\infty_{\mathrm{loc}}((0,\infty); L^1 \cap L^\infty(\mathbb{R}^2 \times V))$ and $\nabla S \in L^\infty_{\mathrm{loc}}((0,\infty); L^p(\mathbb{R}^2))$ for all $2 < p \le \infty$ of the system* (1.6)–(1.9) *with $\epsilon > 0$ fixed but arbitrary.*

3. *Case $n = 3$, $\beta > 0$: Let Assumption* 3.2 *hold with $C_3 = C_5 = 0$. Then there exist global solutions $f, \nabla f \in L^\infty_{\mathrm{loc}}((0,\infty); L^1 \cap L^\infty(\mathbb{R}^3 \times V))$ and $S \in L^\infty_{\mathrm{loc}}((0,\infty); W^{1,p}(\mathbb{R}^3))$ for all $1 \le p \le +\infty$ of the system* (1.6)–(1.9) *with $\epsilon > 0$ fixed but arbitrary.*

4. *Case $n = 3$, $\beta = 0$: Let Assumption* 3.2 *hold with $C_3 = C_5 = 0$. Then there exist global solutions $f, \nabla f \in L^\infty_{\mathrm{loc}}((0,\infty); L^1 \cap L^\infty(\mathbb{R}^3 \times V))$ and $S \in L^\infty_{\mathrm{loc}}((0,\infty); L^p(\mathbb{R}^3))$ for any $3 < p \le \infty$ and $\nabla S \in L^\infty_{\mathrm{loc}}((0,\infty); L^p(\mathbb{R}^3))$ for any $3/2 < p \le \infty$ of the system* (1.6)–(1.9) *with $\epsilon > 0$ fixed but arbitrary.*

Proof. (a) We first consider the case $n = 2$ and $\beta > 0$. Without loss of generality, we assume $\epsilon = 1$. Mass is conserved for ρ, and thus $\|\rho(\cdot,t)\|_{L^1(\mathbb{R}^2)} = \|f_0\|_{L^1(\mathbb{R}^2 \times V)}$.

$$\partial_t f(x,v,t) + v \cdot \nabla_x f(x,v,t) = \int_V T[S](x,v,v',t) f(x,v',t) dv' - \int_V T[S](x,v',v,t) f(x,v,t) dv'.$$

Using Assumption 3.2, we get

$$f(x,v,t) \le f_0(x - vt, v) + C \int_0^t \rho(x - vs, t - s) ds + C f_1(x,v,t) + C f_2(x,v,t),$$

where f_1 and f_2 satisfy

$$\partial_t f_1(x,v,t) + v \cdot \nabla_x f_1(x,v,t) = \int_V [S(x+v,t) + |\nabla S(x+v,t)|] f(x,v',t) dv',$$
$$\partial_t f_2(x,v,t) + v \cdot \nabla_x f_2(x,v,t) = \int_V [S(x-v',t) + |\nabla S(x-v',t)|] f(x,v',t) dv',$$

with initial conditions $f_i(x,v,0) = 0$ for $i = 1, 2$. We first consider f_1. One can easily see that

$$f_1(x,v,t) = \int_0^t [S(x - vs + v, t - s) + |\nabla S(x - vs + v, t - s)|] \rho(x - vs, t - s) ds.$$

After simple calculations, we obtain the following estimates:

$$\|f_1(\cdot,\cdot,t)\|_{L^p(\mathbb{R}^2 \times V)} \le C \sup_{0 \le s \le t} \|S(\cdot,s)\|_{W^{1,p}(\mathbb{R}^2)} \int_0^t \|\rho(\cdot, t - s)\|_{L^p(\mathbb{R}^2)} ds.$$

For the term f_2, we have

$$f_2(x,v,t) = \int_0^t \int_V [S(x - vs - v', t - s) + |\nabla S(x - vs - v', t - s)| f(x - vs, v', t - s) dv' ds.$$

Applying Young's inequality, we get

$$\|(S(\cdot, t - s) + |\nabla S(\cdot, t - s)|) * f(x - vs, \cdot, t - s)\|_{L^\infty(V)} \le \sup_{0 < s < t} \|S(\cdot,s)\|_{W^{1,p}(\mathbb{R}^2)} \|f(x - vs, \cdot, t - s)\|_{L^{p'}(V)},$$

where p and p' are conjugate exponents. If $p \geq 2$, then $p' \leq p$, and so we have, by interpolation between p and 1,

$$\|f(x-vs,\cdot,t-s)\|_{L^{p'}(V)} \leq C(V)\|f(x-vs,\cdot,t-s)\|_{L^p(V)}.$$

Hence,

$$\|f_2(\cdot,\cdot,t)\|_{L^p(\mathbb{R}^2\times V)} \leq \sup_{0<s<t}\|S(\cdot,s)\|_{W^{1,p}(\mathbb{R}^2)}\int_0^t \|f(\cdot,\cdot,t-s)\|_{L^p(\mathbb{R}^2\times V)}ds.$$

Therefore, summing up the estimates above, we obtain for $p \geq 2$

$$\begin{aligned}\|f(\cdot,\cdot,t)\|_{L^p(\mathbb{R}^2\times V)} &\leq \|f_0(\cdot,\cdot)\|_{L^p(\mathbb{R}^2\times V)}\\ &\quad + C\left(1+\sup_{0\leq s\leq t}\|S(\cdot,s)\|_{W^{1,p}(\mathbb{R}^2)}\right)\int_0^t \|f(\cdot,\cdot,s)\|_{L^p(\mathbb{R}^2\times V)}.\end{aligned} \tag{3.9}$$

By Lemma 3.5, we have for $p=2$

$$\begin{aligned}\|f(\cdot,\cdot,t)\|_{L^2(\mathbb{R}^2\times V)} &\leq \|f_0(\cdot,\cdot)\|_{L^2(\mathbb{R}^2\times V)}\\ &\quad + C\left(1+\sup_{0\leq s\leq t}\left[\ln\left(\|f\|^2_{L^2(\mathbb{R}^2\times V)}+1\right)\right]^{1/2}\right)\int_0^t \|f(\cdot,\cdot,s)\|_{L^2(\mathbb{R}^2\times V)}.\end{aligned}$$

Then, applying Gronwall's inequality as in Lemma 3.1, we obtain $f \in L^2(\mathbb{R}^2\times V)$. Now, using bootstrap arguments we obtain the L^∞-estimate by applying repeatedly Lemma 3.4, Young's inequality, and Gronwall's inequality. Next we show L^∞-estimates for the derivatives of f. For convenience let $j=1,2$ be arbitrary but fixed, and we denote by $\tilde{f}$ and $\tilde{T}[S]$ the partial derivatives $\partial_{x_j}f$ and $\partial_{x_j}T[S]$, respectively.

$$\begin{aligned}\partial_t\tilde{f}(x,v,t)+v\cdot\nabla_x\tilde{f}(x,v,t) &= \int_V \tilde{T}[S](x,v,v',t)f(x,v',t)dv'\\ &\quad+\int_V T[S](x,v,v',t)\tilde{f}(x,v',t)dv'\\ &\quad-\int_V \tilde{T}[S](x,v',v,t)f(x,v,t)dv'\\ &\quad-\int_V T[S](x,v',v,t)\tilde{f}(x,v,t)dv'.\end{aligned}$$

Then, in the same manner as before, we obtain

$$\tilde{f}(x,v,t) \leq \tilde{f}_0(x-vt,v)+C\tilde{f}_1(x,v,t)+C\tilde{f}_2(x,v,t)+C\tilde{f}_3(x,v,t)+C\tilde{f}_4(x,v,t),$$

where

$$\begin{aligned}\tilde{f}_1(x,v,t) &= \int_0^t\int_V \tilde{T}[S](x-vs,v,v',t-s)f(x-vs,v',t-s)dv'ds,\\ \tilde{f}_2(x,v,t) &= \int_0^t\int_V T[S](x-vs,v,v',t-s)\tilde{f}(x-vs,v',t-s)dv'ds,\\ \tilde{f}_3(x,v,t) &= -\int_0^t\int_V \tilde{T}[S](x-vs,v',v,t-s)f(x-vs,v,t-s)dv'ds,\\ \tilde{f}_4(x,v,t) &= -\int_0^t\int_V T[S](x-vs,v',v,t-s)\tilde{f}(x-vs,v,t-s)dv'ds.\end{aligned}$$

We consider first $\tilde{f}_1(x,v,t)$. Here we use the fact that the L^∞- and L^p-norms of f, depending on t, are bounded, which was shown above. Therefore we have

$$|\tilde{f}_1(x,v,t)| \le \sup_{0<s<t} \|f(\cdot,s)\|_{L^\infty(\mathbb{R}^2\times V)} \int_0^t \int_V |\tilde{T}[S](x-vs,v,v',t-s)|dv'ds.$$

Using Assumption 3.2, one can easily see

$$\begin{aligned}\|\tilde{f}_1(\cdot,\cdot,t)\|_{L^p(\mathbb{R}^2\times V)} &\le C \sup_{0<s<t} \|f(\cdot,s)\|_{L^\infty(\mathbb{R}^2\times V)} \sup_{0<s<t} \|S(\cdot,s)\|_{W^{2,p}(\mathbb{R}^2)} \\ &\le C \sup_{0<s<t} \|f(\cdot,s)\|_{L^\infty(\mathbb{R}^2\times V)} \sup_{0<s<t} \|\rho(\cdot,s)\|_{L^p(\mathbb{R}^2)} \le C = C(t,|V|),\end{aligned}$$

where we used a standard estimate for the chemo-attractant equation. Since $\tilde{f}_3$ has the same structure as $\tilde{f}_1$, $\tilde{f}_3$ satisfies the estimates above. On the other hand, $\tilde{f}_2$ is estimated, due to Assumption 3.2, as follows:

$$|\tilde{f}_2(x,v,t)| \le \sup_{0<s<t} \|S(\cdot,s)\|_{W^{1,\infty}(\mathbb{R}^2)} \int_0^t \int_V \tilde{f}(x-vs,v',t-s)dv'ds.$$

Again, due to a standard estimate for the chemo-attractant equation, we get

$$|\tilde{f}_2(x,v,t)| \le \sup_{0<s<t} \|\tilde{f}\|_{L^q(\mathbb{R}^2)} \int_0^t \int_V \tilde{f}(x-vs,v',t-s)dv'ds,$$

where q is sufficiently large (i.e., $q>2$). Integration over $\mathbb{R}^2\times V$ yields

$$\|\tilde{f}_2(\cdot,\cdot,t)\|_{L^p(\mathbb{R}^2\times V)} \le C\int_0^t \|\tilde{f}(\cdot,\cdot,t-s)\|_{L^p(\mathbb{R}^2)}ds,$$

where we again used the boundedness of the L^p-norm of f and $C=C(|V|,t)$. $\tilde{f}_4$ can be treated in the same manner, so we omit the details. To sum up, we obtain

$$\|\nabla f(\cdot,\cdot,t)\|_{L^p(\mathbb{R}^2\times V)} \le C(|V|,t) + C(|V|,t)\int_0^t \|\nabla f(\cdot,\cdot,t-s)\|_{L^p(\mathbb{R}^2\times V)}ds.$$

Gronwall's inequality justifies our claim. Repeating this process for higher regularity of f and S, we can easily see that this estimate is valid also in case $p=\infty$. This completes the proof of the case $\beta>0$.

(b) Next we consider the case $n=2$, $\beta=0$. Again, for simplicity, we assume $\epsilon=1$. We first decompose ∇S into two parts,

$$\nabla S = \nabla S^L + \nabla S^S = \rho * \left(-\frac{x}{2\pi|x|^2}\mathbf{I}_{|x|\ge 1}\right) + \rho * \left(-\frac{x}{2\pi|x|^2}\mathbf{I}_{|x|\le 1}\right),$$

where $\mathbf{I_A}$ denotes the characteristic function of a set A. By mass conservation and Young's inequality, we have

$$\|\nabla S^L(t)\|_{L^\infty}(\mathbb{R}^2) \le \frac{1}{2\pi}\|f_0\|_{L^1(\mathbb{R}^2\times V)}.$$

Hence the estimate reduces to considering ∇S^S only, and we may replace ∇S by ∇S^S in the assumption on the turning kernel. Following similar procedures to those described in the case $\beta>0$, we obtain for $p\ge 1$

$$f(x,v,t) \le f_0(x-vt,v) + C\int_0^t \rho(x-vs,t-s)ds + Cf_1(x,v,t),$$

where

$$f_1(x,v,t) = \int_0^t |\nabla S^S(x - vs + v, t - s)|\rho(x - vs, t - s)ds.$$

Simple calculations show

$$\|f_1(\cdot,\cdot,t)\|_{L^p(\mathbb{R}^2\times V)} \le C \sup_{0\le s\le t} \|\nabla S^S(\cdot,s)\|_{L^p(\mathbb{R}^2)} \int_0^t \|\rho(\cdot,t-s)\|_{L^p(\mathbb{R}^2)}ds.$$

To sum up, we obtain

$$\begin{aligned}\|f(\cdot,\cdot,t)\|_{L^p(\mathbb{R}^2\times V)} &\le C + C\left(1 + \sup_{0\le s\le t} \|\nabla S^S(\cdot,s)\|_{L^p(\mathbb{R}^2)}\right)\\ &\quad\times \int_0^t \|f(\cdot,\cdot,t-s)\|_{L^p(\mathbb{R}^2\times V)}ds.\end{aligned} \tag{3.10}$$

Here we note that the above a priori estimate (3.10) holds for all $p \ge 1$. First we choose a specific p with $1 < p < 2$, which ensures, due to Young's inequality, that

$$\|\nabla S^S(\cdot,t)\|_{L^p(\mathbb{R}^2)} \le C\|f_0\|_{L^1(\mathbb{R}^2\times V)}.$$

Then by Gronwall's inequality we get a bound, globally in time, for f in $L^p(\mathbb{R}^2)$ for such chosen p. By bootstrap arguments, we obtain $f \in L^\infty_{\mathrm{loc}}([0,\infty); L^\infty(\mathbb{R}^2 \times V))$.

By similar procedures to those given in the proof of Theorem 3.6, an L^∞-estimate for ∇f can be obtained. $\nabla S \in L^\infty((0,\infty); L^p(\mathbb{R}^2))$, $2 < p \le \infty$, is due to the Hardy–Littlewood–Sobolev theorem (see [20, pp. 119–120]). Since this is also verified by embedding arguments for general elliptic equations, we skip the details.

Remark 3.7. Although similar results, in the theorem above, are expected for nonzero C_2, C_3, C_5 also in case $\beta = 0$, there are some technical difficulties in proving global existence when the chemo-attractant equation is of elliptic type. Indeed, the chemo-attractant equation becomes the Poisson equation without decay term $-\Delta S = \rho$, and thus S has the Newtonian potential representation, i.e., $S = \Gamma * \rho$, where $\Gamma(x) = 1/2\pi \log|x|$. Due to the behavior of Γ at infinity, we cannot, in general, control S in terms of ρ. (We do not have these kind of estimates in Lemma 3.5 if $\beta = 0$.) Thus we leave the global existence as an open question for nonzero C_2, C_3, and C_5 in case $\beta = 0$ and $\tau = 0$.

(c) The three-dimensional case: In this situation, unlike the two-dimensional case in Theorem 3.6, it is not necessary to distinguish proofs for $\beta = 0$ and $\beta \ne 0$. We briefly explain why C_3, C_5 are assumed to be zero in three dimensions. Indeed, as seen in the previous calculations, we end up with the following estimate:

$$\begin{aligned}\|f(\cdot,\cdot,t)\|_{L^p(\mathbb{R}^3\times V)} &\le C + C\left(1 + \sup_{0\le s\le t} \|S^S(\cdot,s)\|_{W^{1,p}(\mathbb{R}^3)}\right)\\ &\quad\times \int_0^t \|f(\cdot,\cdot,t-s)\|_{L^p(\mathbb{R}^3\times V)}ds.\end{aligned} \tag{3.11}$$

On the other hand, in three dimensions, due to behavior of the potential, we have

$$\|S^S(\cdot,s)\|_{W^{1,p}(\mathbb{R}^3)} \le C\|\rho_0\|_{L^1(\mathbb{R}^3)} \quad \text{for } 1 \le p < \frac{3}{2}. \tag{3.12}$$

However, in case C_3 or C_5 are nonzero, one can easily show that estimate (3.11) is still valid provided that $p \geq 2$ (compare the estimates for f_2 and f_4 before), but this does not enable us to use bootstrap arguments to get higher regularity for f because of (3.12). Therefore we assume $C_3 = C_5 = 0$. With this assumption the proof for the case $n = 3$ is similar to the case $n = 2$. ∎

Remark 3.8. It is worth mentioning that Theorem 3.6 also holds in case $n = 3$ when the turning kernel satisfies Assumption 3.2 with $C_2 = C_4 = 0$ instead of $C_3 = C_5 = 0$, namely,

$$0 \leq T[S](x, v, v', t) \leq C\left(1 + S(x - \epsilon v', t) + |\nabla S(x - \epsilon v', t)|\right),$$
$$|\nabla T[S](x, v, v', t)| \leq C(|\nabla S(x - \epsilon v', t)| + |\nabla^2 S(x - \epsilon v', t)|).$$

This can be seen by changing the roles of p and p' in the estimate of f_2 and by following a similar procedure to the one given for the proof of Theorem 3.6.

We do not know if the theorem above is also valid if the turning kernel fulfills the structure conditions (3.2) and (3.3) as in the two-dimensional case.

3.2. Parabolic case: $\tau > 0$. In this part, the parabolic equation for the chemoattractant in (1.8) is considered. From now on we let $\tau = 1$ without loss of generality and, for simplicity, we set here $\alpha = 1$. Then (1.8) for S reads

$$\partial_t S - \Delta S = \rho - \beta S, \quad S(x, 0) = S_0(x), \quad \beta \geq 0. \tag{3.13}$$

To make our arguments simpler, from now on we assume $S_0 = 0$ (compare Remark 3.11 in the following for the case $S_0 \neq 0$).

In the next lemma we recall some basic properties of Γ in two dimensions.

LEMMA 3.9. *Let Γ be the fundamental solution for the operator $\partial_t - \Delta_x + \beta$ in $\mathbb{R}^2$. Then $\Gamma \in L^p(\mathbb{R}^2)$ for any p with $1 \leq p < \infty$, and $\nabla\Gamma \in L^p(\mathbb{R}^2)$ for any q with $1 \leq p < 2$, satisfying*

$$\int_0^t \|\Gamma(\cdot, s)\|_{L^p(\mathbb{R}^2)} ds \leq C(\beta)p, \quad 1 \leq p < \infty,$$
$$\int_0^t \|\nabla\Gamma(\cdot, s)\|_{L^p(\mathbb{R}^2)} ds \leq C(\beta)\frac{2p}{2 - p}, \quad 1 \leq p < 2.$$

Proof. The proof is similar to that of Lemma 3.4, so we omit details. ∎

In the next lemma, we show L^p- and L^2-estimates for S and ∇S, respectively.

LEMMA 3.10. *Let S be a solution of* (3.13) *in $\mathbb{R}^2$ and $S_0 = 0$. Then S satisfies the estimates*

$$\|S(t)\|_{L^p(\mathbb{R}^2)} + \|\nabla S(t)\|_{L^q(\mathbb{R}^2)} \leq C(\beta, p, q)\|\rho_0\|_{L^1(\mathbb{R}^2)}, \tag{3.14}$$

where $1 \leq p < \infty$, $1 \leq q < 2$, and

$$\|\nabla S(t)\|^2_{L^2(\mathbb{R}^2)} \leq C\left(1 + \|\rho_0\|_{L^1(\mathbb{R}^2)}\left(1 + (\ln t)_+ + \sup_{0 \leq \tau \leq t}\left|\ln \|\rho(\tau)\|^2_{L^2(\mathbb{R}^2)}\right|\right)\right), \tag{3.15}$$

where $(f)_+$ indicates the positive part of f.

Proof. By Duhamel's principle and using the fundamental solution Γ in (2.2), we have

$$S(x, t) = \int_0^t \Gamma(\cdot, s) * \rho(\cdot, t - s) ds. \tag{3.16}$$

By using Lemma 3.9, mass conservation, and Young's inequality, we easily get (3.14). To estimate $\|\nabla S\|_{L^2(\mathbb{R}^2)}$, we take the Fourier transform of (3.16) and use Plancherel's equality to get

$$\|\nabla S(t)\|_{L^2(\mathbb{R}^2)} = \|\xi \hat{S}(t)\|_{L^2(\mathbb{R}^2)} \le \int_0^t \||\xi|\hat{\Gamma}(\cdot,s)\hat{\rho}(\cdot,t-s)\|_{L^2(\mathbb{R}^2)}ds = \int_0^r + \cdots + \int_r^t + \cdots,$$

where $r > 0$ will be chosen appropriately later. Note that the Fourier transform of Γ is $\hat{\Gamma}(\xi,s) = \exp(-s(4\xi^2+\beta))$. For $0 < s < r$, due to the Hölder's inequality and Plancherel's equality, we have

$$\begin{aligned}\int_0^r \cdots &\le \int_0^r \||\xi|\exp(-s(4\xi^2+\beta))\|_{L^\infty(\mathbb{R}^2)}\|\hat{\rho}(s)\|_{L^2(\mathbb{R}^2)}ds \\ &\le C \sup_{0\le s\le t}\|\rho\|_{L^2(\mathbb{R}^2)}\int_0^r s^{-1/2}ds \le Cr^{1/2}\sup_{0\le s\le t}\|\rho\|_{L^2(\mathbb{R}^2)}.\end{aligned}$$

For $r < s < t$, due to mass conservation and Hölder's inequality, now applied in the opposite way, we have

$$\begin{aligned}\int_r^t \cdots &\le \int_r^t \||\xi|\exp(-s(4\xi^2+\beta))\|_{L^2(\mathbb{R}^2)}\|\hat{\rho}(s)\|_{L^\infty(\mathbb{R}^2)}ds \\ &\le C\|\rho_0\|_{L^1(\mathbb{R}^2)}\int_r^t \frac{1}{s}ds \le C\|\rho_0\|_{L^1(\mathbb{R}^2)}|\ln t - \ln r|,\end{aligned}$$

where we used $\|\hat{\rho}\|_{L^\infty(\mathbb{R}^2)} \le \|\rho\|_{L^1(\mathbb{R}^2)}$. Therefore we obtain

$$\|\nabla S(t)\|_{L^2(\mathbb{R}^2)} \le C\left(r^{1/2}\sup_{0\le s\le t}\|\rho\|_{L^2(\mathbb{R}^2)} + \|\rho_0\|_{L^1(\mathbb{R}^2)}|\ln t - \ln r|\right).$$

By choosing $r = \min\{(\sup_{0\le s\le t}\|\rho\|_{L^2(\mathbb{R}^2)})^{-2}, t\}$ in the above inequality, we deduce our lemma. □

Remark 3.11. For the case $S_0 \neq 0$, which is assumed to be sufficiently smooth, one has

$$S(x,t) = \int_0^t \Gamma(\cdot,s) * \rho(\cdot,t-s)ds + \int_{\mathbb{R}^2}\Gamma(x-y,t)S_0(y)dy.$$

This gives the following variants of the estimates in the above lemma:

$$\|S(t)\|_{L^p(\mathbb{R}^2)} + \|\nabla S(t)\|_{L^q(\mathbb{R}^2)} \le C\big(\|S_0\|_{L^p(\mathbb{R}^2)} + \|\nabla S_0\|_{L^q(\mathbb{R}^2)} + \|\rho_0\|_{L^1(\mathbb{R}^2)}\big),$$

where $1 \le p < \infty$, $1 \le q < 2$ and

$$\begin{aligned}\|\nabla S(t)\|^2_{L^2(\mathbb{R}^2)} \le C\Bigg(&1 + \|\nabla S_0\|_{L^2(\mathbb{R}^2)} \\ &+ \|\rho_0\|_{L^1(\mathbb{R}^2)}\left(1 + (\ln t)_+ + \sup_{0\le\tau\le t}\left|\ln\left(\|\rho(\tau)\|^2_{L^2(\mathbb{R}^2)}\right)\right|\right)\Bigg).\end{aligned}$$

Since computations are straightforward, we omit the details.

As in the previous elliptic case, we can establish global existence for the system (1.6)–(1.9) with $\tau = 1$. To be more precise, once we have the essential estimate (3.15) for $\|\nabla S\|_{L^2(\mathbb{R}^2)}$, its proof is more or less the same as that for the elliptic case with

$\tau = 0$. Regularity of S is due to standard theory of general parabolic equations. For dimension three, under the weaker assumptions of the turning kernel (Assumption 3.2 with $C_3 = C_5 = 0$) than those in the elliptic case, we can also show global existence of solutions. Since the arguments are straightforward if compared to the elliptic case, we just state the results and skip its proof.

THEOREM 3.12. *Suppose the chemo-attractant equation is of parabolic type. Assume that* $f_0, \nabla f_0 \in (L^1 \cap L^\infty)(\mathbb{R}^n \times V)$.

1. *(Case* $n = 2$*.) Let* $\beta \geq 0$ *and Assumption* 3.2 *hold. Then there exist global solutions* $f, \nabla f \in L^\infty_{\mathrm{loc}}((0,\infty); (L^1 \cap L^\infty)(\mathbb{R}^2 \times V))$ *and* $S, \nabla S \in L^\infty_{\mathrm{loc}}((0,\infty); L^p(\mathbb{R}^2))$ *for all* $1 \leq p \leq +\infty$ *of system* (1.6)–(1.9).

2. *(Case* $n = 3$*.) Let* $\beta \geq 0$ *and Assumption* 3.2 *with* $C_3 = C_5 = 0$. *Then there exist global solutions* $f, \nabla f \in L^\infty_{\mathrm{loc}}((0,\infty); (L^1 \cap L^\infty)(\mathbb{R}^3 \times V))$ *and* $S, \nabla S \in L^\infty_{\mathrm{loc}}((0,\infty); L^p(\mathbb{R}^3))$ *for all* $1 \leq p \leq +\infty$ *of system* (1.6)–(1.9).

4. Diffusion limits of the kinetic model. In this section, the diffusion limit for kinetic models of type (1.6)–(1.9) is presented. First, in a lemma, we review estimates for S which satisfies an equation of elliptic type, i.e.,

$$-\Delta S = \rho - \beta S, \quad \beta \geq 0, \quad \text{in } \mathbb{R}^n, \ n = 2, 3.$$

We use standard arguments, which are known as potential theory. Proofs are straightforward (compare, e.g., [9, Chapters 2 and 8] and [20, Chapter V] for the two-dimensional case, and [16, Chapter 4] and [17, Chapters 4 and 6] for the three-dimensional case).

LEMMA 4.1. *Let* $I = [0, T) \subset \mathbb{R}$ *and* $0 < T < \infty$. *Suppose* $\rho \in L^\infty(I; (W^{1,1}(\mathbb{R}^n) \cap W^{1,q}(\mathbb{R}^n)))$, *where* $q > n$. *Let* S *satisfy the chemo-attractant equation of either elliptic or parabolic type with* $\beta \geq 0$.

(i) *In the case either* $n = 2$, $\beta > 0$ *or* $n = 3$, $\beta \geq 0$, *and* S *fullfils the chemo-attractant equation of either elliptic or parabolic type,*

$$S \in L^\infty(I; W^{2,p}(\mathbb{R}^n)) \cap L^\infty(I; \mathcal{C}^{2+\alpha}(\mathbb{R}^n)), \quad 1 \leq p < \infty, \ 0 < \alpha \leq \frac{q-n}{q},$$

and S *satisfies the estimate*

$$\|S\|_{L^\infty(I; W^{2,p}(\mathbb{R}^n))} + \|S\|_{L^\infty(I; \mathcal{C}^{2+\alpha}(\mathbb{R}^n))} \leq C\big(\|\rho\|_{L^\infty(I; W^{1,1}(\mathbb{R}^n))} + \|\rho\|_{L^\infty(I; W^{1,q}(\mathbb{R}^n))}\big).$$

(ii) *The result of* (i) *is true also for* $n = 2$, $\beta = 0$, *when* S *fullfils the chemo-attractant equation of parabolic type.*

(iii) *In the case* $n = 2$ *and* $\beta = 0$ *and* S *fullfils the chemo-attractant equation of elliptic type,*

$$\nabla S \in L^\infty(I; W^{1,p}(\mathbb{R}^2)) \cap L^\infty(I; \mathcal{C}^{1+\alpha}(\mathbb{R}^2)), \quad 1 \leq p < \infty, \ 0 < \alpha \leq \frac{q-2}{q},$$

and S *satisfies the estimate*

$$\|\nabla S\|_{L^\infty(I; W^{1,p}(\mathbb{R}^2))} + \|\nabla S\|_{L^\infty(I; \mathcal{C}^{1+\alpha}(\mathbb{R}^2))} \leq C\big(\|\rho\|_{L^\infty(I; W^{1,1}(\mathbb{R}^2))} + \|\rho\|_{L^\infty(I; W^{1,q}(\mathbb{R}^2))}\big).$$

As in [3] we need similar assumptions on $\phi^S_\epsilon[S]$ and $\phi^A_\epsilon[S]$, which are the symmetric and antisymmetric parts of $T_\epsilon[S]$ (see Lemma 2.2).

Assumption 4.2. There exist $\gamma > 0$ and a nondecreasing function $\Lambda \in L^\infty_{\mathrm{loc}}$, such that

$$\phi^S_\epsilon[S] \geq \gamma \left(1 - \epsilon\Lambda\left(\|\nabla S\|_{W^{1,\infty}(\mathbb{R}^n)}\right)\right) FF',$$
$$\int_V \frac{\phi^A_\epsilon[S]^2}{F\phi^S_\epsilon[S]} dv' \leq \epsilon^2 \Lambda\left(\|\nabla S\|_{W^{1,\infty}(\mathbb{R}^n)}\right),$$

where $F \in L^\infty(V)$ is a positive velocity distribution satisfying Assumption 2.1.

THEOREM 4.3. *Let Assumptions* 2.1 *and* 4.2 *hold and let $q > n$ with $n = 2, 3$. Suppose that the equation for the chemo-attractant S is either of elliptic ($\tau = 0$) or parabolic type ($\tau \neq 0$). Let one of following conditions hold:*

(i) *If $\tau = 0$, $n = 2$, $\beta > 0$ or if $\tau > 0$, $n = 2$, $\beta \geq 0$, the turning kernel satisfies Assumption* 3.2.

(ii) *If $\tau = 0$, $n = 2$, $\beta = 0$, the turning kernel satisfies Assumption* 3.2 *with $C_2 = C_3 = C_5 = 0$.*

(iii) *If $\tau \geq 0$, $n = 3$, $\beta \geq 0$, the turning kernel satisfies Assumption* 3.2 *with $C_3 = C_5 = 0$.*

Assume further that

$$f_0 \in \Upsilon_q \equiv W^{1,1}(\mathbb{R}^n \times V) \cap W^{1,q}\left(\mathbb{R}^n \times V; \frac{dxdv}{F^{q-1}}\right).$$

Then there exists $t^* > 0$, independent of ϵ, such that the solutions f_ϵ, S_ϵ satisfy

$$f_\epsilon \in L^\infty((0,t^*); \Upsilon_q),$$
$$\nabla S_\epsilon \in L^\infty((0,t^*); W^{1,p}(\mathbb{R}^n) \cap \mathcal{C}^{1+\alpha}(\mathbb{R}^n)), \quad 1 \leq p < \infty,\ \alpha = \frac{q-2}{q}$$
$$\text{if } \tau = 0,\ n = 2,\ \beta = 0.$$
$$S_\epsilon \in L^\infty((0,t^*); W^{2,p}(\mathbb{R}^n) \cap \mathcal{C}^{2+\alpha}(\mathbb{R}^n)), \quad 1 \leq p < \infty,\ \alpha = \frac{q-n}{q} \text{ in all other cases,}$$
$$r_\epsilon = \frac{f_\epsilon - \rho_\epsilon F}{\epsilon} \in L^2\left((0,t^*); \mathbb{R}^n \times V : \frac{dxdvdt}{F}\right). \tag{4.1}$$

Proof. This can be shown by following the same procedure as that given in the proof of Theorem 4 in [3], and therefore we present only a brief sketch of this proof. Simple calculations show

$$\frac{d}{dt}\int_{\mathbb{R}^n}\int_V \frac{f^q_\epsilon}{F^{q-1}} dvdx \leq C\Lambda\big(\|\nabla S\|_{W^{1,\infty}(\mathbb{R}^n)}\big)\int_{\mathbb{R}^n}\int_V \frac{f^q_\epsilon}{F^{q-1}} dvdx.$$

The next step is to estimate S_ϵ:

$$\|\nabla S_\epsilon(\cdot,t)\|_{\mathcal{C}^{1,\alpha}(\mathbb{R}^n)} \leq C\big(1 + \|\nabla\rho_\epsilon(\cdot,t)\|_{L^q(\mathbb{R}^n)}\big) \leq \tilde{C}\big(1 + \|\rho_\epsilon(\cdot,t)\|_{L^q(\mathbb{R}^n)}\big).$$

Here we used the estimates in Lemma 4.1.

$$\frac{d}{dt}\int_{\mathbb{R}^n}\int_V \frac{f^q_\epsilon}{F^{q-1}} dvdx \leq C\left[1 + \left(\int_{\mathbb{R}^n}\int_V \frac{f^q_\epsilon}{F^{q-1}} dvdx\right)^{\frac{1}{q}}\right]\int_{\mathbb{R}^n}\int_V \frac{f^q_\epsilon}{F^{q-1}} dvdx.$$

This shows the first two statements. The rest can be done by using the same method as that given in the proof of Theorem 4 in [3], and thus we omit the details. ∎

Now we are ready to prove the existence of the diffusion limit in a short time interval.

THEOREM 4.4. *Let the assumption of Theorem* 4.3 *hold. Suppose that the equation for the chemo-attractant S is of elliptic ($\tau = 0$) or parabolic ($\tau \neq 0$) type. Assume further that for families (S_ϵ), which are uniformly bounded in $L^\infty_{\mathrm{loc}}([0,\infty);\mathcal{C}^{2+\alpha}(\mathbb{R}^n))$ for some α with $0 < \alpha \leq 1$, such that S_ϵ, ∇S_ϵ, and $\nabla^2 S_\epsilon$ converge to $S_0, \nabla S_0$, and $\nabla^2 S_0$ as $\epsilon \to 0$, respectively, in $L^p_{\mathrm{loc}}([0,\infty);\mathbb{R}^n)$ for some $p > n/(n-1)$ with $n = 2, 3$, we have the convergence*

$$T_\epsilon[S_\epsilon] \to T_0[S_0] \quad \text{in } L^p_{\mathrm{loc}}([0,\infty);\mathbb{R}^n \times \bar{V} \times \bar{V}),$$
$$\frac{\mathcal{T}_\epsilon[S_\epsilon](F)}{\epsilon} = \frac{2}{\epsilon}\int_V \phi^A_\epsilon[S_\epsilon]dv' \to \mathcal{T}_1[S_0](F) \quad \text{in } L^p_{\mathrm{loc}}([0,\infty);\mathbb{R}^n \times \bar{V}). \tag{4.2}$$

Then the solutions f_ϵ and S_ϵ of (1.6)–(1.9) *satisfy*

$$f_\epsilon \to \rho_0 F \quad \text{in } L^\infty((0,t^*);\Upsilon_q) \text{ weak } *,$$

and for $\tau = 0$

$$\nabla S_\epsilon \to \nabla S_0 \quad \text{in } W^{1,q}_{\mathrm{loc}}((0,t^*);\mathbb{R}^n), \quad 1 \leq q < \infty \text{ if } n = 2,\ \beta = 0,$$
$$S_\epsilon \to S_0 \quad \text{in } W^{2,q}_{\mathrm{loc}}((0,t^*);\mathbb{R}^n), \quad 1 \leq q < \infty \text{ otherwise},$$

whereas for $\tau \neq 0$

$$S_\epsilon \to S_0 \quad \text{in } L^q_{\mathrm{loc}}((0,t^*);W^{2,q}(\mathbb{R}^n)), \quad 1 \leq q < \infty.$$

Proof. Since the proof is similar to that of Theorem 5 in [3], we again present only a brief sketch of the procedure. First we note, due to (4.1), that

$$J_\epsilon = \frac{1}{\epsilon}\int_V v f_\epsilon dv = \int_V v r_\epsilon dv \in L^2((0,t^*);L^2(\mathbb{R}^n))$$

uniformly in ϵ. From the cell conservation equation $\partial_t \rho_\epsilon + \operatorname{div} \mathrm{J}_\epsilon = 0$, one can easily see that

$$\partial_t(\nabla S_\epsilon) \in L^2((0,t^*);L^2_{\mathrm{loc}}(\mathbb{R}^n))$$

by considering the gradient of the convolution of (1.8). The strong convergence follows combining the above estimate and the parabolic regularity for the convolutions defining S_ϵ and ∇S_ϵ from ρ_ϵ. Therefore, the kinetic equation (1.7) leads to

$$\epsilon\frac{\partial f_\epsilon}{\partial t} + v \cdot \nabla_x f_\epsilon = -\rho_\epsilon \frac{\mathcal{T}[S_\epsilon](F)}{\epsilon} - \mathcal{T}_\epsilon[S_\epsilon](r_\epsilon).$$

By assumption (4.2) and passing to the limit, we obtain

$$\mathcal{T}_0[S_0](r_0) = -vF \cdot \nabla \rho_0 - \rho_0 \mathcal{T}_1[S_0](F).$$

This equation can be solved due to Lemma 2.3. The limit of the cell conservation equation is $\partial_t \rho_0 + \nabla \cdot J_0 = 0$ with $J_0 = \int_V v r_0 dv$. This completes the proof. ∎

5. Examples. When dealing with chemosensitive movement of biological species, questions of major interest are, How do the individuals "measure" the chemical signal? How is this information processed, and what kind of behavior results? The model we have introduced before and its macroscopic limit give a partial answer to this problem.

First we give a short summary of possible evaluations of the chemical signal by the cells as suggested by Tranquillo and Alt [22] and later discuss related examples. The individuals might evaluate the chemical signal

spatially - the signal is evaluated at (at least) two distinct locations around the individual, which are related to its direction (cf. Examples 5.1, 5.3, and 5.4);

temporal(ly) differential - the signal is evaluated at (at least) two different times (cf. Example 5.1);

positionally - the signal is evaluated momentarily (cf. Example 5.4);

directionally - the signal is evaluated along the individual direction or its relation to a directional signal field, e.g., a spatial gradient at its position (cf. Examples 5.1, 5.3, 5.4, 5.5, and 5.6).

Discussions of possible turning rates of the cells which depend on the given chemical signal in this context are also given in [1], [2], [18] and [10], [19].

In [10], [19] the macroscopic limit is formal. It is assumed that the turning kernel has an expansion in ϵ which is supposed to be given. Here the ϵ-expansion is directly related to possible evaluations of the chemo-attractant by the cells, and thus the connection between the micro- and macroparameters can be derived.

To understand the different influences of the evaluations of the chemical signal, our first example is very general and allows also dependencies on time derivatives of the chemo-attractant. Since we did not prove regularity for S_t so far, the macroscopic limit in this case has to be considered only formal. Nevertheless, from this example the other rigorous examples can be extracted later. Below we only consider the two-dimensional case, to keep the computations simple and since this case is the most interesting one biologically.

Example 5.1 (formal for $\alpha > 0$, rigorous for $\alpha = 0$). Let the turning kernel be of general type:

$$\begin{aligned} T_\epsilon[s] = \phi(&S(x+\epsilon v,t), S(x-\epsilon v',t), S(x,t-\epsilon), \nabla S(x+\epsilon v,t), \nabla S(x-\epsilon v',t),\\ &\partial_t S(x+\epsilon v,t), \partial_t S(x-\epsilon v',t), \partial_t S(x,t-\epsilon), v) + \epsilon\psi\left(\frac{v\cdot v'}{|v||v'|}\right), \end{aligned} \tag{5.1}$$

where $\phi : \mathbb{R}^{12} \to \mathbb{R}$ and $\psi : \mathbb{R} \to \mathbb{R}$ are smooth and $\phi + \epsilon\psi$ is strictly positive (∇S contributes two entries, $\partial_{x_1} S$ and $\partial_{x_2} S$). Here S satisfies the chemo-attractant equation either of elliptic type or of parabolic type with $\alpha \geq 0$ and $\beta > 0$ in two dimensions. For $\alpha = 0$ the S-equation is completely decoupled. In this case the derivation given below is rigorous. We do not include direct dependencies such as $S(x,t), S_t(x,t), \nabla S(x,t)$ at this point. These will be discussed later.

We use the following notational abbreviations:

$$\begin{aligned} \phi[S, \nabla S, \partial_t S, v] &:= \phi(S(x,t), S(x,t), S(x,t), \nabla S(x,t),\\ &\qquad \nabla S(x,t), \partial_t S(x,t), \partial_t S(x,t), \partial_t S(x,t), v),\\ \phi_i[S, \nabla S, \partial_t S, v] &:= \phi_i(S(x,t), S(x,t), S(x,t), \nabla S(x,t), \nabla S(x,t),\\ &\qquad \partial_t S(x,t), \partial_t S(x,t), \partial_t S(x,t), v), \end{aligned}$$

where $\phi_i(\cdots)$ indicates the partial derivative of ϕ with respect to the ith argument for $i = 1, 2, \ldots, 12$. By the asymptotic expansion of $T_\epsilon = T_0 + \epsilon T_1 + O(\epsilon^2)$, one can easily see that $T_0 = T_0[S, v] = \phi[S, \nabla S, \partial_t S, v]$ and

$$\begin{aligned} T_1 &= T_1[S, v, v'] \\ &= (\phi_1[S, \nabla S, \partial_t S, v]v - \phi_2[S, \nabla S, \partial_t S, v]v') \cdot \nabla S + \phi_3[S, \nabla S, \partial_t S, v]\partial_t S \\ &\quad + (\phi_{3+i}[S, \nabla S, \partial_t S, v]v - \phi_{5+i}[S, \nabla S, \partial_t S, v]v') \cdot \nabla S_{x_i} \\ &\quad + (\phi_8[S, \nabla S, \partial_t S, v]v - \phi_9[S, \nabla S, \partial_t S, v]v') \cdot \nabla S_t \\ &\quad - \phi_{10}[S, \nabla S, \partial_t S, v]\partial_t^2 S + \psi\left(\frac{v \cdot v'}{|v||v'|}\right), \end{aligned}$$

where we used the summation convention, which is understood over repeated indices running from 1 to 2. Furthermore, we define Φ, $\tilde{\Phi}$, $\hat{\Phi}$, and $\bar{\Phi}$ as follows:

$$\Phi[S_0, \nabla S_0, \partial_t S_0] := \int_V T_0[S_0, v']dv', \qquad \tilde{\Phi}[S_0, \nabla S_0, \partial_t S_0, v] := \int_V T_1[S_0, v', v]dv',$$

$$\hat{\Phi}[S_0, \nabla S_0, \partial_t S_0, v] := \int_V T_1[S_0, v, v']f_0(v', x, t)dv',$$

$$\bar{\Phi}[S_0, \nabla S_0, \partial_t S_0, v] := \frac{1}{\Phi[S_0, \nabla S_0, \partial_t S_0]} \int_V T_0[S_0, v']T_1[S_0, v, v']dv'.$$

From $\mathcal{T}_0[S_0](f_0) = 0$, we have

$$f_0(v, x, t) = \frac{\phi[S_0, \nabla S_0, \partial_t S_0, v]\rho_0(x, t)}{\Phi[S_0, \nabla S_0, \partial_t S_0]},$$

and therefore it is easy to see $\hat{\Phi}(v) = \bar{\Phi}(v)\rho_0$.

Due to $\mathcal{T}_0[S_0](f_1) = -\mathcal{T}_1[S_0](f_0) - v \cdot \nabla f_0$, we have

$$\begin{aligned} f_1(v, x, t) = \frac{1}{\Phi[S_0, \nabla S_0, \partial_t S_0]}(&- v \cdot \nabla f_0(v, x, t) - \tilde{\Phi}[S_0, \nabla S_0, \partial_t S_0, v]f_0(v, x, t) \\ &+ \hat{\Phi}[S_0, \nabla S_0, \partial_t S_0, v]). \end{aligned}$$

Computing $J_\epsilon = \int_V v f_1(v, x, t)dv$, we obtain

$$\begin{aligned} J_\epsilon = &- \int_V \frac{v^i v^j \partial_{x_j} f_0}{\Phi[S_0, \nabla S_0, \partial_t S_0]}dv - \int_V \frac{v^i \tilde{\Phi}[S_0, \nabla S_0, \partial_t S_0, v]f_0}{\Phi[S_0, \nabla S_0, \partial_t S_0]}dv \\ &+ \int_V \frac{v^i \hat{\Phi}[S_0, \nabla S_0, \partial_t S_0, v]\rho_0}{\Phi[S_0, \nabla S_0, \partial_t S_0]}dv. \end{aligned} \tag{5.2}$$

The first integral in (5.2) becomes

$$\begin{aligned} \int_V \frac{v^i v^j \partial_{x_j} f_0}{\Phi[S_0, \nabla S_0, \partial_t S_0]}dv &= \frac{\rho_0}{\Phi[S_0, \nabla S_0, \partial_t S_0]} \int_V \left(v^i v^j \partial_{x_j}\left(\frac{\phi[S_0, \nabla S_0, \partial_t S_0, v]}{\Phi[S_0, \nabla S_0, \partial_t S_0]}\right)\right)dv \\ &\quad + \frac{\partial_{x_j}\rho_0}{\Phi^2[S_0, \nabla S_0, \partial_t S_0]} \int_V v^i v^j \phi[S_0, \nabla S_0, \partial_t S_0, v]dv \\ &= \frac{A_i}{\Phi}\rho_0 + \frac{B_{ij}}{\Phi^2}\partial_{x_j}\rho_0, \end{aligned}$$

where

$$A_i = A_i[S_0, \nabla S_0, \partial_t S_0] = \int_V v^i v^j \partial_{x_j} \left(\frac{\phi[S_0, \nabla S_0, \partial_t S_0, v]}{\Phi[S_0, \nabla S_0, \partial_t S_0]} \right) dv, \tag{5.3}$$

$$B_{ij} = B_{ij}[S_0, \nabla S_0, \partial_t S_0] = \int_V v^i v^j \phi[S_0, \nabla S_0, \partial_t S_0, v] dv. \tag{5.4}$$

The second integral in (5.2) leads to

$$\int_V \frac{v^i \tilde{\Phi}[S_0, \nabla S_0, \partial_t S_0, v] f_0(v)}{\Phi[S_0, \nabla S_0, \partial_t S_0]} dv = \frac{C_i}{\Phi^2(S_0, \nabla S_0, \partial_t S_0)} \rho_0,$$

where

$$C_i = C_i[S_0, \nabla S_0, \partial_t S_0] = \int_V v^i \tilde{\Phi}[S_0, \nabla S_0, \partial_t S_0, v] \phi[S_0, \nabla S_0, \partial_t S_0, v] dv. \tag{5.5}$$

The last integral in (5.2) becomes

$$\int_V \frac{v^i \hat{\Phi}[S_0, \nabla S_0, \partial_t S_0, v]}{\Phi[S_0, \nabla S_0, \partial_t S_0]} dv = \int_V \frac{v^i \bar{\Phi}[S_0, \nabla S_0, \partial_t S_0, v] \rho_0}{\Phi[S_0, \nabla S_0, \partial_t S_0]} dv = \frac{D_i}{\Phi} \rho_0,$$

where

$$D_i = D_i[S_0, \nabla S_0, \partial_t S_0] = \int_V v^i \bar{\Phi}[S_0, \nabla S_0, \partial_t S_0, v] dv. \tag{5.6}$$

Summing up, we obtain the macroscopic equation

$$\partial_t \rho_0 = \partial_{x_i} \left(\frac{A_i}{\Phi} \rho_0 + \frac{B_{ij}}{\Phi^2} \partial_{x_j} \rho_0 + \frac{C_i}{\Phi^2} \rho_0 - \frac{D_i}{\Phi} \rho_0 \right), \quad \Phi = \Phi[S_0, \nabla S_0, \partial_t S_0],$$

where A_i, B_{ij}, C_i, and D_i are defined in (5.3)–(5.6).

Remark 5.2. If we drop out the explicit dependence of the last argument v in the functional ϕ in (5.1), then the term $\psi(v \cdot v'/|v||v'|)$ does not influence the resulting macroscopic equation anymore. This is due to the fact that only C_i and D_i depend on ψ (A_i, B_i do not), and $C_i = D_i = 0$ when ϕ is independent of v. This is to be expected from a biological point of view since reorientations without any bias cannot have a macroscopic effect.

In the following we will see how to evaluate A_i, B_{ij}, C_i, and D_i more specifically.

Example 5.3 (rigorous for $\alpha \geq 0$). Let

$$T_\epsilon[S] = \phi(S(x + \epsilon v, t), S(x - \epsilon v', t), \nabla S(x + \epsilon v, t), \nabla S(x - \epsilon v', t)), \tag{5.7}$$

where S satisfies chemo-attractant equation of elliptic type with $\beta > 0$ in two dimensions. Note that $\phi : \mathbb{R}^2 \times \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ is an even function with respect to the variable ∇S, and increasing and decreasing for the first and second argument, respectively. Also assume the structure condition of Assumptions 2.1 and 3.2, i.e.,

$$|T_\epsilon[S](x, v, v', t)| \leq C(1 + S(x + \epsilon v, t) + S(x - \epsilon v', t) + |\nabla S(x + \epsilon v, t)| + |\nabla S(x - \epsilon v', t)|).$$

Using the asymptotic expansion of the turning kernel, i.e., $T_\epsilon[S] = T_0[S] + \epsilon T_1[S] + O(\epsilon^2)$, we can easily see that $T_0[S] = \phi(S(x,t), S(x,t), \nabla S(x,t), \nabla S(x,t))$, and

$$\begin{aligned} T_1[S] = {} & (\phi_1(S, S, \nabla S, \nabla S) v - \phi_2(S, S, \nabla S, \nabla S) v') \cdot \nabla S \\ & + \sum_{i=1}^{2} (\phi_{2+i}(S, S, \nabla S, \nabla S) v - \phi_{4+i}(S, S, \nabla S, \nabla S) v') \cdot \nabla S_{x_i}. \end{aligned}$$

Here ϕ_k, $k = 1, 2, \ldots, 6$, indicates differentiation of ϕ with respect to the kth argument. The symmetric $\phi_\epsilon^A[S]$ and antisymmetric part $\phi_\epsilon^S[S]$ of the turning kernel satisfy

$$\phi_\epsilon^S[S] \geq \gamma\big(1 - \epsilon\Lambda\big(\|\nabla S\|_{W^{1,\infty}(\mathbb{R}^n)}\big)\big)FF', \quad \int_V \frac{\phi_\epsilon^A[S]^2}{F\phi_\epsilon^S[S]}dv' \leq \epsilon^2\Lambda\left(\|\nabla S\|_{W^{1,\infty}(\mathbb{R}^n)}\right),$$

where $\gamma > 0$ and $\Lambda \in L^\infty_{\mathrm{loc}}$ is a nondecreasing function. By asymptotic expansion of f_ϵ and S_ϵ, the leading order equation becomes $f_0(x, v, t) = \rho_0(x, t)/|V|$. Here f_0 is independent of v. Since the ϵ-order equation is

$$\mathcal{T}_0[S_0](f_1) = -(v \cdot \nabla\rho_0)/|V| - \mathcal{T}_1[S_0](f_0),$$

we have to calculate

$$\mathcal{T}_1[S_0](f_0) = -\rho_0(\phi_1 + \phi_2)\nabla S_0 \cdot v - \sum_{i=1}^{2} \rho_0(\phi_{2+i} + \phi_{4+i})\nabla S_{0,x_i} \cdot v.$$

Therefore,

$$\mathcal{T}_0[S_0](f_1) = -\frac{v \cdot \nabla\rho_0}{|V|} + \rho_0(\phi_1 + \phi_2)\nabla S_0 \cdot v + \sum_{i=1}^{2} \rho_0(\phi_{2+i} + \phi_{4+i})\nabla S_{0,x_i} \cdot v,$$

due to the solvability condition, and thus we get

$$f_1 = -\frac{v \cdot \nabla\rho_0}{|V|^2\phi} + \frac{\rho_0(\phi_1 + \phi_2)\nabla S_0 \cdot v}{|V|\phi} + \frac{\rho_0(\phi_{2+i} + \phi_{4+i})\nabla S_{0,x_i} \cdot v}{|V|\phi}.$$

Let $\mu = \int_V |v|^2 dv$. Using the above results, we obtain the flux density $J_\epsilon = \int_V v f_1 dv + O(\epsilon)$, where

$$J_\epsilon = -\frac{\mu}{2|V|^2}\frac{\nabla\rho_0}{\phi} + \frac{\mu}{2|V|}\frac{(\phi_1 + \phi_2)\rho_0\nabla S_0}{\phi} + \sum_{i=0}^{2}\frac{\mu}{2|V|}\frac{(\phi_{2+i} + \phi_{4+i})\rho_0\nabla S_{0,x_i}}{\phi}.$$

Hence the diffusion limit is

$$\frac{\partial}{\partial t}\rho_0 = \nabla \cdot \left(D\nabla\rho_0 - \chi\rho_0\nabla S_0 - \sum_{i=1}^{2}\tilde{\chi}_i\rho_0\nabla S_{0,x_i}\right) \tag{5.8}$$

with

$$D = \frac{\mu}{2|V|^2\phi}, \quad \chi = \frac{\mu(\phi_1 + \phi_2)}{2|V|\phi}, \quad \tilde{\chi}_i = \frac{\mu(\phi_{2+i} + \phi_{4+i})}{2|V|\phi}, \quad i = 1, 2,$$

coupled to $-\Delta S_0 = \rho_0 - \beta S_0$. It is not known whether solutions for the macroscopic equation (5.8) blow up in finite time or not.

Example 5.4. If we choose an appropriate turning kernel, then the classical Keller–Segel model with constant coefficients can also be obtained. Indeed, if the turning kernel (5.7) is replaced by $T_\epsilon[s] = \phi(S(x,t), S(x+\epsilon v, t), \nabla S(x+\epsilon v, t), \nabla S(x-\epsilon v', t))$, then, by following similar computations to those given above, we have

$$\frac{\partial}{\partial t}\rho_0 = \nabla \cdot \left(\frac{\mu}{2|V|^2\phi}\nabla\rho_0 - \frac{\mu\phi_2}{2|V|\phi}\rho_0\nabla S_0 - \sum_{i=1}^{2}\frac{\mu(\phi_{2+i} + \phi_{4+i})}{2|V|\phi}\rho_0\nabla S_{0,x_i}\right). \tag{5.9}$$

Now let

$$\phi(x_1,x_2,x_3,x_4,x_5,x_6)=\varphi(x_2-x_1)+\varphi(x_5-x_3)+\varphi(x_6-x_4), \tag{5.10}$$

where $\varphi(x)=C_1\sqrt{1+x^2}+C_2x$, $C_1>C_2>0$.

Concerning the gradient terms, this example seems a bit artificial, but it shows how higher order terms might cancel out. Since $\varphi(0)=C_1$, $\varphi'(0)=C_2$, we have $\phi=C_1$, $\phi_2=C_2$, $\phi_3=\phi_4=-C_2$, and $\phi_5=\phi_6=C_2$. Therefore (5.9) leads to

$$\frac{\partial}{\partial t}\rho_0=\nabla\cdot\left(\frac{\mu}{2|V|^2C_1}\nabla\rho_0-\frac{\mu C_2}{2|V|C_1}\rho_0\nabla S_0\right),$$

which is the classical version of the Keller–Segel model. The diffusion coefficient and chemotactic sensitivity, respectively, are $D=\mu/(2|V|^2C_1)$, $\chi=(\mu C_2)/(2|V|C_1)$, which are both constants in this case.

Example 5.5 (rigorous, $\alpha\geq 0$, $\beta>0$). The next example considers time variations of the chemical S.

$$T_\epsilon=\sigma S(x+\epsilon v,t)+h(\partial_t S(x,t),\nabla S(x,t),v)+C_2, \tag{5.11}$$

where $\sigma\geq 0$ is a fixed constant and $h:\mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}$, $n=2,3$, is smooth and bounded, say $-C_1\leq h\leq C_1$ with $0<C_1<C_2$. Note that the turning kernel satisfies the structure condition in Assumption 3.2. Skipping the details of the calculations, *the macroscopic equation reads*

$$\begin{aligned}\partial_t\rho_0=\nabla\cdot\Bigg(&\frac{1}{\sigma S_0|V|+H[S_0]}\left[\nabla\left(\frac{\mu(\sigma S_0+C_2)}{\sigma S_0|V|+H[S_0]}\rho_0\right)+(A_{ij}[S_0]\rho_0)_{x_j}\right]\\&-\frac{\sigma\mu}{\sigma S_0|V|+H[S_0]}\rho_0\nabla S_0\Bigg).\end{aligned} \tag{5.12}$$

This equation is rigorously derived with related turning kernel (5.11) since it satisfies Assumption 4.2.

As a specific example, we consider the case

$$h(\partial_t S,\nabla S,v)=C_1\frac{\gamma\partial_t S+v\cdot\nabla S}{\mathcal{N}(S)},\quad \mathcal{N}(S)=\sqrt{1+\gamma^2|\partial_t S|^2+|\nabla S|^2},$$

where γ is a fixed constant. Then one can easily see

$$H[S_0]=\frac{C_1\gamma\partial_t S_0|V|}{\mathcal{N}(S_0)}+C_2|V|,\qquad A_{ij}[S_0]=\frac{C_1\mu\gamma\partial_t S_0}{(\sigma S_0|B_1|+H[S_0])\mathcal{N}(S_0)}.$$

Therefore, *the macroscopic equation* (5.12) can be explicitly calculated, namely, *for* $\gamma=0$ ($H[S_0]=C_2|V|$ and $A_{ij}[S_0]=0$),

$$\partial_t\rho_0=\nabla\cdot\left(\frac{\mu}{(\sigma S_0+C_2)|V|^2}\nabla\rho_0-\frac{\sigma\mu}{(\sigma S_0+C_2)|V|}\rho_0\nabla S_0\right). \tag{5.13}$$

On the other hand, if $\sigma=0$, then the last term in (5.12) vanishes and (5.12) reads

$$\partial_t\rho_0=\nabla\cdot\left(\frac{1}{H[S_0]}\nabla\left(\frac{\mu C_2}{H[S_0]}\rho_0\right)+(A_{ij}[S_0]\rho_0)_{x_j}\right),$$

where $A_{ij}[S_0] = \int_V v^i v^j h(\partial_t S, \nabla S, v)/H[S_0]dv$. In case h is odd with respect to v, then $A_{ij} = 0$ in (5.12).

In the next example we discuss the influence of nonlocal terms in h.

Example 5.6 (formal for $\alpha > 0$, rigorous for $\alpha = 0$). Consider $T_\epsilon = \sigma S(x + \epsilon v, t) + h(\partial_t S(x+\epsilon v, t), v\cdot\nabla S(x+\epsilon v, t)) + C_2$, where $h : \mathbb{R}\times\mathbb{R}\to\mathbb{R}$, $n = 2, 3$, is smooth and bounded, say $-C_1 \le h \le C_1$ with $0 < C_1 < C_2$. The structure condition in Assumption 3.2 is satisfied.

Again, skipping the detailed calculations, *the macroscopic equation reads*

$$\partial_t\rho_0 = -\nabla\cdot J_\epsilon = \nabla\cdot\left(\frac{1}{\sigma S_0|V| + H[S_0]}\left(\int_V v_i v_j \partial_{x_j} f_0 dv + K[S_0]\int_V v_i f_0 dv - \rho_0\int_V v_i T_1[S_0, v]dv\right)\right).$$

Next we consider a specific example of the turning kernel above. Let

$$h = h(\partial_t S(x+\epsilon v, t-\epsilon), v\cdot\nabla S(x+\epsilon v, t)) = \frac{C_1 v\cdot\nabla S(x+\epsilon v, t)}{\sqrt{1 + (v\cdot\nabla S(x+\epsilon v, t))^2}}.$$

Therefore, *the macroscopic equation reads*

$$\partial_t\rho_0 = \nabla\cdot\left(\frac{\mu}{(\sigma S_0 + C)|V|^2}\nabla\rho_0 - \frac{\sigma\mu}{(\sigma S_0 + C)|V|}\rho_0\nabla S_0 + \frac{L[S_0](L[S_0]\Delta S_0 - M[S_0]|\nabla S_0|^2\Delta S_0)}{(\sigma S_0 + C)^2|V|^2}\rho_0\nabla S_0\right),$$

where

$$L[S_0] = \frac{1}{n}\int_V \frac{|v|^2}{\sqrt{1 + (v\cdot\nabla S_0)^2}}dv, \qquad M[S_0] = \frac{1}{n^2}\int_V \frac{|v|^4}{(1 + (v\cdot\nabla S_0)^2)^{\frac{3}{2}}}dv. \tag{5.14}$$

The third term in the macroscopic equation is completely due to the nonlocal dependencies of h. Compare (5.13) for the local formulation.

Acknowledgment. We would like to thank B. Perthame for pointing out Remark 3.8 during his visit at MPI MIS in Leipzig.

REFERENCES

[1] W. Alt, *Biased random walk models for chemotaxis and related diffusion approximations*, J. Math. Biol., 9 (1980), pp. 147–177.

[2] W. Alt, *Singular perturbation of differential integral equations describing biased random walks*, J. Reine Angew. Math., 322 (1981), pp. 15–41.

[3] F. A. C. C. Chalub, P. Markowich, B. Perthame, and C. Schmeiser, *Kinetic models for chemotaxis and their drift-diffusion limits*, Monatsh. Math., 142 (2004), pp. 123–141.

[4] F. A. C. C. Chalub, P. Markowich, B. Perthame, and C. Schmeiser, *On the Derivation of Drift-Diffusion Model for Chemotaxis from Kinetic Equations*, ANUM preprint 14/02, Vienna Technical University, 2002.

[5] F. A. C. C. Chalub, P. Markowich, B. Perthame, and C. Schmeiser, *Global Existence and Macroscopic Limits for Kinetic Models of Chemotaxis*, ANUM preprint 16/02, Vienna Technical University, 2002.

[6] Y. Dolak and C. Schmeiser, *Kinetic Models for Chemotaxis: Hydrodynamic Limits and the Back-of-the-Wave Problem*, ANUM preprint 5/03, Vienna Technical University, 2003.

[7] L. C. Evans, *Partial Differential Equations*, AMS, Providence, RI, 1998.

[8] G. B. Folland, *Real Analysis. Modern Techniques and Their Applications*, Pure Appl. Math. (N.Y.), John Wiley & Sons, New York, 1984.
[9] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Grundlehren Math. Wiss. 224, Springer-Verlag, Berlin, 1983.
[10] T. Hillen and H. G. Othmer, *The diffusion limit of transport equations derived from velocity-jump processes*, SIAM J. Appl. Math., 61 (2000), pp. 751–775.
[11] D. Horstmann, *From* 1970 *until present: The Keller-Segel model in chemotaxis and its consequences*, Jahresber. Deutsch. Math.-Verein., 105 (2003), pp. 103–165.
[12] H. Hwang, K. Kang, and A. Stevens, *Drift-diffusion limits of kinetic models for chemotaxis: A generalization*, Discrete Contin. Dyn. Syst. Ser. B., to appear.
[13] H. Hwang, K. Kang, and A. Stevens, *Global existence of classical solutions for a hyperbolic chemotaxis model and its parabolic limit*, Indiana Univ. Math. J., submitted.
[14] E. F. Keller and L. A. Segel, *Initiation of slide mold aggregation viewed as an instability*, J. Theoret. Biol., 26 (1970), pp. 399–415.
[15] E. F. Keller and L. A. Segel, *Model for chemotaxis*, J. Theoret. Biol., 30 (1971), pp. 225–234.
[16] O. A. Ladyženskaja, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monogr. 23, AMS, Providence, RI, 1967.
[17] G. M. Lieberman, *Second Order Parabolic Differential Equations*, World Scientific, River Edge, NJ, 1996.
[18] H. G. Othmer, S. R. Dunbar, and W. Alt, *Models of dispersal in biological systems*, J. Math. Biol., 26 (1988), pp. 263–298.
[19] H. G. Othmer and T. Hillen, *The diffusion limit of transport equations* II: *Chemotaxis equations*, SIAM J. Appl. Math., 62 (2002), pp. 1222–1250.
[20] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.
[21] A. Stevens, *The derivation of chemotaxis equations as limit dynamics of moderately interacting stochastic many-particle systems*, SIAM J. Appl. Math., 61 (2000), pp. 183–212.
[22] B. Tranquillo and W. Alt, *Glossary of terms concerning oriented movement*, in Proceedings of a Workshop on Biological Motion held in Königswinter, 1989, W. Alt and G. Hoffmann, eds., Lecture Notes in Biomath. 89, Springer-Verlag, Heidelberg, 1990, pp. 510–517.

SIAM J. MATH. ANAL.
Vol. 36, No. 4, pp. 1200–1225

GLOBAL EXISTENCE RESULT FOR PAIR DIFFUSION MODELS*

A. GLITZKY† AND R. HÜNLICH†

Abstract. In this paper we prove a global existence result for pair diffusion models in two dimensions. Such models describe the transport of charged particles in semiconductor heterostructures. The underlying model equations are continuity equations for mobile and immobile species coupled with a nonlinear Poisson equation. The continuity equations for the mobile species are nonlinear parabolic PDEs involving drift, diffusion, and reaction terms; the corresponding equations for the immobile species are ODEs containing reaction terms only. Forced by applications to semiconductor technology, these equations have to be considered with nonsmooth data and kinetic coefficients additionally depending on the state variables.

Our proof is based on regularizations, on a priori estimates which are obtained by estimates of the free energy and by Moser iteration, as well as on existence results for the regularized problems. These are obtained by applying the Banach fixed point theorem for the equations of the immobile species, and the Schauder fixed point theorem for the equations of the mobile species.

Key words. reaction-diffusion systems for charged particles, pair diffusion models, global existence, a priori estimates, fixed point theorems

AMS subject classifications. 35K45, 35K57, 35R05, 35D05, 35B45, 80A30

DOI. 10.1137/S0036141002417590

1. The model.

Pair diffusion models describe the transport of charged particles (dopant atoms, point defects, dopant-defect pairs) in semiconductors [4, 7]. In [11] we treated a rather general model of this kind, which we continue to study in this paper. We consider m species X_i. The first $l \le m$ species are mobile, the other ones are immobile. The particle densities u_i of the ith species and some additional potential ψ are the primary unknown functions. They have to satisfy the following initial boundary value problem:

$$
\left\{
\begin{aligned}
&\frac{\partial u_i}{\partial t} + \nabla \cdot j_i + \sum_{(\alpha,\beta)\in\mathcal{R}^\Omega} (\alpha_i - \beta_i)\, R^\Omega_{\alpha\beta} = 0 && \text{on } (0,\infty)\times\Omega,\\
&\nu \cdot j_i - \sum_{(\alpha,\beta)\in\mathcal{R}^\Gamma} (\alpha_i - \beta_i)\, R^\Gamma_{\alpha\beta} = 0 && \text{on } (0,\infty)\times\Gamma,\\
& && i = 1,\dots,l;\\
&\frac{\partial u_i}{\partial t} + \sum_{(\alpha,\beta)\in\mathcal{R}^\Omega} (\alpha_i - \beta_i)\, R^\Omega_{\alpha\beta} = 0 && \text{on } (0,\infty)\times\Omega,\\
& && i = l+1,\dots,m;\\
&-\nabla\cdot(\varepsilon\nabla\psi) + e(\cdot,\psi) - \sum_{i=1}^{m} Q_i(\psi) u_i = f && \text{on } (0,\infty)\times\Omega,\\
&\nu\cdot(\varepsilon\nabla\psi) = 0 && \text{on } (0,\infty)\times\Gamma;\\
&u_i(0) = U_i && \text{on } \Omega,\ i = 1,\dots,m.
\end{aligned}
\right. \tag{1.1}
$$

*Received by the editors November 12, 2002; accepted for publication (in revised form) March 12, 2004; published electronically February 3, 2005.
http://www.siam.org/journals/sima/36-4/41759.html
†Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstraße 39, D-10117 Berlin, Germany (glitzky@wias-berlin.de, huenlich@wias-berlin.de).

Here Γ denotes the boundary of the domain $\Omega \subset \mathbb{R}^2$, and ν is the outer unit normal. For specifying the flux densities j_i and the reaction rates $R^{\Omega}_{\alpha\beta}$, $R^{\Gamma}_{\alpha\beta}$ we introduce other functions related to the primary ones, namely, the chemical activities b_i and electrochemical activities a_i,

$$b_i = \frac{u_i}{p_{0i}}, \quad a_i = b_i \mathrm{e}^{P_i(\psi)}, \qquad P_i(\psi) = \int_0^{\psi} Q_i(s)\mathrm{d}s, \quad i = 1, \dots, m, \tag{1.2}$$

where p_{0i} denotes a reference density, and Q_i is the charge number of the ith species which depends on ψ. Then the flux densities of the mobile species are given by

$$j_i = -D_i(\cdot, b, \psi)\, p_{0i} \left(\nabla b_i + Q_i(\psi)\, b_i \nabla \psi\right), \quad i = 1, \dots, l,$$

where D_i is the diffusivity which depends on the state variables $b = (b_1, \dots, b_m)$ and ψ. All m continuity equations contain volume source terms generated by mass action–type reactions of the form

$$\alpha_1 X_1 + \cdots + \alpha_m X_m \rightleftharpoons \beta_1 X_1 + \cdots + \beta_m X_m, \quad (\alpha, \beta) \in \mathcal{R}^{\Omega},$$

where $\alpha, \beta \in \mathbb{Z}^m_+$ are the vectors of stoichiometric coefficients, and $\mathcal{R}^{\Omega}$ describes the set of all reactions under consideration. The corresponding reaction rates are given by

$$R^{\Omega}_{\alpha\beta} = k^{\Omega}_{\alpha\beta}(x, b_1, \dots, b_m, \psi) \left[\prod_{i=1}^{m} a_i^{\alpha_i} - \prod_{i=1}^{m} a_i^{\beta_i}\right], \quad (\alpha, \beta) \in \mathcal{R}^{\Omega}.$$

The continuity equations for the mobile species include additional boundary source terms generated by boundary reactions with reaction rates given by

$$R^{\Gamma}_{\alpha\beta} = k^{\Gamma}_{\alpha\beta}(x, b_1, \dots, b_l, \psi) \left[\prod_{i=1}^{l} a_i^{\alpha_i} - \prod_{i=1}^{l} a_i^{\beta_i}\right], \quad (\alpha, \beta) \in \mathcal{R}^{\Gamma}.$$

Finally, the nonlinear Poisson equation contains various source terms, namely, the fixed charge density f, the charge density e depending on ψ, and the charge density $\sum_{i=1}^{m} Q_i u_i$ of all particles; ε is the dielectric permittivity.

In heterostructures, which we want to include in our considerations, the reference densities p_{0i} (and other quantities such as D_i, $k^{\Omega}_{\alpha\beta}$, $k^{\Gamma}_{\alpha\beta}$, ε, and e) depend on x, and they may jump when crossing interfaces between different materials. The densities u_i may jump, too, but for the chemical activities b_i of the mobile species and for the potential ψ one can assume that $b_1(t, \cdot), \dots, b_l(t, \cdot)$, $\psi(t, \cdot)$ belong to $H^1(\Omega)$. This regularity can be improved slightly. For example, one obtains that $\psi(t, \cdot)$ belongs to $W^{1,2+\delta}(\Omega)$ if ε is an arbitrary element of $L^{\infty}(\Omega)$ (see [14]), but $W^{2,p}$-regularity can not be expected in general.

It is the aim of the present paper to show that the initial boundary value problem (1.1) has a global solution in a sense which is precisely defined in section 2. In that section all needed assumptions are also listed, and they are motivated by considering the special model in [4, 7]. Here let us only observe that the charge numbers Q_i (see (2.8) for the special model) are monotonic decreasing functions of ψ. This property as well as the special structure of the kinetic relations and natural assumptions on the kinetic coefficients ensure that the evolution problem (see $(\mathcal{P})$ later on) as well as needed regularizations of this problem (see $(\mathcal{P}_N)$, $(\mathcal{P}_M)$ later on) have a convex

Lyapunov function, namely, the free energy. Section 3 contains the proof of the existence result and related assertions. Here we make use of some results derived in our earlier papers [11, 16]. There we did not prove any existence result, but we studied qualitative properties of (possible) solutions, especially the long-time behavior. We give a short summary of these properties in section 4 of the present paper.

Let us yet discuss some simplified models for which results are already known. At first we consider the case that each species has a constant charge number, $Q_i(\psi) = q_i = \text{const}$. Then we arrive at a model which we have studied in [13, 8, 9, 10]. There we assumed that all species are mobile, $l = m$, that the diffusivities do not depend on b, and that the initial values are strongly positive, $U_i \geq c > 0$. The continuity equations were reformulated by introducing the electrochemical potentials $\zeta_i = \ln a_i = \ln b_i + q_i\psi$ (defined for $a_i > 0$). Then the kinetic relations were obtained as

$$j_i = -D_i\, u_i\, \nabla\zeta_i,$$

$$R^\Omega_{\alpha\beta} = k^\Omega_{\alpha\beta}\left[\mathrm{e}^{\sum_{i=1}^m \alpha_i\zeta_i} - \mathrm{e}^{\sum_{i=1}^m \beta_i\zeta_i}\right], \quad R^\Gamma_{\alpha\beta} = k^\Gamma_{\alpha\beta}\left[\mathrm{e}^{\sum_{i=1}^l \alpha_i\zeta_i} - \mathrm{e}^{\sum_{i=1}^l \beta_i\zeta_i}\right].$$

We proved the global existence and uniqueness of a solution and studied its asymptotic behavior. The methods developed in the present paper allow us to handle this class of models also in the case that $l < m$, that $D_1, \dots, D_l$ may depend on b, and that only $U_i \geq 0$ is assumed.

At last we consider the case of a homogeneous material where no physical parameter explicitly depends on x. Especially $p_{0i} = \text{const} > 0$ holds, and for the mobile species $u_i(t,\cdot) \in H^1(\Omega)$ follows. Therefore the flux densities can be rewritten as

$$j_i = -D_i\,(\nabla u_i + Q_i(\psi)\, u_i \nabla\psi), \quad i = 1, \dots, l.$$

Under some additional assumptions (two-dimensional domain with smooth boundary, kinetic coefficients depend only on ψ) global existence and uniqueness results were proven in [21] (all species are mobile, $l = m$) and in [12] (some species can be immobile, $l \leq m$). There higher regularity of the solution was obtained. For example, $\psi(t,\cdot)$ belongs to $W^{2,p}(\Omega)$, $p \geq 2$. One may find in [1] a local existence result for the same simplified model, but in arbitrary space dimension. A special pair diffusion model for uncharged species (and without the Poisson equation) was studied in [15]. The case $l < m$ was treated by passing to the limit $D_i \to 0$, $i = l+1, \dots, m$. Several different asymptotic limits for variants of such a model are discussed in [17].

2. Notation, assumptions, and main result.

2.1. Notation. The notation of function spaces corresponds to that in [18]. By $\mathbb{Z}^k_+, \mathbb{R}^k_+, L^p_+, H^1_+$ we denote the cones of nonnegative elements. If $u \in \mathbb{R}^k$, then $u \geq 0$ ($u > 0$) means $u_i \geq 0\ \forall i$ ($u_i > 0\ \forall i$). $\sqrt{u}$ denotes the vector $\{\sqrt{u_i}\}_{i=1,\dots,k}$. For the scalar product in $\mathbb{R}^k$ we use a centered dot. If $u, v \in \mathbb{R}^k$, then $uv = \{u_i v_i\}_{i=1,\dots,k}$, and u/v is to be understood analogously. If $u \in \mathbb{R}^k_+$ and $\alpha \in \mathbb{Z}^k_+$, then u^α means the product $\prod_{i=1}^k u_i^{\alpha_i}$. In our estimates positive constants, which depend at most on the data of the problem, are denoted by c. Some auxiliary results which are relevant for the paper are collected in the appendix. Finally, we make use of the following.

DEFINITION 2.1. *Consider a function $f : (x,y) \in M \times E \to f(x,y) \in \mathbb{R}$, where E is a closed subset of $\mathbb{R}^k$. We say that $f(x,\cdot)$ is locally Lipschitz continuous uniformly with respect to (w.r.t.) x if $f(x,\cdot)$ is Lipschitz continuous on each compact set $K \subset E$ where the Lipschitz constant depends on K, but not on x.*

2.2. Assumptions. Let us summarize all needed assumptions, which we will apply throughout the rest of the paper:

$$\begin{cases} \Omega \subset \mathbb{R}^2 \text{ is a bounded Lipschitzian domain, } \Gamma = \partial\Omega, \\ m, l \in \mathbb{N},\ 1 \le l \le m,\ U \in L^\infty_+(\Omega, \mathbb{R}^m),\ f \in L^2(\Omega); \end{cases} \tag{2.1}$$

$$\begin{cases} \varepsilon \in L^\infty(\Omega),\ \operatorname{ess\,inf}_{x\in\Omega} \varepsilon(x) > 0, \\ e\colon \Omega \times \mathbb{R} \to \mathbb{R} \text{ satisfies the Carathéodory conditions,} \\ e(x,\cdot) \text{ is locally Lipschitz continuous uniformly w.r.t. } x, \\ |e(x,\psi)| \le c\, e^{c|\psi|} \text{ for a.a. } x \in \Omega,\ \forall \psi \in \mathbb{R} \text{ with some } c > 0, \\ e(x,\psi) - e(x,\overline{\psi}) \ge e_0(x)\,(\psi - \overline{\psi}) \text{ for a.a. } x \in \Omega,\ \forall \psi, \overline{\psi} \in \mathbb{R} \text{ with } \psi \ge \overline{\psi}, \\ \text{where } e_0 \in L^\infty_+(\Omega),\ \|e_0\|_{L^1} > 0; \end{cases} \tag{2.2}$$

$$\begin{cases} \text{for } i = 1, \dots, m\colon \\ p_{0i} \in L^\infty(\Omega),\ \operatorname{ess\,inf}_{x\in\Omega} p_{0i}(x) \ge \epsilon_0 > 0, \\ Q_i \in C^1(\mathbb{R}),\ |Q_i(\psi)| \le c,\ Q_i'(\psi) \le 0 \quad \forall \psi \in \mathbb{R}; \end{cases} \tag{2.3}$$

$$\begin{cases} \mathcal{R}^\Sigma \text{ is a finite subset of } \mathbb{Z}^{m_\Sigma}_+ \times \mathbb{Z}^{m_\Sigma}_+,\ \Sigma = \Omega, \Gamma,\ m_\Omega = m,\ m_\Gamma = l; \\ \text{for } \Sigma = \Omega, \Gamma,\ (\alpha,\beta) \in \mathcal{R}^\Sigma : \\ k^\Sigma_{\alpha\beta}\colon \Sigma \times \mathbb{R}^{m_\Sigma}_+ \times \mathbb{R} \to \mathbb{R}_+ \text{ satisfies the Carathéodory conditions,} \\ k^\Sigma_{\alpha\beta}(x,\cdot,\cdot) \text{ is locally Lipschitz continuous uniformly w.r.t. } x, \\ \forall R > 0 \text{ there exists a constant } c_R > 0 \text{ such that} \\ k^\Sigma_{\alpha\beta}(x,b,\psi) \le c_R \text{ for a.a. } x \in \Sigma,\ \forall b \in \mathbb{R}^{m_\Sigma}_+,\ \forall \psi \in [-R,R]; \end{cases} \tag{2.4}$$

$$\begin{cases} (\alpha_{l+1} + \dots + \alpha_m)\,(\beta_{l+1} + \dots + \beta_m) = 0 \quad \forall (\alpha,\beta) \in \mathcal{R}^\Omega, \\ \text{there exists a constant } c > 0 \text{ such that} \\ \max\limits_{k=1,\dots,m} \big\{(a^\alpha - a^\beta)(\beta_k - \alpha_k)\big\} \le c \sum\limits_{k=1}^{m} a_k^2 + c \quad \forall\, a \in \mathbb{R}^m_+,\ (\alpha,\beta) \in \mathcal{R}^\Omega, \\ \max\limits_{k=1,\dots,l} \big\{(a^\alpha - a^\beta)(\beta_k - \alpha_k)\big\} \le c \sum\limits_{k=1}^{l} a_k + c \quad \forall\, a \in \mathbb{R}^l_+,\ (\alpha,\beta) \in \mathcal{R}^\Gamma; \end{cases} \tag{2.5}$$

$$\begin{cases} \text{for } i = 1, \dots, l\colon \\ D_i\colon \Omega \times \mathbb{R}^m_+ \times \mathbb{R} \to \mathbb{R}_+ \text{ satisfies the Carathéodory conditions,} \\ D_i(x,b,\psi) \ge c > 0 \text{ for a.a. } x \in \Omega,\ \forall b \in \mathbb{R}^m_+,\ \forall \psi \in \mathbb{R}, \\ \forall R > 0 \text{ there exists a constant } c_R > 0 \text{ such that} \\ D_i(x,b,\psi) \le c_R \text{ for a.a. } x \in \Omega,\ \forall b \in \mathbb{R}^m_+,\ \forall \psi \in [-R,R]; \end{cases} \tag{2.6}$$

$$(2.7)\quad \begin{cases} \text{for } i = l+1,\dots,m\text{:} \\ \text{there is a reaction of the form} \\ R^{\Omega}_{\alpha_{(i)}\beta_{(i)}}(x,b,\psi) = k^{\Omega}_{\alpha_{(i)}\beta_{(i)}}(x,b,\psi)\left[\prod_{k=1}^{l} a_k^{\alpha_{(i)k}} - a_i^2\right], \\ \forall R>0 \text{ there exists a constant } c_R>0 \text{ such that} \\ k^{\Omega}_{\alpha_{(i)}\beta_{(i)}}(x,b,\psi) \ge c_R \text{ for a.a. } x\in\Omega,\ \forall b\in\mathbb{R}^m_+,\ \forall \psi\in[-R,R]. \end{cases}$$

2.3. Example. We would like to comment on our assumptions by considering the special pair diffusion model in [4, 7]. This example is concerned with the redistribution of a single dopant (phosphorus) in a homogeneous semiconductor (silicon). Here we have to deal with $m = 5$ species, namely, $X_1 = I$ (silicon interstitials), $X_2 = V$ (vacancies in the silicon lattice), $X_3 = PI$ (phosphorus–interstitial pairs), $X_4 = PV$ (phosphorus–vacancy pairs), and $X_5 = P$ (phosphorus atoms on lattice site). The only immobile species is X_5, and hence $l = 4$.

The charge density e (of electrons and holes, except for the sign) is given by $e(\psi) = e_1\mathrm{e}^{\psi} - e_2\mathrm{e}^{-\psi}$, $e_1, e_2 = \text{const} > 0$, and fulfills the properties required in (2.2). The reference densities p_{0i}, the charge numbers Q_i, and the diffusivities D_i (of the mobile species) are expressions of the form

$$(2.8)\quad p_{0i} = \sum_{k=1}^{k_i} K_{ik},\ Q_i(\psi) = \frac{\sum_{k=1}^{k_i} q_{ik}K_{ik}\mathrm{e}^{-q_{ik}\psi}}{\sum_{k=1}^{k_i} K_{ik}\mathrm{e}^{-q_{ik}\psi}},\ D_i(\psi) = \frac{\sum_{k=1}^{k_i} D_{ik}K_{ik}\mathrm{e}^{-q_{ik}\psi}}{\sum_{k=1}^{k_i} K_{ik}\mathrm{e}^{-q_{ik}\psi}},$$

where $k_i \ge 1$, $K_{ik}, D_{ik} = \text{const} > 0$, $q_{ik} = \text{const}$. Obviously, all properties required in (2.3) and (2.6) are fulfilled. Finally, volume reactions as described in Table 2.1 are considered. The set $\mathcal{R}^{\Omega}$ contains the corresponding (α,β)-pairs. The kinetic coefficients $k^{\Omega}_{\alpha\beta}$ are finite sums of the form

$$k^{\Omega}_{\alpha\beta}(\psi) = \sum_{\gamma} k_{\alpha\beta,\gamma}\mathrm{e}^{-q_{\alpha\beta,\gamma}\psi},\ k_{\alpha\beta,\gamma} = \text{const} > 0,\ q_{\alpha\beta,\gamma} = \text{const}$$

and satisfy the requirements in (2.4). All reactions fulfill the first condition in (2.5), namely, $\alpha_5\beta_5 = 0$. Moreover, all reactions are of second order and provide the first growth condition in (2.5). Reaction no. 6 in Table 2.1 is that which fulfills the assumptions in (2.7). Since here boundary reactions do not occur, we set $\mathcal{R}^{\Gamma} = \emptyset$, and the flux boundary conditions in (1.1) result in $\nu \cdot j_i = 0$, $i = 1,\dots,l$. Thus we see that our basic assumptions can be verified when considering this special model.

TABLE 2.1
Volume reactions, stoichiometric coefficients, and reaction rates.

No.	Reaction	α	β	Reaction rate $R^{\Omega}_{\alpha\beta}$
1	$I + P \rightleftharpoons PI$	$(1,0,0,0,1)$	$(0,0,1,0,0)$	$k^{\Omega}_{\alpha\beta}(\psi)\,(a_1a_5 - a_3)$
2	$V + P \rightleftharpoons PV$	$(0,1,0,0,1)$	$(0,0,0,1,0)$	$k^{\Omega}_{\alpha\beta}(\psi)\,(a_2a_5 - a_4)$
3	$I + V \rightleftharpoons 0$	$(1,1,0,0,0)$	$(0,0,0,0,0)$	$k^{\Omega}_{\alpha\beta}(\psi)\,(a_1a_2 - 1)$
4	$I + PV \rightleftharpoons P$	$(1,0,0,1,0)$	$(0,0,0,0,1)$	$k^{\Omega}_{\alpha\beta}(\psi)\,(a_1a_4 - a_5)$
5	$V + PI \rightleftharpoons P$	$(0,1,1,0,0)$	$(0,0,0,0,1)$	$k^{\Omega}_{\alpha\beta}(\psi)\,(a_2a_3 - a_5)$
6	$PI + PV \rightleftharpoons 2P$	$(0,0,1,1,0)$	$(0,0,0,0,2)$	$k^{\Omega}_{\alpha\beta}(\psi)\,(a_3a_4 - a_5^2)$

For various reasons we formulated our assumptions in somewhat more general forms. First, we want to take into account more immobile dopants and corresponding mobile dopant–defect pairs as well as dopant–dopant pairs, if necessary (see, e.g., [20]). Then the number of volume reactions increases. Second, we allow the appearance of boundary reactions of first order (see the second growth condition in (2.5)). Third, the kinetic coefficients D_i, $k^\Sigma_{\alpha\beta}$ should depend on the state variables u_i (or b_i), too. Fourth, we attach importance to the study of heterostructures where physical parameters explicitly depend on x, namely, discontinuously.

Finally, we give reasons for our assumption that the domain Ω should be two-dimensional. In the continuity equations appear quadratic drift terms and quadratic source terms coming from the reactions. We use Gröger's result concerning the $W^{1,2+\delta}$-regularity of the potential ψ, $\delta > 0$ small [14], and the two-dimensional version of the Gagliardo–Nirenberg inequality (5.3) to estimate the $L^\infty(\mathbb{R}_+, L^p(\Omega))$-norms of u_i (or b_i) for $p \geq 2$. In three dimensions the corresponding procedure does not work. Moreover, in two dimensions we can use Trudinger's imbedding result [25] to simplify the investigation of the nonlinear Poisson equation and of the free energy functional.

2.4. Formulation of the problem. We use the function spaces

$$Y = L^2(\Omega, \mathbb{R}^m), \quad X = \big\{b \in Y\colon b_i \in H^1(\Omega),\ i = 1, \dots, l\big\}$$

and define operators

$$B\colon Y \to Y,\ A, R\colon [X \cap L^\infty_+(\Omega, \mathbb{R}^m)] \times [H^1(\Omega) \cap L^\infty(\Omega)] \to X^*,\ E\colon H^1(\Omega) \times Y \to H^1(\Omega)^*$$

by the relations

$$(Bb, \overline{b})_Y = \int_\Omega \sum_{i=1}^m p_{0i}\, b_i\, \overline{b}_i \,\mathrm{d}x, \quad \overline{b} \in Y,$$

$$\langle A(b,\psi), \overline{b}\rangle_X = \int_\Omega \sum_{i=1}^l D_i(\cdot, b, \psi) p_{0i} (\nabla b_i + b_i Q_i(\psi) \nabla\psi) \cdot \nabla \overline{b}_i \,\mathrm{d}x, \quad \overline{b} \in X,$$

$$\begin{aligned}\langle R(b,\psi), \overline{b}\rangle_X = &\int_\Omega \sum_{(\alpha,\beta)\in\mathcal{R}^\Omega} R^\Omega_{\alpha\beta}(\cdot, b_1, \dots, b_m, \psi) \sum_{i=1}^m (\beta_i - \alpha_i)\, \overline{b}_i \,\mathrm{d}x \\ &+ \int_\Gamma \sum_{(\alpha,\beta)\in\mathcal{R}^\Gamma} R^\Gamma_{\alpha\beta}(\cdot, b_1, \dots, b_l, \psi) \sum_{i=1}^l (\beta_i - \alpha_i)\, \overline{b}_i \,\mathrm{d}\Gamma, \quad \overline{b} \in X,\end{aligned}$$

$$\langle E(\psi, u), \overline{\psi}\rangle_{H^1} = \int_\Omega \left\{ \varepsilon \nabla\psi \cdot \nabla\overline{\psi} + e(\cdot,\psi)\overline{\psi} - \sum_{i=1}^m u_i Q_i(\psi)\overline{\psi} - f\overline{\psi} \right\} \mathrm{d}x, \quad \overline{\psi} \in H^1(\Omega).$$

Let us recall that

$$R^\Sigma_{\alpha\beta}(x, b, \psi) = k^\Sigma_{\alpha\beta}(x, b, \psi)\,(a^\alpha - a^\beta),\ x \in \Sigma,\ b \in \mathbb{R}^{m_\Sigma}_+,\ \psi \in \mathbb{R},\ \Sigma = \Omega, \Gamma,$$

where a is related to b and ψ according to (1.2), $a_i = b_i \mathrm{e}^{P_i(\psi)}$, $i = 1, \dots, m_\Sigma$.

The precise formulation of the initial boundary value problem (1.1) reads as follows:

$$(\mathcal{P})\quad \begin{cases} u'(t) + A(b(t),\psi(t)) = R(b(t),\psi(t)), \\ E(\psi(t),u(t)) = 0,\ u(t) = Bb(t) \text{ for a.a. } t>0,\ u(0)=U, \\ u \in H^1_{\mathrm{loc}}(\mathbb{R}_+, X^*) \cap L^2_{\mathrm{loc}}(\mathbb{R}_+, Y),\ b \in L^2_{\mathrm{loc}}(\mathbb{R}_+, X) \cap L^\infty_{\mathrm{loc}}(\mathbb{R}_+, L^\infty_+(\Omega,\mathbb{R}^m)), \\ \psi \in L^2_{\mathrm{loc}}(\mathbb{R}_+, H^1(\Omega)) \cap L^\infty_{\mathrm{loc}}(\mathbb{R}_+, L^\infty(\Omega)). \end{cases}$$

Remark 2.2. Let (u,b,ψ) be a solution of $(\mathcal{P})$. Lemma 5.1(ii), (iii) ensure that $u, b \in C(\mathbb{R}_+, Y)$. Furthermore one easily obtains that $u, b \in C(\mathbb{R}_+, (L^\infty(\Omega,\mathbb{R}^m), w^*))$, and $\psi \in C(\mathbb{R}_+, H^1(\Omega))$; see Lemma 3.1, too. These properties imply that the relations

$$\begin{aligned} E(\psi(t),u(t)) &= 0 \quad \text{in } H^1(\Omega)^*, \\ u(t) = p_0\, b(t) \text{ in } L^\infty(\Omega,\mathbb{R}^m), &\quad u(t),\ b(t) \ge 0 \text{ a.e. on } \Omega \end{aligned} \tag{2.9}$$

are fulfilled not only for a.a. $t \in \mathbb{R}_+$, but also $\forall t \in \mathbb{R}_+$.

2.5. Main result. Now we formulate the main result of the paper.

THEOREM 2.3. *There exists a solution of* $(\mathcal{P})$.

Further qualitative properties of the evolution problem $(\mathcal{P})$ are summarized in section 4.

3. Proofs.

3.1. The nonlinear Poisson equation. We start with some results concerning the Poisson equation which we need in what follows.

LEMMA 3.1. *For any $u \in Y$ there exists a unique solution $\psi \in H^1(\Omega)$ of the equation $E(\psi,u^+)=0$. Moreover, there is an exponent $q>2$, a positive constant c, and a monotonously increasing function $d\colon \mathbb{R}_+ \to \mathbb{R}_+$ such that*

$$\|\psi - \overline{\psi}\|_{H^1} \le c\,\|u-\overline{u}\|_Y \quad \forall u,\overline{u} \in Y,\ E(\psi,u^+) = E(\overline{\psi},\overline{u}^+) = 0, \tag{3.1}$$

$$\|\psi\|_{L^\infty} \le c\left\{1 + \sum_{i=1}^m \|u_i^+ \ln u_i^+\|_{L^1} + d(\|\psi\|_{H^1})\right\} \quad \forall u \in Y,\ E(\psi,u^+)=0, \tag{3.2}$$

$$\|\psi\|_{W^{1,q}} \le c\left\{1 + \sum_{i=1}^m \|u_i\|_{L^{2q/(2+q)}} + d(\|\psi\|_{H^1})\right\} \quad \forall u \in Y,\ E(\psi,u^+)=0, \tag{3.3}$$

$$\|\psi\|_{H^1} \le c(1+\|u\|_Y) \quad \forall u \in Y,\ E(\psi,u^+)=0. \tag{3.4}$$

Finally, let $S=[0,T]$, $T>0$. Then for every $u \in L^2(S,Y)$ there exists a unique $\psi \in L^2(S,H^1(\Omega))$ such that

$$E(\psi(t),u^+(t)) = 0 \text{ for a.a. } t \in S.$$

If $u \in C(S,Y)$, then $\psi \in C(S,H^1(\Omega))$ follows and the last equation holds $\forall t \in S$.

Proof. For the first existence result and the estimates (3.1), (3.2) we refer to [16, Lemma 1]. The estimate (3.3) is a consequence of Gröger's regularity result for elliptic equations [14, Theorem 1] and of Trudinger's imbedding theorem [25]. Moreover, let ψ_0 be the (unique) solution of $E(\psi_0,0)=0$. According to (3.1) we have $\|\psi-\psi_0\|_{H^1} \le c\|u\|_Y$ if $u \in Y$ and $E(\psi,u^+)=0$. Thus (3.4) follows. The last assertions result from the pointwise existence result and (3.1). ∎

3.2. First regularized problem ($\mathcal{P}_N$). In order to prove Theorem 2.3 we shall consider two regularized problems which are defined on an arbitrary given interval $S = [0,T]$. First we introduce a problem $(\mathcal{P}_N)$ as follows. Let $N \in \mathbb{R}$, $N > 0$, be given and let $\rho_N\colon \mathbb{R}^{m+1} \to [0,1]$ be a Lipschitz continuous function with

$$\rho_N(y) = \begin{cases} 0 & \text{if} \quad |y|_\infty \geq N, \\ 1 & \text{if} \quad |y|_\infty \leq N/2, \end{cases} \qquad |y|_\infty = \max\{|y_1|, \dots, |y_{m+1}|\}.$$

We define the functions $r_i^\Sigma\colon \Sigma \times \mathbb{R}_+^{m_\Sigma} \times \mathbb{R} \to \mathbb{R}$, $i = 1, \dots, m_\Sigma$, $\Sigma = \Omega, \Gamma$, by

$$r_i^\Omega(x,b,\psi) = \rho_N(b,\psi) \sum_{(\alpha,\,\beta)\in\mathcal{R}^\Omega} R_{\alpha\beta}^\Omega(x,b,\psi)(\beta_i - \alpha_i),$$

$$r_i^\Gamma(x,b_1,\dots,b_l,\psi) = \rho_N(b_1,\dots,b_l,0,\dots,0,\psi) \sum_{(\alpha,\,\beta)\in\mathcal{R}^\Gamma} R_{\alpha\beta}^\Gamma(x,b_1,\dots,b_l,\psi)(\beta_i - \alpha_i).$$

These functions satisfy the Carathéodory conditions, and the functions $r_i^\Sigma(x,\cdot,\cdot)$ are Lipschitz continuous uniformly w.r.t. x since $R_{\alpha\beta}^\Sigma(x,\cdot,\cdot)$ are locally Lipschitz continuous uniformly w.r.t. x and ρ_N is a Lipschitz continuous function with compact support. Further important properties of these functions are

$$|r_i^\Sigma(x,b,\psi)| \leq c(N) \quad \text{for a.a. } x \in \Sigma,\ \forall (b,\psi) \in \mathbb{R}_+^{m_\Sigma} \times \mathbb{R},\ i = 1, \dots, m_\Sigma, \tag{3.5}$$

$$\sum_{i=1}^{m_\Sigma} r_i^\Sigma(x,b,\psi)\ (\ln b_i + P_i(\psi)) \leq 0 \quad \text{for a.a. } x \in \Sigma,\ \forall (b,\psi) \in \mathbb{R}_+^{m_\Sigma} \times \mathbb{R},\ b > 0. \tag{3.6}$$

We define the operator $R_N\colon X_+ \times H^1(\Omega) \to X^*$ by

$$\begin{aligned} \langle R_N(b,\psi), \overline{b}\rangle_X = &\int_\Omega \sum_{i=1}^m r_i^\Omega(\cdot, b_1, \dots, b_m, \psi)\, \overline{b}_i \,\mathrm{d}x \\ &+ \int_\Gamma \sum_{i=1}^l r_i^\Gamma(\cdot, b_1, \dots, b_l, \psi)\, \overline{b}_i \,\mathrm{d}\Gamma,\ \overline{b} \in X. \end{aligned}$$

Now our first regularized problem is formulated as follows:

$$(\mathcal{P}_N) \qquad \begin{cases} u'(t) + A(b(t),\psi(t)) = R_N(b(t),\psi(t)), \\ E(\psi(t),u(t)) = 0,\ u(t) = Bb(t) \quad \text{for a.a. } t \in S,\ u(0) = U, \\ u \in H^1(S,X^*) \cap L^2(S,Y),\ b \in L^2(S,X) \cap L^\infty(S, L_+^\infty(\Omega,\mathbb{R}^m)), \\ \psi \in L^2(S,H^1(\Omega)) \cap L^\infty(S, L^\infty(\Omega)). \end{cases}$$

3.3. Energy estimates for solutions of ($\mathcal{P}_N$). We summarize some results which can be obtained as in [11, 16]. Let $\widetilde{F}_1, \widetilde{F}_2 : Y \to \overline{\mathbb{R}}$ be given by

$$\widetilde{F}_1(u) = \int_\Omega \left\{ \frac{\varepsilon}{2}|\nabla\psi|^2 + g(\cdot,\psi) - \sum_{i=1}^m u_i h_i(\psi) \right\} \mathrm{d}x, \quad u \in Y_+, \tag{3.7}$$

where the functions g, h_i are defined by

$$g(x,\psi) = e(x,\psi)\psi - \int_0^\psi e(x,z)\,\mathrm{d}z, \quad x \in \Omega,\ \psi \in \mathbb{R},$$
$$h_i(\psi) = Q_i(\psi)\psi - \int_0^\psi Q_i(z)\,\mathrm{d}z = P_i'(\psi)\psi - P_i(\psi),\ \psi \in \mathbb{R}, \quad i = 1,\dots,m,$$

and $\psi \in H^1(\Omega) \cap L^\infty(\Omega)$ is the unique solution of the Poisson equation $E(\psi, u) = 0$,

$$\widetilde{F}_2(u) = \int_\Omega \sum_{i=1}^m \left\{ u_i \left[\ln \frac{u_i}{p_{0i}} - 1 \right] + p_{0i} \right\} \mathrm{d}x, \quad u \in Y_+, \tag{3.8}$$

$$\widetilde{F}_1(u) = +\infty, \quad \widetilde{F}_2(u) = +\infty, \quad u \in Y \setminus Y_+.$$

Finally, we define the functionals

$$F_k = (\widetilde{F}_k^*|_X)^* : X^* \to \overline{\mathbb{R}},\ k = 1,2,\ F = F_1 + F_2 : X^* \to \overline{\mathbb{R}}. \tag{3.9}$$

The value $F(u)$ represents the free energy of the state $u \in X^*$.

LEMMA 3.2. *The functional $F = F_1 + F_2 : X^* \to \overline{\mathbb{R}}$ is proper, convex, and lower semicontinuous. For $u \in Y_+$ it can be evaluated according to* (3.7), (3.8).

For the proof see [11, Lemma 3.2]. Next, we introduce the functional $D : M_D \to \mathbb{R}$ by the formula

$$D(u) = 4 \int_\Omega \sum_{i=1}^l D_i(\cdot, b, \psi) p_i(\cdot, \psi) |\nabla \sqrt{a_i}|^2 \,\mathrm{d}x,\ u \in M_D, \tag{3.10}$$

$$M_D = \Big\{ u \in L^\infty_+(\Omega, \mathbb{R}^m) : \sqrt{a} \in X, \text{ where } a = u/p(\cdot,\psi) \text{ and } E(\psi, u) = 0 \Big\}, \tag{3.11}$$

where $p_i(\cdot,\psi) = p_{0i}\,\mathrm{e}^{-P_i(\psi)}$, $i = 1,\dots,m$. The functional D is a nonnegative lower estimate for the dissipation rate of problem $(\mathcal{P}_N)$ (and also of problem $(\mathcal{P})$; see (4.3)), where the contributions arising from the reactions have been omitted in view of (3.6). Following the ideas in [16, section 5]) and [11] (see also the similar proof of Lemma 3.15) we obtain the following.

LEMMA 3.3. *Along any solution (u, b, ψ) of $(\mathcal{P}_N)$ the free energy $F(u)$ remains bounded from above and decreases monotonously; more precisely,*

$$F(u(t_2)) + \int_{t_1}^{t_2} D(u(t))\,\mathrm{d}t \le F(u(t_1)) \le F(U), \quad 0 \le t_1 \le t_2 \le T, \tag{3.12}$$

holds. Moreover, there exist constants $c, c_{3.13} > 0$ depending only on the data but not on N and T such that

$$\begin{aligned} &\sum_{i=1}^m \|u_i \ln u_i\|_{L^\infty(S,L^1(\Omega))} \le c, \quad \|u\|_{L^\infty(S,L^1(\Omega,\mathbb{R}^m))} \le c, \\ &\|\psi\|_{L^\infty(S,H^1(\Omega))} \le c, \quad \|\psi\|_{L^\infty(S,L^\infty(\Omega))}, \quad \|\psi\|_{L^\infty(S,L^\infty(\Gamma))} \le c_{3.13} \end{aligned} \tag{3.13}$$

for any solution of $(\mathcal{P}_N)$.

Remark 3.4. Note that the last two estimates of Lemma 3.3 together with the assumptions (2.4), (2.6), (2.7) ensure the existence of constants $c, \tilde{\epsilon}, \epsilon > 0$ such that

$$k^{\Sigma}_{\alpha\beta}(\cdot, b_1, \dots, b_{m_\Sigma}, \psi) \le c \text{ a.e. in } S \times \Sigma,\ (\alpha, \beta) \in \mathcal{R}^{\Sigma},\ \Sigma = \Omega, \Gamma,$$
$$\tilde{\epsilon} \le 2k^{\Omega}_{\alpha_{(i)}\beta_{(i)}}(\cdot, b, \psi)\, \mathrm{e}^{P_i(\psi)} \text{ a.e. in } S \times \Omega,\ i = l+1, \dots, m,$$
$$\epsilon \le D_i(\cdot, b, \psi) p_{0i} \le c \text{ a.e. in } S \times \Omega,\ i = 1, \dots, l,$$

for any solution (u, b, ψ) of $(\mathcal{P}_N)$.

3.4. Further a priori estimates for solutions of $(\mathcal{P}_N)$. The constants in the estimates of this subsection will depend on T. Therefore it is not possible to use these results to obtain global (w.r.t. time) bounds for solutions of $(\mathcal{P})$. Such global bounds are derived in [11] by a modified method.

LEMMA 3.5. *There is a constant $c_{3.14} > 0$ not depending on N such that*

$$\|b_i(t)\|_{L^2} \le c_{3.14} \quad \forall t \in S, \quad i = 1, \dots, m, \tag{3.14}$$

for any solution (u, b, ψ) of $(\mathcal{P}_N)$.

Proof. Let (u, b, ψ) be a solution of $(\mathcal{P}_N)$.

1. Choosing q as in Lemma 3.1 we obtain by Lemmas 3.1 and 3.3 that

$$\|\psi(t)\|_{W^{1,q}} \le c \left(1 + \sum_{i=1}^{m} \|b_i(t)\|_{L^{2q/(2+q)}}\right) \text{ for a.a. } t \in S. \tag{3.15}$$

2. We take into account the assumptions (2.5) and (2.7) concerning the order of the source terms of the reactions and the presence of reactions with quadratic sink terms for the immobile species, respectively. Since $\|\psi\|_{L^\infty(S, L^\infty(\Sigma))} \le c_{3.13}$, $\Sigma = \Omega, \Gamma$, and $|\rho_N(b, \psi)| \le 1$ we find that

$$\int_\Omega \sum_{i=1}^{m} r_i^{\Omega}(\cdot, b, \psi)\, b_i \,\mathrm{d}x$$
$$\le \int_\Omega \rho_N(b, \psi) \sum_{k=l+1}^{m} \left\{ c \sum_{i=1}^{l} (b_i^3 + b_i^2 b_k + b_i b_k^2 + b_k^2 + 1) - \tilde{\epsilon}\, b_k^3 \right\} \mathrm{d}x \le c \sum_{i=1}^{l} \|b_i\|_{L^3}^3 + c,$$
$$\int_\Gamma \sum_{i=1}^{l} r_i^{\Gamma}(\cdot, b_1, \dots, b_l, \psi)\, b_i \,\mathrm{d}\Gamma \le c \sum_{i=1}^{l} \|b_i\|_{L^2(\Gamma)}^2 + c.$$

3. Testing the evolution equation in $(\mathcal{P}_N)$ with $2b$, and using the estimates from step 2 as well as (5.1), (5.3), and Young's inequality, we obtain

$$\sum_{i=1}^{m} \left(\epsilon_0 \|b_i(t)\|_{L^2}^2 - c\|U_i\|_{L^2}^2\right)$$
$$\le \int_0^t \sum_{i=1}^{l} \left\{ -2\epsilon \|b_i\|_{H^1}^2 + c(\|b_i\|_{L^r} \|\psi\|_{W^{1,q}} \|b_i\|_{H^1} + \|b_i\|_{L^3}^3 + \|b_i\|_{L^2(\Gamma)}^2 + 1) \right\} \mathrm{d}s$$
$$\le \int_0^t \sum_{i=1}^{l} \left\{ -\epsilon \|b_i\|_{H^1}^2 + \bar{c}(\|b_i\|_{L^r} \|\psi\|_{W^{1,q}} \|b_i\|_{H^1} + \|b_i\|_{L^2}^4 + 1) \right\} \mathrm{d}s \quad \forall t \in S,$$

where $r = 2q/(q-2)$. Using the estimate $\|b_k\|_{L^{2q/(2+q)}} \le \|b_k\|_{L^1}^{(r-2)/r} \|b_k\|_{L^2}^{2/r}$ as well as Lemma 3.3 and (3.15), (5.3) we calculate

$$\overline{c}\|b_i\|_{L^r}\|\psi\|_{W^{1,q}}\|u_i\|_{H^1} \le c\|b_i\|_{L^2}^{2/r}\left(1+\sum_{k=1}^m \|b_k\|_{L^2}^{2/r}\right)\|b_i\|_{H^1}^{2(r-1)/r}$$
$$\le \epsilon\|b_i\|_{H^1}^2 + c\|b_i\|_{L^2}^2\sum_{k=1}^m \|b_k\|_{L^2}^2 + c.$$

Therefore we can continue the first estimate in step 3 as

$$\sum_{k=1}^m \|b_k(t)\|_{L^2}^2 \le c\int_0^t \sum_{k=1}^m\sum_{i=1}^l \|b_i\|_{L^2}^2\|b_k\|_{L^2}^2\,\mathrm{d}s + c \quad \forall t\in S.$$

Let i, $1\le i\le l$, be fixed. By Lemma 3.3 and (3.10) we find that

$$\|\nabla\sqrt{a_i}\|_{L^2(S,L^2)} \le c,\ \|u_i\|_{L^\infty(S,L^1)} \le c,\ \|\psi\|_{L^\infty(S,L^\infty)} \le c$$

and $\|\sqrt{a_i}\|_{L^2(S,H^1)} \le c$, $\|\sqrt{a_i}\|_{L^\infty(S,L^2)} \le c$. Thus interpolation yields $\|\sqrt{a_i}\|_{L^4(S,L^4)} \le c$ and $\|b_i\|_{L^2(S,L^2)} \le c$. A special form of Gronwall's lemma (see [26, pp. 14, 15]) leads to the desired result. □

Again, let q be chosen as in Lemma 3.1. Since $2q/(2+q) < 2$ we obtain from (3.14), (3.15) the estimate $\|\psi\|_{L^\infty(S,W^{1,q})} \le c_q$. We define

$$\kappa = c_q^{2r} + 1, \text{ where } r = 2q/(q-2),\ q \text{ is as in Lemma 3.1.} \tag{3.16}$$

LEMMA 3.6. *There is a constant $c_{3.17} \ge 1$ not depending on N such that*

$$\|b_i(t)\|_{L^4} \le c_{3.17} \quad \forall t\in S,\ i=1,\dots,m, \tag{3.17}$$

for any solution (u,b,ψ) of $(\mathcal{P}_N)$.

Proof. Let (u,b,ψ) be a solution of $(\mathcal{P}_N)$. We use the test function $4(b_1^3,\dots,b_m^3)$. Using an argument similar to that in step 2 of the proof of Lemma 3.5, we find that

$$\int_\Omega \sum_{i=1}^m r_i^\Omega(\cdot,b,\psi)\,b_i^3\,\mathrm{d}x$$
$$\le \int_\Omega \rho_N(b,\psi)\sum_{k=l+1}^m \Big\{ c\sum_{i=1}^l \big((b_i^2+1)b_k^3 + (b_k^2+1)b_i^3 + b_i^5\big) - \tilde{\epsilon}b_k^5\Big\}\mathrm{d}x \le c\sum_{i=1}^l \|b_i\|_{L^5}^5 + c,$$
$$\int_\Gamma \sum_{i=1}^l r_i^\Gamma(\cdot,b_1,\dots,b_l,\psi)\,b_i^3\,\mathrm{d}\Gamma \le c\sum_{i=1}^l \|b_i\|_{L^4(\Gamma)}^4 + c.$$

Therefore we obtain $\forall t\in S$

$$\sum_{i=1}^m \big(\epsilon_0\|b_i(t)\|_{L^4}^4 - c\|U_i\|_{L^4}^4\big)$$
$$\le \int_0^t \sum_{i=1}^l \big\{ -2\epsilon\|b_i^2\|_{H^1}^2 + c(\|\nabla\psi\|_{L^q}\|\nabla(b_i^2)\|_{L^2}\|b_i^2\|_{L^r} + \|b_i\|_{L^5}^5 + \|b_i\|_{L^4(\Gamma)}^4 + 1)\big\}\,\mathrm{d}s.$$

We apply the trace inequality (5.1), Gagliardo–Nirenberg's inequality (5.3), (3.16), and Young's inequality,

$$\epsilon_0 \sum_{i=1}^{m} \|b_i(t)\|_{L^4}^4 \le \int_0^t \sum_{i=1}^{l} \Big\{ -\frac{\epsilon}{2} \|b_i^2\|_{H^1}^2 + c(\|\psi\|_{W^{1,q}} \|b_i^2\|_{L^1}^{1/r} \|b_i^2\|_{H^1}^{2-1/r} + \|b_i^2\|_{L^1} \|b_i^2\|_{H^1}^{3/2} + \|b_i^2\|_{L^1}^{1/2} \|b_i^2\|_{H^1}^{3/2} + 1) \Big\} \, \mathrm{d}s + c$$
$$\le c \int_0^t \sum_{i=1}^{l} \left(\kappa \|b_i^2\|_{L^1}^2 + \|b_i^2\|_{L^1}^4 + \|b_i^2\|_{L^1}^2 + 1 \right) \mathrm{d}s + c \quad \forall t \in S,$$

and the assertion follows from Lemma 3.5. □

THEOREM 3.7. *There exists a constant $c_{3.18} > 0$ not depending on N such that*

$$\begin{aligned} &\|b_i(t)\|_{L^\infty} \le c_{3.18} \quad \forall t \in S, \quad i = 1, \dots, m, \\ &\|b_i\|_{L^\infty(S, L^\infty(\Gamma))} \le c_{3.18}, \quad i = 1, \dots, l, \end{aligned} \tag{3.18}$$

for any solution (u, b, ψ) of $(\mathcal{P}_N)$.

Proof. The proof will be done in two steps. First, by Moser iteration we establish global upper bounds for the mobile species. Then, using these bounds we derive global upper bounds for the immobile species. Let (u, b, ψ) be a solution of $(\mathcal{P}_N)$. Let $K = \max\{1, \max_{i=1,\dots,m} \|U_i/p_{0i}\|_{L^\infty}\}$ and define $z_i = (b_i - K)^+$, $i = 1, \dots, m$.

1. *Bounds for the mobile species.* Let $p \ge 8$. We use $p(z_1^{p-1}, \dots, z_l^{p-1}, 0, \dots, 0)$ as the test function and define $w_i = z_i^{p/2}$, $i = 1, \dots, l$. At first let us remark that

$$\sum_{i=1}^{l} r_i^\Omega(\cdot, b, \psi) \, z_i^{p-1} \le c \sum_{i=1}^{l} \sum_{k=1}^{m} (b_k^2 + 1) z_i^{p-1} \le c \sum_{i=1}^{l} \left(z_i^{p+1} + \sum_{k=l+1}^{m} z_i^{p-1} z_k^2 \right) + c.$$

With Lemma 3.6 and Hölder's inequality we can estimate

$$\int_\Omega z_i^{p-1} z_k^2 \, \mathrm{d}x \le \|z_i\|_{L^{2(p-1)}}^{p-1} \|z_k\|_{L^4}^2 \le c_{3.17}^2 \|w_i\|_{L^{4(p-1)/p}}^{2(p-1)/p}.$$

Therefore we obtain $\forall t \in S$

$$\epsilon_0 \sum_{i=1}^{l} \|w_i(t)\|_{L^2}^2 \le \int_0^t \sum_{i=1}^{l} \{ -2\epsilon \|w_i\|_{H^1}^2 + cp(\|\nabla\psi\|_{L^q} \|\nabla w_i\|_{L^2} (\|w_i\|_{L^r} + 1) + \|w_i\|_{L^{2(p+1)/p}}^{2(p+1)/p} + c_{3.17}^2 \|w_i\|_{L^{4(p-1)/p}}^{2(p-1)/p} + \|w_i\|_{L^2(\Gamma)}^2 + 1) \} \, \mathrm{d}s.$$

We apply for $k = 1$ and $\widetilde{p} = r$, $\widetilde{p} = 2(p+1)/p$, and $\widetilde{p} = 4(p-1)/p$, respectively, Gagliardo–Nirenberg's inequality (5.3) and continue:

$$\begin{aligned} &\epsilon_0 \sum_{i=1}^{l} \|w_i(t)\|_{L^2}^2 \\ &\le \int_0^t \sum_{i=1}^{l} \{ -\epsilon \|w_i\|_{H^1}^2 + cp^{2r} (\|\psi\|_{W^{1,q}}^{2r} + 1)(\|w_i\|_{L^1}^2 + 1) + cp(\|w_i\|_{H^1}^{(p+2)/p} \|w_i\|_{L^1} \\ &\qquad + c_{3.17}^2 \|w_i\|_{H^1}^{(3p-4)/2p} \|w_i\|_{L^1}^{1/2} + \|w_i\|_{H^1}^{3/2} \|w_i\|_{L^1}^{1/2} + 1) \} \mathrm{d}s \\ &\le \int_0^t \sum_{i=1}^{l} c \{ p^{2r} \kappa (\|w_i\|_{L^1}^2 + 1) + p^4 \|w_i\|_{L^1}^{2p/(p-2)} + p^4 c_{3.17}^8 \|w_i\|_{L^1}^{2p/(p+4)} + p^4 \|w_i\|_{L^1}^2 \} \mathrm{d}s \\ &\le cp^{2r} (\kappa + c_{3.17}^8) \int_0^t \sum_{i=1}^{l} \left(\|w_i\|_{L^1}^{2p/(p-2)} + 1 \right) \mathrm{d}s \\ &\le cp^{2r} \left(\kappa + c_{3.17}^8 \right) T \sum_{i=1}^{l} \left(\sup_{s \in S} \|z_i(s)\|_{L^{p/2}}^{p^2/(p-2)} + 1 \right). \end{aligned}$$

Therefore the iteration formula

$$\sum_{i=1}^{l} \|z_i(t)\|_{L^p}^p + 1 \le \overline{c}\, p^{2r} \left(\kappa + c_{3.17}^8\right)\, T \left(\sum_{i=1}^{l} \sup_{s\in S} \|z_i(s)\|_{L^{p/2}}^{p/2} + 1 \right)^{2p/(p-2)} \qquad \forall t \in S$$

results, where $\overline{c} > 1$ depends only on the data, and $\kappa, r, c_{3.17}$ are defined in (3.16) and Lemma 3.6. Now we set $p = 2^k$, $k \in \mathbb{N}$, $k \ge 3$. The iteration formula yields

$$\gamma_k \le \left(2^{4r} \left(\kappa + c_{3.17}^8\right)\, T\, \overline{c}\, \gamma_2\right)^{c_0 2^k}, \quad \gamma_k = \sum_{i=1}^{l} \sup_{s\in S} \|z_i(s)\|_{L^{2^k}}^{2^k} + 1, \quad c_0 = \prod_{j=1}^{\infty} \frac{2^j}{2^j - 1}.$$

Passing to the limit $k \to \infty$ we obtain

$$\sum_{i=1}^{l} \|z_i(t)\|_{L^\infty} \le \sqrt{l} \left(2^{4r} \left(\kappa + c_{3.17}^8\right)\, T\, \overline{c} \left(\sum_{i=1}^{l} \sup_{s\in S} \|z_i(s)\|_{L^4}^4 + 1 \right) \right)^{c_0} \qquad \forall t \in S.$$

With Lemma 3.6 and (5.2) the desired estimates for b_i, $i = 1, \dots, l$, are verified.

2. *Bounds for the immobile species.* Now, let $p \ge 2$. We use the test function $p(0, \dots, 0, z_{l+1}^{p-1}, \dots, z_m^{p-1})$. Taking into account the assumptions (2.7), (2.5), the estimates for b_i, $i = 1, \dots, l$, obtained in step 1, as well as the inequalities $b_k \ge z_k \ge 0$ we find that

$$\begin{aligned} \sum_{k=l+1}^{m} r_k^\Omega(\cdot, b, \psi)\, z_k^{p-1} &\le c \sum_{i=1}^{l} \sum_{k=l+1}^{m} (b_i^2 + b_i + 1) z_k^{p-1} - \tilde{\epsilon} \sum_{k=l+1}^{m} z_k^{p+1} \\ &\le \tilde{c} \sum_{k=l+1}^{m} z_k^{p-1} - \tilde{\epsilon} \sum_{k=l+1}^{m} z_k^{p+1} \le (m-l) \frac{\tilde{c}^{(p+1)/2}}{\tilde{\epsilon}^{(p-1)/2}}. \end{aligned}$$

The last estimate follows from Young's inequality. Therefore we obtain

$$\epsilon_0 \sum_{k=l+1}^{m} \|z_k(t)\|_{L^p}^p \le pT|\Omega|(m-l)\, \frac{\tilde{c}^{(p+1)/2}}{\tilde{\epsilon}^{(p-1)/2}} \quad \forall t \in S.$$

And consequently,

$$\|z_k(t)\|_{L^p} \le \left(pT|\Omega|(m-l)/\epsilon_0\right)^{1/p}\ \frac{\tilde{c}^{(p+1)/2p}}{\tilde{\epsilon}^{(p-1)/2p}} \le c_\infty\ \ \forall t \in S,\ k = l+1, \dots, m.$$

Passing to the limit $p \to \infty$ we get $\|z_k(t)\|_{L^\infty} \le c_\infty$ $\forall t \in S$, $k = l+1, \dots, m$, which leads to the desired estimates for b_k, $k = l+1, \dots, m$. ∎

3.5. Second regularized problem ($\mathcal{P}_M$). We prove the solvability of ($\mathcal{P}_N$) for fixed $N > 0$ by means of a second regularization ($\mathcal{P}_M$). Let

$$M^* = \max\left\{N + 1, \max_{i=1,\dots,m} \|U_i/p_{0i}\|_{L^\infty}\right\}, \tag{3.19}$$

and let $M \ge M^*$. We denote by σ_M the projection from $\mathbb{R}$ onto $[-M, M]$,

$$\sigma_M(y) = \operatorname{sign}(y) \min\{|y|, M\},\ \ y \in \mathbb{R}.$$

Moreover, we introduce the functions

$$D_{iM}(x,b,\psi) = D_i(x,b^+,\sigma_M(\psi)),\ i=1,\dots,l,\ x\in\Omega,\ b\in\mathbb{R}^m,\ \psi\in\mathbb{R},$$

define the operator $A_M\colon X\times H^1(\Omega)\to X^*$,

$$\langle A_M(b,\psi),\overline{b}\rangle_X = \int_\Omega \sum_{i=1}^{l} D_{iM}(\cdot,b,\psi)p_{0i}(\nabla b_i + [\sigma_M(b_i)]^+ Q_i(\psi)\nabla\psi)\cdot\nabla\overline{b}_i\,\mathrm{d}x,\ \overline{b}\in X,$$

and consider the following problem:

$$(\mathcal{P}_M)\qquad \begin{cases} u'(t) + A_M(b(t),\psi(t)) = R_N(b^+(t),\psi(t)), \\ E(\psi(t),u^+(t)) = 0, \quad u(t) = Bb(t) \quad \text{for a.a. } t\in S, \quad u(0) = U, \\ u\in H^1(S,X^*)\cap L^2(S,Y),\ b\in L^2(S,X),\ \psi\in L^2(S,H^1(\Omega)). \end{cases}$$

Let us remark that we have $u,b\in C(S,Y)$, $\psi\in C(S,H^1(\Omega))$ for solutions of $(\mathcal{P}_M)$.

3.6. Existence result for $(\mathcal{P}_M)$. First we derive an equivalent formulation of $(\mathcal{P}_M)$. We write b in the form $b=(v,w)$, where $v=(b_1,\dots,b_l)$ and $w=(b_{l+1},\dots,b_m)$ denote the chemical activities of the mobile and immobile species, respectively. We introduce the spaces

$$Y^l = L^2(\Omega,\mathbb{R}^l),\ Y^{m-l} = L^2(\Omega,\mathbb{R}^{m-l}),\ X^l = H^1(\Omega,\mathbb{R}^l)$$

and the operators $B_{\mathrm{mob}}\colon L^2(S,Y^l)\to L^2(S,Y^l)$, $B_{\mathrm{imm}}\colon L^2(S,Y^{m-l})\to L^2(S,Y^{m-l})$,

$$\langle (B_{\mathrm{mob}}v)(t),\overline{v}\rangle_{Y^l} = \int_\Omega \sum_{i=1}^{l} p_{0i}\, v_i(t)\,\overline{v}_i\,\mathrm{d}x,\quad \overline{v}\in Y^l,$$

$$\langle (B_{\mathrm{imm}}w)(t),\overline{w}\rangle_{Y^{m-l}} = \int_\Omega \sum_{i=1}^{m-l} p_{0(l+i)}\, w_i(t)\,\overline{w}_i\,\mathrm{d}x,\quad \overline{w}\in Y^{m-l},\quad t\in S.$$

Moreover, we define the operators

$$\begin{aligned} &R_{\mathrm{mob}}\colon L^2(S,X^l)\times L^2(S,Y^{m-l})\times L^2(S,H^1(\Omega))\to L^2(S,X^{l*}),\\ &R_{\mathrm{imm}}\colon L^2(S,X^l)\times L^2(S,Y^{m-l})\times L^2(S,H^1(\Omega))\to L^2(S,Y^{m-l}),\\ &A_{\mathrm{mob}}\colon L^2(S,X^l)\times L^2(S,X^l)\times L^2(S,Y^{m-l})\times L^2(S,H^1(\Omega))\to L^2(S,X^{l*}),\\ &A^0_{\mathrm{mob}}\colon L^2(S,X^l)\times L^2(S,Y^{m-l})\times L^2(S,H^1(\Omega))\to L^2(S,X^{l*}) \end{aligned}$$

as follows:

$$\langle R_{\mathrm{mob}}(v,w,\psi),\overline{v}\rangle_{L^2(S,X^l)} = \int_S \langle R_N(v^+,w^+,\psi),(\overline{v},0)\rangle_X\,\mathrm{d}s,\quad \overline{v}\in L^2(S,X^l),$$

$$\langle R_{\mathrm{imm}}(v,w,\psi),\overline{w}\rangle_{L^2(S,Y^{m-l})} = \int_S \langle R_N(v^+,w^+,\psi),(0,\overline{w})\rangle_X\,\mathrm{d}s,\quad \overline{w}\in L^2(S,Y^{m-l}),$$

$$\langle A_{\mathrm{mob}}(v;\widehat{v},w,\psi),\overline{v}\rangle_{L^2(S,X^l)} = \int_S\int_\Omega \sum_{i=1}^{l}(D_{iM}(\cdot,\widehat{v},w,\psi)\,p_{0i}\,\nabla v_i\cdot\nabla\overline{v}_i + v_i\overline{v}_i)\,\mathrm{d}x\,\mathrm{d}s,$$

$$\begin{aligned}\langle A^0_{\mathrm{mob}}(v,w,\psi),\overline{v}\rangle_{L^2(S,X^l)} = \int_S\int_\Omega \sum_{i=1}^{l}(&D_{iM}(\cdot,v,w,\psi)\,p_{0i}\,[\sigma_M(v_i)]^+ Q_i(\psi)\nabla\psi\cdot\nabla\overline{v}_i\\ &- v_i\overline{v}_i)\,\mathrm{d}x\,\mathrm{d}s,\quad \overline{v}\in L^2(S,X^l).\end{aligned}$$

For any given $v \in L^2(S,Y^l)$, $w \in L^2(S,Y^{m-l})$ we have that $(B_{\rm mob}v, B_{\rm imm}w) \in L^2(S,Y)$, and by Lemma 3.1 we find a unique $\psi \in L^2(S,H^1(\Omega)) \cap L^\infty(S,L^\infty(\Omega))$ such that

$$E(\psi(t),((B_{\rm mob}v)^+(t),(B_{\rm imm}w)^+(t))) = 0 \quad \text{for a.a. } t \in S.$$

We denote by $\mathcal{N}\colon L^2(S,Y^l)\times L^2(S,Y^{m-l}) \to L^2(S,H^1(\Omega))$ the corresponding solution operator, $\psi = \mathcal{N}(v,w)$. Problem $(\mathcal{P}_M)$ is obviously equivalent to the system of equations

$$\begin{aligned}(B_{\rm mob}v)' + A_{\rm mob}(v;v,w,\mathcal{N}(v,w)) &= R_{\rm mob}(v,w,\mathcal{N}(v,w))\\ &\quad - A^0_{\rm mob}(v,w,\mathcal{N}(v,w)),\\ (B_{\rm mob}v)(0) = U_{\rm mob} &= (U_1,\dots,U_l), \quad v \in W^l,\end{aligned} \tag{3.20}$$

where $W^l = \{v \in L^2(S,X^l)\colon (B_{\rm mob}v)' \in L^2(S,X^{l*})\} \subset C(S,Y^l)$,

$$\begin{aligned}(B_{\rm imm}w)' &= R_{\rm imm}(v,w,\mathcal{N}(v,w)),\\ (B_{\rm imm}w)(0) = U_{\rm imm} &= (U_{l+1},\dots,U_m), \quad B_{\rm imm}w \in H^1(S,Y^{m-l}).\end{aligned} \tag{3.21}$$

Let us briefly outline how these equations will be solved. We start with some fixed $\widehat{v} \in W^l$. First we solve the initial value problem

$$\begin{aligned}(B_{\rm imm}w)' &= R_{\rm imm}(\widehat{v},w,\mathcal{N}(\widehat{v},w)),\\ (B_{\rm imm}w)(0) &= U_{\rm imm},\ B_{\rm imm}w \in H^1(S,Y^{m-l}).\end{aligned} \tag{3.22}$$

This problem has a unique solution $w = \mathcal{T}\widehat{v}$ (see Lemma 3.8). Next we solve the initial boundary value problem

$$\begin{aligned}(B_{\rm mob}v)' + A_{\rm mob}(v;\widehat{v},\mathcal{T}\widehat{v},\mathcal{N}(\widehat{v},\mathcal{T}\widehat{v})) &= R_{\rm mob}(\widehat{v},\mathcal{T}\widehat{v},\mathcal{N}(\widehat{v},\mathcal{T}\widehat{v}))\\ &\quad - A^0_{\rm mob}(\widehat{v},\mathcal{T}\widehat{v},\mathcal{N}(\widehat{v},\mathcal{T}\widehat{v})),\\ (B_{\rm mob}v)(0) &= U_{\rm mob},\ v \in W^l.\end{aligned} \tag{3.23}$$

Also this problem has a unique solution $v = \mathcal{Q}\widehat{v}$ (see Lemma 5.2). The operator $\mathcal{Q}$ is compact and continuous (see Lemma 3.10). Using Schauder's fixed point theorem we obtain a fixed point v of $\mathcal{Q}$ (see Lemma 3.11). Then $(v,\mathcal{T}v)$ is a solution of (3.20), (3.21).

Now we give the detailed proofs. The constants c in the estimates of this subsection can depend on the data and on M,N,T.

LEMMA 3.8. *For any $\widehat{v} \in W^l$ there exists a unique solution w of problem* (3.22), *and w belongs to $H^1(S,Y^{m-l}) \subset C(S,Y^{m-l})$.*

Proof. Let $\widehat{v} \in W^l$ be fixed. The initial value problem (3.22) is obviously equivalent to the initial value problem

$$w' + G\,w = 0, \quad w(0) = (B_{\rm imm})^{-1}U_{\rm imm}, \quad w \in H^1(S,Y^{m-l}), \tag{3.24}$$

where $G\colon L^2(S,Y^{m-l}) \to L^2(S,Y^{m-l})$ is defined by

$$G\,w = -(B_{\rm imm})^{-1}[R_{\rm imm}(\hat{v},w,\mathcal{N}(\widehat{v},w))], \quad w \in L^2(S,Y^{m-l}).$$

For $\widehat{v} \in L^2(S, Y^l)$, $w \in L^2(S, Y^{m-l})$ we have $\mathcal{N}(\widehat{v}, w) \in L^2(S, H^1(\Omega))$ and

$$\|\mathcal{N}(\widehat{v}, w^1) - \mathcal{N}(\widehat{v}, w^2)\|_{L^2(S,H^1)} \le c\|w^1 - w^2\|_{L^2(S,Y^{m-l})} \quad \forall w^1,\, w^2 \in L^2(S, Y^{m-l}).$$

Since the functions $r_i^\Omega(x, \cdot, \cdot)$ are Lipschitz continuous uniformly w.r.t. x, the estimate

$$\|G\, w^1 - G\, w^2\|_{L^2(S,Y^{m-l})} \le L\|w^1 - w^2\|_{L^2(S,Y^{m-l})} \quad \forall w^1,\, w^2 \in L^2(S, Y^{m-l})$$

follows, and [6, Chapter V, Theorem 1.3] ensures the existence of a unique solution of (3.24). In principle, this result was obtained by means of the Banach fixed point theorem. ∎

We denote by $\mathcal{T}\colon W^l \to H^1(S, Y^{m-l})$ the operator which assigns to $\widehat{v}$ the solution w of (3.22).

LEMMA 3.9. *There exists a constant $c > 0$ such that the following estimates hold:*

$$\|\mathcal{T}\widehat{v}^1 - \mathcal{T}\widehat{v}^2\|_{C(S,Y^{m-l})} \le c\|\widehat{v}^1 - \widehat{v}^2\|_{L^2(S,Y^l)} \quad \forall\, \widehat{v}^1, \widehat{v}^2 \in W^l,$$
$$\|\mathcal{T}\widehat{v}\|_{C(S,Y^{m-l})} \le c \quad \forall\, \widehat{v} \in W^l.$$

Proof. Let $w^k = \mathcal{T}\widehat{v}^k$, $k = 1, 2$. Using the test function $\overline{w} = w^1 - w^2$ for the corresponding problems (3.22), the Lipschitz continuity of r_i^Ω, Hölder's inequality, and Lemma 3.1, we find that

$$\begin{aligned}
\|\overline{w}(t)\|_{Y^{m-l}}^2 &\le c\|(B_{\mathrm{imm}}(\overline{w}(t))\|_{Y^{m-l}}^2 \\
&\le c\int_0^t \left(\|\overline{w}\|_{Y^{m-l}} + \|\widehat{v}^1 - \widehat{v}^2\|_{Y^l} + \|\mathcal{N}(\widehat{v}^1, w^1) - \mathcal{N}(\widehat{v}^2, w^2)\|_{H^1}\right) \|\overline{w}\|_{Y^{m-l}}\, \mathrm{d}s \\
&\le c\int_0^t \left(\|\overline{w}\|_{Y^{m-l}}^2 + \|\widehat{v}^1 - \widehat{v}^2\|_{Y^l}^2\right)\, \mathrm{d}s.
\end{aligned}$$

Gronwall's lemma leads to the first assertion. Next, testing (3.22) with $w = \mathcal{T}\widehat{v}$ and using (3.5), the estimate

$$\|w(t)\|_{Y^{m-l}}^2 \le c + c\int_0^t \|w(s)\|_{Y^{m-l}}\, \mathrm{d}s \le c + \int_0^t \|w(s)\|_{Y^{m-l}}^2 \quad \forall t \in S$$

follows, where c does not depend on $\widehat{v}$. Again applying Gronwall's lemma the second assertion is obtained. ∎

Next we conclude that for given $\widehat{v} \in W^l$ the initial boundary value problem (3.23) has a unique solution. This follows from Lemma 5.2 since B_{mob} and A_{mob} are diagonal and the right-hand side belongs to $L^2(S, X^{l*})$. We denote by $\mathcal{Q}\colon W^l \to W^l$ the operator which assigns to $\widehat{v}$ the solution v of (3.23).

LEMMA 3.10. *The operator $\mathcal{Q}$ is compact and continuous.*

Proof. 1. Let $\{\widehat{v}_n\} \subset W^l$ be bounded. We show that $\{\mathcal{Q}\widehat{v}_n\} \subset W^l$ contains a convergent subsequence. Because of Lemma 5.1(v) we may assume that there exists an element $\widehat{v} \in W^l$ such that $\widehat{v}_n \to \widehat{v}$ in $L^2(S, Y^l)$ as well as in $L^2(S, L^2(\Gamma, \mathbb{R}^l))$. Let

$$v_n = \mathcal{Q}\widehat{v}_n,\ v = \mathcal{Q}\widehat{v},\ w_n = \mathcal{T}\widehat{v}_n,\ w = \mathcal{T}\widehat{v},\ \psi_n = \mathcal{N}(\widehat{v}_n, w_n),\ \psi = \mathcal{N}(\widehat{v}, w).$$

From Lemmas 3.9 and 3.1 it follows that

$$w_n \to w \text{ in } L^2(S, Y^{m-l}), \quad \psi_n \to \psi \text{ in } L^2(S, H^1).$$

Using the test function $v_n - v$ we obtain

$$\begin{aligned}
&\frac{\epsilon_0}{2}\|(v_n-v)(t)\|_{Y^l}^2 + \int_0^t \epsilon\|v_n-v\|_{X^l}^2\,\mathrm{d}s \\
&\quad\le \int_0^t c\Big\{ \big(\|\widehat{v}_n-\widehat{v}\|_{L^2(\Gamma,\mathbb{R}^l)} + \|\psi_n-\psi\|_{L^2(\Gamma)}\big)\,\|v_n-v\|_{L^2(\Gamma,\mathbb{R}^l)} \\
&\qquad + \big(\|\widehat{v}_n-\widehat{v}\|_{Y^l} + \|w_n-w\|_{Y^{m-l}} + \|\psi_n-\psi\|_{L^2}\big)\,\|v_n-v\|_{Y^l} \\
&\qquad + \int_\Omega \sum_{i=1}^{l} \{ (|\nabla v_i| + |\nabla\psi|)\,|D_{iM}(\cdot,\widehat{v}_n,w_n,\psi_n) - D_{iM}(\cdot,\widehat{v},w,\psi)|\,|\nabla(v_{ni}-v_i)| \\
&\qquad\quad + \big(|Q_i(\psi_n)-Q_i(\psi)| + |[\sigma_M(\widehat{v}_{ni})]^+ - [\sigma_M(\widehat{v}_i)]^+|\big)\,|\nabla\psi|\,|\nabla(v_{ni}-v_i)| \\
&\qquad\quad + |\nabla(\psi_n-\psi)|\,|\nabla(v_{ni}-v_i)|\}\,\mathrm{d}x\Big\}\,\mathrm{d}s \quad \forall t\in S.
\end{aligned}$$

Applying Hölder's inequality and Lemma 3.9 we arrive at

$$\begin{aligned}
\|v_n-v\|_{L^2(S,X^l)}^2 \le c\|v_n-v\|_{L^2(S,X^l)}\{&\|\widehat{v}_n-\widehat{v}\|_{L^2(S,Y^l)} + \|\widehat{v}_n-\widehat{v}\|_{L^2(S,L^2(\Gamma,\mathbb{R}^l))} \\
&+ \|\psi_n-\psi\|_{L^2(S,H^1)} + I_n\},
\end{aligned}$$

$$\begin{aligned}
I_n = \sum_{i=1}^{l}\Big\{ &\Big[\int_0^T\int_\Omega |D_{iM}(\cdot,\widehat{v}_n,w_n,\psi_n) - D_{iM}(\cdot,\widehat{v},w,\psi)|^2\,|\nabla v_i|^2\,\mathrm{d}x\mathrm{d}s\Big]^{1/2} \\
&+\Big[\int_0^T\int_\Omega |D_{iM}(\cdot,\widehat{v}_n,w_n,\psi_n) - D_{iM}(\cdot,\widehat{v},w,\psi)|^2\,|\nabla\psi|^2\,\mathrm{d}x\mathrm{d}s\Big]^{1/2} \\
&+\Big[\int_0^T\int_\Omega |Q_i(\psi_n)-Q_i(\psi)|^2\,|\nabla\psi|^2\,\mathrm{d}x\mathrm{d}s\Big]^{1/2} \\
&+\Big[\int_0^T\int_\Omega |[\sigma_M(\widehat{v}_{ni})]^+ - [\sigma_M(\widehat{v}_i)]^+|^2\,|\nabla\psi|^2\,\mathrm{d}x\mathrm{d}s\Big]^{1/2}\Big\}.
\end{aligned}$$

Properties of superposition operators ensure that the last four bracketed terms tend to zero if $n\to\infty$. Thus in summary we find that $v_n\to v$ in $L^2(S,X^l)$. Next we obtain

$$\begin{aligned}
\|(B_{\mathrm{mob}}v_n)'-(B_{\mathrm{mob}}v)'\|_{L^2(S,X^{l*})} \le{}& \|R_{\mathrm{mob}}(\widehat{v}_n,w_n,\psi_n) - R_{\mathrm{mob}}(\widehat{v},w,\psi)\|_{L^2(S,X^{l*})} \\
&+\|A_{\mathrm{mob}}(v_n;\widehat{v}_n,w_n,\psi_n) - A_{\mathrm{mob}}(v;\widehat{v},w,\psi)\|_{L^2(S,X^{l*})} \\
&+\|A^0_{\mathrm{mob}}(\widehat{v}_n,w_n,\psi_n) - A^0_{\mathrm{mob}}(\widehat{v},w,\psi)\|_{L^2(S,X^{l*})} \\
\le{}& c\{\,\|v_n-v\|_{L^2(S,X^l)} + \|\widehat{v}_n-\widehat{v}\|_{L^2(S,Y^l)} \\
&+\|\widehat{v}_n-\widehat{v}\|_{L^2(S,L^2(\Gamma,\mathbb{R}^l))} + \|w_n-w\|_{L^2(S,Y^{m-l})} \\
&+\|\psi_n-\psi\|_{L^2(S,H^1)} + I_n\},
\end{aligned}$$

and we arrive at $v_n\to v$ in W^l.

2. The continuity of the operator $\mathcal{Q}$ can be shown by similar arguments. ∎

LEMMA 3.11. *The operator $\mathcal{Q}$ has a fixed point.*

Proof. Let $\widehat{v} \in W^l$, $\psi = \mathcal{N}(\widehat{v}, \mathcal{T}\widehat{v})$, and $v = \mathcal{Q}\widehat{v}$. We use v as the test function for (3.23), take into account the properties of D_{iM}, Q_i, and apply Lemma 3.1, (5.1), Lemma 3.9, the boundedness of r_i^{Σ}, and Young's inequality. Thus we obtain

$$\begin{aligned}
&\epsilon_0\|v(t)\|^2_{Y^l} - c\|(U_1,\dots,U_l)\|^2_{Y^l} + 2\epsilon\int_0^t \|v\|^2_{X^l}\,\mathrm{d}s \\
&\quad\le c\int_0^t \big(1 + \|v\|^2_{Y^l} + \|\widehat{v}\|^2_{Y^l} + \|\psi\|_{H^1}\|v\|_{X^l} + \|v\|^2_{L^2(\Gamma,\mathbb{R}^l)}\big)\mathrm{d}s \qquad (3.25)\\
&\quad\le \int_0^t \big(\epsilon\|v\|^2_{X^l} + c\,\big(1 + \|v\|^2_{Y^l} + \|\widehat{v}\|^2_{Y^l}\big)\big)\mathrm{d}s \quad \forall t \in S.
\end{aligned}$$

Therefore we find a constant $\overline{c} > 0$ such that $\forall k > 0$

$$\begin{aligned}
&e^{-kt}\left(\|v(t)\|^2_{Y^l} + \int_0^t \|v\|^2_{X^l}\,\mathrm{d}s\right)\\
&\le \overline{c} + \overline{c}e^{-kt}\int_0^t \left\{\left\{\|v\|^2_{Y^l} + \|\widehat{v}\|^2_{Y^l} + \int_0^s \left(\|v\|^2_{X^l} + \|\widehat{v}\|^2_{X^l}\right)\,\mathrm{d}\tau\right\}e^{-ks}e^{ks}\right\}\mathrm{d}s\\
&\le \overline{c} + \overline{c}e^{-kt}\sup_{s\in S}\left\{\left\{\|v(s)\|^2_{Y^l} + \|\widehat{v}(s)\|^2_{Y^l} + \int_0^s \left(\|v\|^2_{X^l} + \|\widehat{v}\|^2_{X^l}\right)\,\mathrm{d}\tau\right\}e^{-ks}\right\}\frac{e^{kt}-1}{k}.
\end{aligned}$$

Choosing now $k \ge 3\overline{c}$ we obtain

$$\begin{aligned}
&\sup_{t\in S} e^{-kt}\left(\|v(t)\|^2_{Y^l} + \int_0^t \|v(s)\|^2_{X^l}\,\mathrm{d}s\right)\\
&\quad\le \frac{3}{2}\overline{c} + \frac{1}{2}\sup_{t\in S}\left\{e^{-kt}\left(\|\widehat{v}(t)\|^2_{Y^l} + \int_0^t \|\widehat{v}(s)\|^2_{X^l}\,\mathrm{d}s\right)\right\}.
\end{aligned}$$

Again using Lemmas 3.1 and 3.9 we estimate

$$\begin{aligned}
&\|(B_{\mathrm{mob}}v)'\|_{L^2(S,X^{l*})}\\
&= \sup_{\|\overline{v}\|_{L^2(S,X^l)}\le 1} \langle R_{\mathrm{mob}}(\widehat{v},\mathcal{T}\widehat{v},\psi) - A_{\mathrm{mob}}(v;\widehat{v},\mathcal{T}\widehat{v},\psi) - A^0_{\mathrm{mob}}(\widehat{v},\mathcal{T}\widehat{v},\psi),\overline{v}\rangle_{L^2(S,X^l)}\\
&\le c\,\left(\|v\|_{L^2(S,X^l)} + \|\psi\|_{L^2(S,H^1)} + 1\right) \le c\,\left(\|v\|_{L^2(S,X^l)} + \|\widehat{v}\|_{L^2(S,Y^l)} + 1\right)\\
&\le \widetilde{c}\left(\|v\|_{L^2(S,X^l)} + \left[\sup_{t\in S}\left\{e^{-kt}\left(\|\widehat{v}(t)\|^2_{Y^l} + \int_0^t \|\widehat{v}(s)\|^2_{X^l}\,\mathrm{d}s\right)\right\}e^{kT}\right]^{1/2} + 1\right).
\end{aligned}$$

Now we define the set

$$\begin{aligned}
\mathcal{B} = \Bigg\{ v \in W^l\colon &\sup_{t\in S}\left\{e^{-kt}\left(\|v(t)\|^2_{Y^l} + \int_0^t \|v\|^2_{X^l}\,\mathrm{d}s\right)\right\} \le 3\overline{c},\\
&\|(B_{\mathrm{mob}}v)'\|_{L^2(S,X^{l*})} \le \widetilde{c}\left(2\sqrt{3\overline{c}e^{kT}} + 1\right)\Bigg\}.
\end{aligned}$$

This set is a nonempty, bounded, closed, and convex subset of W^l with the property that $\mathcal{Q}(\mathcal{B}) \subset \mathcal{B}$. Since the operator $\mathcal{Q}$ is compact and continuous, the assertion follows from the Schauder fixed point theorem. ∎

THEOREM 3.12. *There exists a solution (u,b,ψ) of $(\mathcal{P}_M)$.*

Proof. Because of Lemma 3.11 there exists a solution v of the problem

$$(B_{\mathrm{mob}}v)'+A_{\mathrm{mob}}(v;v,\mathcal{T}v,\mathcal{N}(v,\mathcal{T}v)) = R_{\mathrm{mob}}(v,\mathcal{T}v,\mathcal{N}(v,\mathcal{T}v))-A^0_{\mathrm{mob}}(v,\mathcal{T}v,\mathcal{N}(v,\mathcal{T}v)),$$

$$(B_{\mathrm{mob}}v)(0) = (U_1,\dots,U_l), \quad v \in W^l.$$

We set $w = \mathcal{T}v \in H^1(S,Y^{m-l})$. Then the pair (v,w) fulfills (3.20) and (3.21), which represent an equivalent formulation of problem $(\mathcal{P}_M)$. ∎

3.7. Energy estimates for solutions of $(\mathcal{P}_M)$. First, we prove the following.

LEMMA 3.13. *For any solution (u,b,ψ) of $(\mathcal{P}_M)$ and for every $t \in S$ the inequalities $b(t), u(t) \geq 0$ a.e. on Ω are fulfilled.*

Proof. Let (u,b,ψ) be a solution of $(\mathcal{P}_M)$. We test the evolution equation with the function $-b^-$. Taking into account that

$$(\nabla b_i + [\sigma_M(b_i)]^+ Q_i(\psi)\nabla\psi)\cdot\nabla b_i^- \leq 0, \quad i = 1,\dots,l,$$

$$-r_i^{\Sigma}(\cdot, b_1^+,\dots,b_{m_\Sigma}^+,\psi)\, b_i^- \leq 0, \quad i = 1,\dots,m_\Sigma,\ \Sigma = \Omega,\Gamma,$$

we find that $\|b^-(t)\|_Y^2 \leq 0\ \forall t \in S$. ∎

Next, we introduce a regularized free energy functional F_M, which is adapted to the regularizations in problem $(\mathcal{P}_M)$. We define the function

$$l_M(y) = \begin{cases} \ln y & \text{if} \quad 0 < y \leq M, \\ \ln M - 1 + \dfrac{y}{M} & \text{if} \quad y > M \end{cases}$$

and the functional $\widetilde{F}_{M2} : Y \to \overline{\mathbb{R}}$ by

$$\widetilde{F}_{M2}(u) = \begin{cases} \displaystyle\int_\Omega \sum_{i=1}^m \int_{p_{0i}}^{u_i} l_M(z/p_{0i})\,\mathrm{d}z\,\mathrm{d}x & \text{if} \quad u \in Y_+, \\ +\infty & \text{if} \quad u \in Y \setminus Y_+. \end{cases} \tag{3.26}$$

Moreover, we set

$$F_{M2} = (\widetilde{F}^*_{M2}|_X)^* : X^* \to \overline{\mathbb{R}}, \quad F_M = F_1 + F_{M2} : X^* \to \overline{\mathbb{R}},$$

where F_1 was introduced in subsection 3.3. Since the function l_M has the same essential properties as the function ln which occurs in the definition of the functional F_2, arguments similar to those in [11] give the following results.

LEMMA 3.14. *The functional $F_M = F_1 + F_{M2} : X^* \to \overline{\mathbb{R}}$ is proper, convex, and lower semicontinuous. For $u \in Y_+$ it can be evaluated according to* (3.7), (3.26). *The restriction $F_M|_{Y_+}$ is continuous. If $u \in Y_+$, then $P(\psi) \in \partial F_1(u)$, where ψ is the solution of $E(\psi,u) = 0$. If $u \in Y$, $u \geq \delta > 0$ and $u/p_0 \in X$, then $l_M(u/p_0) \in \partial F_{M2}(u)$, where $l_M(b)$ denotes the vector $\{l_M(b_i)\}_{i=1,\dots,m}$.*

By the definition of l_M the inequality

$$\int_{p_{0i}}^{y} l_{Mi}(z/p_{0i})\,\mathrm{d}z \geq y\ln\frac{y}{p_{0i}} - y + p_{0i} \quad \text{a.e. on } \Omega,\ \forall y \in \mathbb{R}_+$$

holds. Therefore it follows that

$$F_M(u) \geq c_1 \left\{ \|\psi\|_{H^1}^2 + \sum_{i=1}^m \|u_i \ln u_i\|_{L^1} \right\} - c_2, \ u \in Y_+. \tag{3.27}$$

LEMMA 3.15. *Along any solution (u, b, ψ) of $(\mathcal{P}_M)$ the regularized free energy $F_M(u)$ remains bounded by its initial value and decreases monotonously,*

$$F_M(u(t_2)) \leq F_M(u(t_1)) \leq F_M(U) = F(U), \quad 0 \leq t_1 \leq t_2 \leq T.$$

Moreover, there exist positive constants $c, c_{3.28}, c_{3.29}$ not depending on M, such that

$$\sum_{i=1}^m \|u_i \ln u_i\|_{L^\infty(S,L^1(\Omega))} \leq c, \quad \|u\|_{L^\infty(S,L^1(\Omega,\mathbb{R}^m))} \leq c,$$

$$\sum_{i=1}^m \|b_i \ln b_i\|_{L^\infty(S,L^1(\Omega))} \leq c_{3.28}, \tag{3.28}$$

$$\|\psi\|_{L^\infty(S,H^1(\Omega))} \leq c, \quad \|\psi\|_{L^\infty(S,L^\infty(\Omega))}, \quad \|\psi\|_{L^\infty(S,L^\infty(\Gamma))} \leq c_{3.29} \tag{3.29}$$

for any solution of $(\mathcal{P}_M)$.

Proof. Let (u, b, ψ) be a solution of $(\mathcal{P}_M)$.

1. We know that $u \in H^1(S, X^*)$, $\psi \in L^2(S, H^1(\Omega))$, $P(\psi) \in L^2(S, X)$, $\nabla P(\psi) = Q(\psi)\nabla\psi$. By Lemma 3.14 we find that $P(\psi(t)) \in \partial F_1(u(t))$ for a.a. $t \in S$. Thus, the function $t \mapsto F_1(u(t))$ is absolutely continuous on S, and according to the chain rule (see [3, Lemma 3.3]) we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t} F_1(u(t)) = \langle u'(t), P(\psi(t)) \rangle_X \quad \text{for a.a. } t \in S.$$

2. We choose some $\delta \in (0, 1)$ and define $u^\delta = u + \delta p_0$, $b^\delta = u^\delta / p_0 = b + \delta$. Then we find that $u^\delta \in H^1(S, X^*)$, $l_M(b^\delta) \in L^2(S, X)$, $\nabla l_M(b_i^\delta) = \nabla b_i / \sigma_M(b_i^\delta)$, $i = 1, \ldots, l$. Lemma 3.14 ensures that $l_M(b^\delta(t)) \in \partial F_{M2}(u^\delta(t))$ for a.a. $t \in S$. Thus, the function $t \mapsto F_{M2}(u^\delta(t))$ is absolutely continuous on S and

$$\frac{\mathrm{d}}{\mathrm{d}t} F_{M2}(u^\delta(t)) = \big\langle u'(t), l_M(b^\delta(t)) \big\rangle_X \quad \text{for a.a. } t \in S.$$

3. We set $\zeta_M^\delta = l_M(b^\delta) + P(\psi)$ and obtain

$$\begin{aligned} \big[F_1(u(t)) + F_{M2}(u^\delta(t))\big] \Big|_{t_1}^{t_2} &= \int_{t_1}^{t_2} \langle u'(t), \zeta_M^\delta(t) \rangle_X \, \mathrm{d}t \\ &= \int_{t_1}^{t_2} \langle R_N(b(t), \psi(t)) - A_M(b(t), \psi(t)), \zeta_M^\delta(t) \rangle_X \, \mathrm{d}t. \end{aligned}$$

The volume integral in the definition of $\langle R_N(b, \psi), \zeta_M^\delta \rangle_X$, namely,

$$I = \int_\Omega \rho_N(b, \psi) \sum_{(\alpha,\, \beta) \in \mathcal{R}^\Omega} k_{\alpha\beta}^\Omega(\cdot, b, \psi)(a^\alpha - a^\beta) \sum_{i=1}^m (\beta_i - \alpha_i) \zeta_{Mi}^\delta \, \mathrm{d}x, \ a_i = b_i \mathrm{e}^{P_i(\psi)},$$

is estimated as follows. Since for $|(b, \psi)|_\infty > N$ the integrand vanishes we may assume that $|(b, \psi)|_\infty \leq N$ and thus $b_i \leq N$, $b_i^\delta \leq N + 1 \leq M$, $\zeta_{Mi}^\delta = \ln a_i^\delta$ with $a_i^\delta = b_i^\delta \mathrm{e}^{P_i(\psi)}$,

$i = 1, \dots, m$, $|\psi| \le N$. Then we have

$$\big[(a^\delta)^\alpha - (a^\delta)^\beta\big] \sum_{i=1}^m (\beta_i - \alpha_i) \ln a_i^\delta \le 0,$$

$$\left| \big[a^\alpha - a^\beta - (a^\delta)^\alpha + (a^\delta)^\beta\big] \sum_{i=1}^m (\beta_i - \alpha_i) \ln a_i^\delta \right| \le c_N \delta (1 + |\ln \delta|)$$

and $I \le c_N \delta(1 + |\ln \delta|)$. The boundary integral is handled analogously and, in summary, we obtain

$$\langle R_N(b(t), \psi(t)), \zeta_M^\delta(t) \rangle_X \le h_1^\delta = c_N \delta (1 + |\ln \delta|) \quad \text{for a.a. } t \in S.$$

Next we consider the term $-\langle A_M(b, \psi), \zeta_M^\delta \rangle_X$, i.e., the integral

$$-\int_\Omega \sum_{i=1}^l D_{iM}(\cdot, b, \psi) p_{0i} (\nabla b_i + \sigma_M(b_i) Q_i(\psi) \nabla \psi) \cdot \nabla \zeta_{Mi}^\delta \, \mathrm{d}x.$$

Here we write

$$\nabla b_i + \sigma_M(b_i) Q_i(\psi) \nabla \psi = \sigma_M(b_i^\delta) \nabla \zeta_{Mi}^\delta + \big[\sigma_M(b_i) - \sigma_M(b_i^\delta)\big] Q_i(\psi) \nabla \psi,$$

and in view of $D_{iM}(\cdot, b, \psi) \le c_M$ we obtain

$$-\langle A_M(b(t), \psi(t)), \zeta_M^\delta(t) \rangle_X \le h_2^\delta(t) \quad \text{for a.a. } t \in S,$$

$$h_2^\delta(t) = c_M \int_\Omega \sum_{i=1}^l \delta |\nabla \psi(t)| \left[|\nabla \psi(t)| + \frac{1}{\sigma_M(b_i^\delta(t))} |\nabla b_i(t)| \right] \mathrm{d}x.$$

The last estimates ensure that

$$\big[F_1(u(t)) + F_{M2}(u^\delta(t))\big] \big|_{t_1}^{t_2} \le \int_{t_1}^{t_2} \big(h_1^\delta + h_2^\delta(t)\big) \, \mathrm{d}t,$$

and letting $\delta \to 0$, the inequality $F_M(u(t_2)) - F_M(u(t_1)) \le 0$ follows. The choice of M guarantees that $F_M(U) = F(U)$. The remaining assertions of the lemma are a consequence of (3.27), Lemma 3.1, and (5.2). ∎

3.8. Further estimates for solutions of ($\mathcal{P}_M$). We prove the following.

THEOREM 3.16. *There is a constant $c_{3.30} > 0$ not depending on M such that*

$$\|b\|_{L^\infty(S, L^\infty(\Omega, \mathbb{R}^m))} \le c_{3.30}, \quad \|b_i\|_{L^\infty(S, L^\infty(\Gamma))} \le c_{3.30}, \quad i = 1, \dots, l, \tag{3.30}$$

for any solution (u, b, ψ) of $(\mathcal{P}_M)$.

Proof. Let (u, b, ψ) be a solution of $(\mathcal{P}_M)$. Let $q > 2$ be chosen as in Lemma 3.1, $r = 2q/(q-2)$, $r' = 2q/(2+q)$. Other constants in the following estimates can depend on the data and on N, T.

1. Testing $(\mathcal{P}_M)$ with $(0, \dots, 0, b_{l+1}, \dots, b_m)$ we obtain in view of (3.5) that

$$\|b_i(t)\|_{L^2} \le c \quad \forall t \in S, \quad i = l+1, \dots, m, \tag{3.31}$$

which ensures that $\|u_i(t)\|_{L^{r'}} \le c \; \forall t \in S$, $i = l+1, \dots, m$. Hence we get

$$\|\psi(t)\|_{W^{1,q}} \le c \left[1 + \sum_{i=1}^m \|u_i(t)\|_{L^{r'}}\right] \le c \left[1 + \sum_{i=1}^l \|b_i(t)\|_{L^{r'}}\right] \quad \forall t \in S. \tag{3.32}$$

2. Testing $(\mathcal{P}_M)$ with $2(b_1,\dots,b_l,0,\dots,0)$, estimating $[\sigma_M(b_i)]^+$ by b_i, using (3.5), (3.32), (5.1), (5.3), Young's inequality, and Lemma 3.15, we find that

$$\begin{aligned}
&\sum_{i=1}^{l}\left(\epsilon_0\|b_i(t)\|_{L^2}^2 - c\|U_i\|_{L^2}^2\right)\\
&\quad\le \int_0^t \sum_{i=1}^{l}\{-2\epsilon\|b_i\|_{H^1}^2 + c(\|b_i\|_{L^r}\|\psi\|_{W^{1,q}}\|b_i\|_{H^1} + \|b_i\|_{L^2}^2 + \|b_i\|_{L^2(\Gamma)}^2 + 1)\}\mathrm{d}s\\
&\quad\le \int_0^t \sum_{i=1}^{l}\left\{-\epsilon\|b_i\|_{H^1}^2 + \overline{c}\|b_i\|_{L^r}\|b_i\|_{H^1}\sum_{k=1}^{l}\|b_k\|_{L^{r'}} + c\right\}\mathrm{d}s.
\end{aligned}$$

Using the inequality (5.4) for $p=2$ and Lemma 3.15 we have

$$\begin{aligned}
\overline{c}\sum_{i=1}^{l}\|b_i\|_{L^r}\|b_i\|_{H^1}\sum_{k=1}^{l}\|b_k\|_{L^{r'}} &\le \sum_{i=1}^{l}\left\{\frac{\epsilon}{2}\|b_i\|_{H^1}^2 + c\|b_i\|_{L^2}^2\sum_{k=1}^{l}\|b_k\|_{L^2}^2\right\}\\
&\le \sum_{i=1}^{l}\left\{\frac{\epsilon}{2}\|b_i\|_{H^1}^2 + \left[\frac{\sqrt{\epsilon}}{2c_{3.28}}\|b_i\ln b_i\|_{L^1}\|b_i\|_{H^1} + c\|b_i\|_{L^1}\right]^2\right\} \le \sum_{i=1}^{l}\epsilon\|b_i\|_{H^1}^2 + c.
\end{aligned}$$

The previous estimates and the inequalities (3.31), (3.32) ensure the existence of positive constants $c,\widetilde{\kappa}$ not depending on M such that

$$\|b_i(t)\|_{L^2}\le c,\ i=1,\dots,m,\ \|\psi(t)\|_{W^{1,q}}^{2r}+1\le\widetilde{\kappa}\quad\forall t\in S. \tag{3.33}$$

3. Following the estimates in the proofs of Lemma 3.6 and Theorem 3.7, but estimating $[\sigma_M(b_i)]^+$ by b_i and using $\widetilde{\kappa}$ from (3.33) instead of κ, we find that

$$\begin{aligned}
&\|b_i(t)\|_{L^4}\le\widetilde{c},\ i=1,\dots,m,\\
&\sum_{i=1}^{l}\|(b_i-K)^+(t)\|_{L^\infty}\le\sqrt{l}\left(2^{4r}\left(\widetilde{\kappa}+\widetilde{c}^8\right)cT\left(\sum_{i=1}^{l}\sup_{s\in S}\|(b_i-K)^+(s)\|_{L^4}^4+1\right)\right)^{c_0},\\
&\|(b_i-K)^+(t)\|_{L^\infty}\le c,\ i=l+1,\dots,m,\quad\forall t\in S,
\end{aligned}$$

where the constants K, c_0 have the same meaning as in the proof of Theorem 3.7. This provides the desired estimates. ∎

3.9. Existence result for $(\mathcal{P}_N)$.

THEOREM 3.17. *There exists a solution of* $(\mathcal{P}_N)$.

Proof. Let $N>0$ be fixed. We choose $M=\max\{M^*,c_{3.29},c_{3.30}\}$ (see (3.19), Lemma 3.15, and Theorem 3.16). By Theorem 3.12 there is a solution (u,b,ψ) of $(\mathcal{P}_M)$. Since $b\ge 0$ (see Lemma 3.13) and

$$\begin{aligned}
&\|\psi\|_{L^\infty(S,L^\infty(\Omega))},\ \|\psi\|_{L^\infty(S,L^\infty(\Gamma))}\le M,\\
&\|b_i\|_{L^\infty(S,L^\infty)}\le M,\ i=1,\dots,m,\quad \|b_i\|_{L^\infty(S,L^\infty(\Gamma))}\le M,\ i=1,\dots,l
\end{aligned}$$

(see Lemma 3.15 and Theorem 3.16) this solution is a solution of $(\mathcal{P}_N)$, too. ∎

3.10. Existence result for $(\mathcal{P})$.

Proof of Theorem 2.3. It suffices to prove the existence of a solution of $(\mathcal{P})$ on any finite time interval $S=[0,T]$. Such problems are denoted by $(\mathcal{P}_S)$. We choose $N=2\max\{c_{3.13},c_{3.18}\}$ (see Lemma 3.3 and Theorem 3.7). Then according to Theorem 3.17 there is a solution (u,b,ψ) of $(\mathcal{P}_N)$. The choice of N guarantees that $R_N(b,\psi)=R(b,\psi)$. Therefore (u,b,ψ) is a solution of $(\mathcal{P}_S)$, too. ∎

4. Remarks. For the sake of completeness we summarize here some results of our earlier papers [11, 16] concerned with qualitative properties of solutions of $(\mathcal{P})$, the existence of which is now established by the results of the present paper.

4.1. Uniqueness. Under the restrictive assumption that

$$\begin{cases} D_i\colon \Omega \times \mathbb{R} \to \mathbb{R}_+ \text{ does not depend on } b, \\ D_i(x,\cdot) \text{ is locally Lipschitz continuous uniformly w.r.t. } x,\ i = 1,\dots,l, \end{cases} \tag{4.1}$$

the solution of $(\mathcal{P})$ is unique (see [11, Lemma 7.2]).

4.2. Steady states. We regard the set $\mathcal{R}^\Gamma \subset \mathbb{Z}_+^l \times \mathbb{Z}_+^l$ as a subset of $\mathbb{Z}_+^m \times \mathbb{Z}_+^m$ by setting $\alpha_i = \beta_i = 0$, $i = l+1,\dots,m$, $(\alpha,\beta) \in \mathcal{R}^\Gamma$, and introduce the stoichiometric subspace $\mathcal{S}$ belonging to all reactions,

$$\mathcal{S} = \operatorname{span}\{\alpha - \beta : (\alpha,\beta) \in \mathcal{R}^\Omega \cup \mathcal{R}^\Gamma\} \subset \mathbb{R}^m.$$

Every solution of $(\mathcal{P})$ fulfills the invariance property $\int_\Omega \{u(t) - U\}\,\mathrm{d}x \in \mathcal{S}$, $t \in \mathbb{R}_+$. Therefore we consider the following stationary problem:

$$\begin{cases} A(b,\psi) = R(b,\psi),\ E(\psi,u) = 0,\ u = Bb,\ \displaystyle\int_\Omega \{u - U\}\,\mathrm{d}x \in \mathcal{S}, \\ u \in Y, \quad b \in X \cap L^\infty_+(\Omega,\mathbb{R}^m), \quad \psi \in H^1(\Omega) \cap L^\infty(\Omega). \end{cases} \tag{S}$$

In addition to the assumptions in subsection 2.2 we suppose that

$$\begin{cases} \displaystyle\int_\Omega U \cdot \overline{\zeta}\,\mathrm{d}x > 0 \quad \forall \overline{\zeta} \in \mathcal{S}^\perp,\ \overline{\zeta} \ge 0,\ \overline{\zeta} \ne 0; \\ \Big\{ a \in \mathbb{R}^m_+ :\ a^\alpha = a^\beta\ \forall (\alpha,\beta) \in \mathcal{R}^\Omega \cup \mathcal{R}^\Gamma,\ \displaystyle\int_\Omega (u - U)\,\mathrm{d}x \in \mathcal{S},\ \text{where} \\ u = ap(\cdot,\psi) \text{ and } \psi \text{ is the solution of } E(\psi,u) = 0 \Big\} \cap \partial\mathbb{R}^m_+ = \emptyset; \\ k^\Sigma_{\alpha\beta}(x,b,\psi) \ge b^\Sigma_{\alpha\beta,R}(x) \text{ for a.a. } x \in \Sigma,\ \forall b \in \mathbb{R}^{m\Sigma}_+,\ \forall \psi \in [-R,R],\ R > 0, \\ \text{where } b^\Sigma_{\alpha\beta,R} \in L^\infty_+(\Sigma),\ \|b^\Sigma_{\alpha\beta,R}\|_{L^1(\Sigma)} > 0 \quad \forall (\alpha,\beta) \in \mathcal{R}^\Sigma,\ \Sigma = \Omega,\Gamma. \end{cases} \tag{4.2}$$

Here $p(\cdot,\psi)$ denotes the vector $p_i(\cdot,\psi) = p_{0i}\,\mathrm{e}^{-P_i(\psi)}$, $i = 1,\dots,m$. Then the stationary problem (S) has a unique solution (u^*,b^*,ψ^*), which is a thermodynamic equilibrium (see [11, Theorem 3.1]).

4.3. Energy estimates. The free energy functional F related to problem $(\mathcal{P})$ is again given by (3.7)–(3.9), while the dissipation functional D now reads as

$$\begin{aligned} D(u) = {} & 4\int_\Omega \sum_{i=1}^{l} D_i(\cdot,b,\psi) p_i(\cdot,\psi) |\nabla \sqrt{a_i}|^2\,\mathrm{d}x \\ & + 4\int_\Omega \sum_{(\alpha,\beta)\in\mathcal{R}^\Omega} k^\Omega_{\alpha\beta}(\cdot,b_1,\dots,b_m,\psi) |\sqrt{a}^{\,\alpha} - \sqrt{a}^{\,\beta}|^2\,\mathrm{d}x \\ & + 4\int_\Gamma \sum_{(\alpha,\beta)\in\mathcal{R}^\Gamma} k^\Gamma_{\alpha\beta}(\cdot,b_1,\dots,b_l,\psi) |\sqrt{a}^{\,\alpha} - \sqrt{a}^{\,\beta}|^2\,\mathrm{d}\Gamma, \quad u \in M_D, \end{aligned} \tag{4.3}$$

where the set M_D is the same as in (3.11). The free energy F is a Lyapunov function for problem $(\mathcal{P})$; more precisely, the inequality (3.12) holds for any solution of $(\mathcal{P})$ where D has to be replaced by the expression (4.3) (see [11, Theorem 3.2]). Using the additional assumption (4.2) the following estimate of the free energy by the dissipation rate can be derived (see [16, Theorem 3]):

$$F(u)-F(u^*)\le c(R)D(u)\quad \forall u\in\left\{u\in M_D\colon F(u)-F(u^*)\le R,\ \int_\Omega (u-U)\,\mathrm{d}x\in\mathcal{S}\right\}.$$

This is a logarithmic Sobolev inequality adapted to the structure of $(\mathcal{P})$ and guarantees the exponential decay of the free energy along any solution of $(\mathcal{P})$,

$$0\le F(u(t))-F(u^*)\le \mathrm{e}^{-\lambda t}\,(F(U)-F(u^*))\quad \forall t\in\mathbb{R}_+$$

(see [11, Theorem 3.3]).

4.4. Global upper and lower bounds. Again under the additional assumption (4.2) there exists a constant $c>0$ such that

$$\|u_i(t)\|_{L^\infty},\ \|b_i(t)\|_{L^\infty},\ \|\psi(t)\|_{L^\infty}\le c\quad \forall t\in\mathbb{R}_+,\ i=1,\dots,m,$$

if (u,b,ψ) is a solution of $(\mathcal{P})$ (see [11, Theorem 4.1, Theorem 3.2]). Moreover, if the initial densities are strictly positive, $U_i\ge c_0>0$, $i=1,\dots,m$, then we find a constant $c>0$ such that

$$\operatorname{ess\,inf}_{x\in\Omega} u_i(t)\ge c>0\quad \forall t\in\mathbb{R}_+,\ i=1,\dots,m,$$

for any solution of $(\mathcal{P})$ (see [11, Corollary 5.1]).

4.5. Asymptotic behavior. Let q be fixed as in Lemma 3.1, $p\in[1,+\infty)$. Then there exist constants $c,\lambda_p,\lambda>0$ such that

$$\begin{aligned}&\|u(t)-u^*\|_{L^p(\Omega,\mathbb{R}^m)},\ \|b(t)-b^*\|_{L^p(\Omega,\mathbb{R}^m)}\le c\,\mathrm{e}^{-\lambda_p t}\quad \forall\, t\in\mathbb{R}_+,\\ &\|\psi(t)-\psi^*\|_{W^{1,q}},\ \|\psi(t)-\psi^*\|_{L^\infty}\le c\,\mathrm{e}^{-\lambda t}\quad \forall\, t\in\mathbb{R}_+\end{aligned}$$

if (u,b,ψ) is a solution of $(\mathcal{P})$ (see [11, Theorem 6.1]; again (4.2) has to be assumed).

4.6. Example. Let us consider the example in subsection 2.3. Here we have

$$\mathcal{S}^\perp=\operatorname{span}\{(0,0,1,1,1),(1,-1,1,-1,0)\}\subset\mathbb{R}^5,$$

and the first assumption in (4.2) is fulfilled if we require that

$$\int_\Omega (U_3+U_4+U_5)\,\mathrm{d}x>0.$$

Then the second assumption in (4.2) is fulfilled, too. In general, this assumption means that the underlying reaction system has no false equilibria in the sense of [23].

5. Appendix. We assume that $\Omega\subset\mathbb{R}^2$ is a bounded Lipschitzian domain. We apply Sobolev's imbedding theorems (see [18]) and some other imbedding results, especially the trace inequalities

$$\|w\|^q_{L^q(\Gamma)}\le c_\Omega\, q\,\|w\|^{q-1}_{L^{2(q-1)}(\Omega)}\|w\|_{H^1(\Omega)}\quad \forall w\in H^1(\Omega),\ q\ge 2, \tag{5.1}$$

$$\|w\|_{L^\infty(\Gamma)} \le \|w\|_{L^\infty(\Omega)} \quad \forall w \in H^1(\Omega) \cap L^\infty(\Omega), \tag{5.2}$$

and the following version of the Gagliardo–Nirenberg inequality (see [5, 22]):

$$\|w\|_{L^p} \le c_{p,k} \|w\|_{L^k}^{k/p} \|w\|_{H^1}^{1-k/p} \quad \forall w \in H^1(\Omega),\ 1 \le k < p < \infty. \tag{5.3}$$

As an extended form of this inequality one obtains that for any $\delta > 0$ and any $p \in (1, \infty)$ there exists a constant $c_{\delta,p} > 0$ such that

$$\|w\|_{L^p}^p \le \delta \|w \ln |w|\|_{L^1} \|w\|_{H^1}^{p-1} + c_{\delta,p} \|w\|_{L^1} \quad \forall w \in H^1(\Omega). \tag{5.4}$$

This inequality is proven in [2] for bounded domains with smooth boundary and $p = 3$. But (5.4) is valid also for bounded Lipschitzian domains and $p \in (1, \infty)$, since (5.3) is true in this case, too.

Let $p_0 \in L^\infty(\Omega)$, $\operatorname{ess\,inf}_{x\in\Omega} p_0(x) > 0$. We define $B\colon L^2(\Omega) \longrightarrow L^2(\Omega)$ by

$$(Bw, \overline{w})_{L^2} = \int_\Omega p_0\, w\, \overline{w} \mathrm{d}x, \quad \overline{w} \in L^2(\Omega).$$

Let $S = [0, T]$ be a compact interval. The extended operator $B\colon L^2(S, L^2(\Omega)) \longrightarrow L^2(S, L^2(\Omega))$ is defined by $(Bw)(t) = B(w(t))$ for a.a. $t \in S$. Because of properties of p_0 the operator B is linear, continuous, self-adjoint, and positive definite, and there exists the inverse operator $B^{-1}\colon L^2(S, L^2(\Omega)) \longrightarrow L^2(S, L^2(\Omega))$ with the same properties. Let

$$W_B = \{w \in L^2(S, H^1)\colon (Bw)' \in L^2(S, H^{1*})\}.$$

The following assertions can be verified as in [6, 19, 24].

LEMMA 5.1.

(i) *Equipped with the scalar product*

$$(w, \overline{w})_{W_B} = (w, \overline{w})_{L^2(S,H^1)} + ((Bw)', (B\overline{w})')_{L^2(S,H^{1*})},$$

the linear space W_B is a Hilbert space.

(ii) *W_B is continuously embedded in $C(S, L^2(\Omega))$.*

(iii) *The operator $B\colon W_B \to C(S, L^2(\Omega))$ is continuous.*

(iv) *For $w \in W_B$ and $t_1, t_2 \in S$ the formula*

$$\int_{t_1}^{t_2} \big\langle (Bw)'(s), w(s) \big\rangle_{H^1} \mathrm{d}s = \frac{1}{2} \left((Bw)(t_2), w(t_2)\right)_{L^2} - \frac{1}{2} \left((Bw)(t_1), w(t_1)\right)_{L^2}$$

holds.

(v) *The imbeddings of W_B in $L^2(S, L^2(\Omega))$ and in $L^2(S, L^2(\Gamma))$, respectively, are compact.*

Finally, the following existence result can be obtained as in [6, Chapter VI].

LEMMA 5.2. *Let $A\colon L^2(S, H^1(\Omega)) \to L^2(S, H^1(\Omega)^*)$ be the operator*

$$\langle Aw, \overline{w} \rangle_{L^2(S,H^1)} = \int_0^T \int_\Omega (a\, \nabla w \cdot \nabla \overline{w} + d\, w\, \overline{w}) \mathrm{d}x \mathrm{d}s,\ w, \overline{w} \in L^2(S, H^1(\Omega)),$$

where $a, d \in L^\infty(S \times \Omega)$ with $a(t,x), d(t,x) \ge c > 0$ a.e. on $S \times \Omega$. Then for every $f \in L^2(S, H^1(\Omega)^*)$ and every $U \in L^2(\Omega)$ there is a unique solution of the problem

$$(Bw)' + Aw = f, \quad (Bw)(0) = U, \quad w \in W_B.$$

REFERENCES

[1] R. Bader and W. Merz, *Local existence result of the single dopant diffusion in arbitrary space dimension*, Z. Anal. Anwendungen, 21 (2002), pp. 91–111.

[2] P. Biler, W. Hebisch, and T. Nadzieja, *The Debye system: Existence and large time behavior of solutions*, Nonlinear Anal., 23 (1994), pp. 1189–1209.

[3] H. Brézis, *Opérateurs maximaux monotones et semi-groups de contractions dans les espaces de Hilbert*, North-Holland Math. Stud. 5, North-Holland, Amsterdam, 1973.

[4] S. T. Dunham, *A quantitative model for the coupled diffusion of phosphorus and point defects in silicon*, J. Electrochem. Soc., 139 (1992), pp. 2628–2636.

[5] E. Gagliardo, *Ulteriori proprietà di alcune classi di funzioni in più variabili*, Ricerche Mat., 8 (1959), pp. 24–51.

[6] H. Gajewski, K. Gröger, and K. Zacharias, *Nichtlineare Operatorgleichungen und Operatordifferentialgleichungen*, Akademie-Verlag, Berlin, 1974.

[7] K. Ghaderi and G. Hobler, *Simulation of phosphorus diffusion in silicon using a pair diffusion model with a reduced number of parameters*, J. Electrochem. Soc., 142 (1995), pp. 1654–1658.

[8] A. Glitzky and R. Hünlich, *Energy estimates and asymptotics for electro-reaction-diffusion systems*, Z. Angew. Math. Mech., 77 (1997), pp. 823–832.

[9] A. Glitzky and R. Hünlich, *Global estimates and asymptotics for electro-reaction-diffusion systems in heterostructures*, Appl. Anal., 66 (1997), pp. 205–226.

[10] A. Glitzky and R. Hünlich, *Electro-reaction-diffusion systems including cluster reactions of higher order*, Math. Nachr., 216 (2000), pp. 95–118.

[11] A. Glitzky and R. Hünlich, *Global properties of pair diffusion models*, Adv. Math. Sci. Appl., 11 (2001), pp. 293–321.

[12] A. Glitzky and W. Merz, *Single dopant diffusion in semiconductor technology*, Math. Methods Appl. Sci., 27 (2004), pp. 133–154.

[13] A. Glitzky, K. Gröger, and R. Hünlich, *Free energy and dissipation rate for reaction diffusion processes of electrically charged species*, Appl. Anal., 60 (1996), pp. 201–217.

[14] K. Gröger, *A $W^{1,p}$-estimate for solutions to mixed boundary value problems for second order elliptic differential equations*, Math. Ann., 283 (1989), pp. 679–687.

[15] S. L. Hollis and J. J. Morgan, *Partly dissipative reaction-diffusion systems and a model of phosphorus diffusion in silicon*, Nonlinear Anal., 19 (1992), pp. 427–440.

[16] R. Hünlich and A. Glitzky, *On energy estimates for electro-diffusion equations arising in semiconductor technology*, in Partial Differential Equations. Theory and Numerical Solution, W. Jäger, J. Nečas, O. John, K. Najzar, and J. Stará, eds., Chapman & Hall/CRC Res. Notes Math. 406, Chapman & Hall/CRC, Boca Raton, FL, 2000, pp. 158–174.

[17] J. R. King, *Asymptotic analysis of a model for the diffusion of dopant-defect pairs*, in Semiconductors. Part I, W. M. Coughran, Jr., J. Cole, P. Lloyd, and J. K. White, eds., IMA Vol. Math. Appl. 58, Springer, New York, 1994, pp. 49–66.

[18] A. Kufner, O. John, and S. Fučik, *Function Spaces*, Academia, Prague, 1977.

[19] J. L. Lions, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod Gauthier-Villars, Paris, 1969.

[20] B. Margesin, R. Canteri, S. Solmi, A. Armigliato, and F. Baruffaldi, *Boron and antimony codiffusion in silicon*, J. Mater. Res., 6 (1991), pp. 2353–2361.

[21] W. Merz, A. Glitzky, R. Hünlich, and K. Pulverer, *Strong solutions for pair diffusion models in homogeneous semiconductors*, Nonlinear Anal. Real World Appl., 2 (2001), pp. 541–567.

[22] L. Nirenberg, *An extended interpolation inequality*, Ann. Scuola Norm. Sup. Pisa, 20 (1966), pp. 733–737.

[23] I. Prigogine and R. Defay, *Chemical Thermodynamics*, Longmans, London, 1954.

[24] R. Temam, *Navier–Stokes Equations. Theory and Numerical Analysis*, Stud. Math. Appl. 2, North-Holland, Amsterdam, 1979.

[25] N. S. Trudinger, *On imbeddings into Orlicz spaces and some applications*, J. Math. Mech., 17 (1967), pp. 473–483.

[26] W. Walter, *Differential and Integral Inequalities*, Ergeb. Math. Grenzgeb. 55, Springer, Berlin, 1970.

SIAM J. MATH. ANAL.
Vol. 36, No. 4, pp. 1226–1258

A SEMILINEAR FOURTH ORDER ELLIPTIC PROBLEM WITH EXPONENTIAL NONLINEARITY*

GIANNI ARIOLI†, FILIPPO GAZZOLA†, HANS-CHRISTOPH GRUNAU‡, AND ENZO MITIDIERI§

Abstract. We study a semilinear fourth order elliptic problem with exponential nonlinearity. Motivated by a question raised in [P.-L. Lions, *SIAM Rev.*, 24 (1982), pp. 441–467], we partially extend results known for the corresponding second order problem. Several new difficulties arise and many problems still remain to be solved. We list those of particular interest in the final section.

Key words. biharmonic operator, super-subsolutions, computer assisted proof

AMS subject classifications. 35G30, 35J40

DOI. 10.1137/S0036141002418534

1. Introduction. In the last forty years a great deal has been written about existence and multiplicity of solutions to nonlinear second order elliptic problems in bounded and unbounded domains of $\mathbb{R}^n$ ($n \geq 2$). Important achievements on this topic have been made by applying various combinations of analytical techniques, and among all of them we mention only the variational and topological methods. For the latter, especially when the main interest is focused on the existence of positive solutions, the fundamental tool which has been used is the maximum principle [A1] and its consequences [GNN].

For higher order problems, a possible failure of the maximum principle causes several technical difficulties. This fact is very likely the reason why the knowledge on higher order nonlinear problems is far from being reasonably complete, as it is in the second order case.

One of the most interesting and intensively studied second order model problems that exhibits several peculiar features of most nonlinear elliptic equations is the so-called Gel'fand problem [G, section 15],

$$\begin{cases} -\Delta u = \lambda e^u & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases} \tag{1}$$

Here Ω is a bounded smooth domain in $\mathbb{R}^n$ ($n \geq 3$) and $\lambda \geq 0$ is a parameter. This problem appears in connection with combustion theory [G, JL] and stellar structure [C]. From a mathematical point of view, one of the main interests is that it may have both unbounded (singular) solutions and bounded (regular) solutions (see [BV, GMP, MP1, MP2]): by the results in [CR, BCMR] it is known that there exists $\lambda^* > 0$ such that if $\lambda > \lambda^*$ there exists no solution of (1) (neither regular nor singular),

*Received by the editors November 25, 2002; accepted for publication (in revised form) April 23, 2004; published electronically February 3, 2005.
http://www.siam.org/journals/sima/36-4/41853.html
†Dipartimento di Matematica del Politecnico, via Bonardi 9, 20133 Milano, Italy (gianni@mate.polimi.it, gazzola@mate.polimi.it). The work of these authors was supported by the MURST project "Metodi Variazionali ed Equazioni Differenziali non Lineari."
‡Fakultät für Mathematik, Otto-von-Guericke-Universität, Postfach 4120, 39016 Magdeburg, Germany (Hans-Christoph.Grunau@mathematik.uni-magdeburg.de).
§Dipartimento di Scienze Matematiche, via A. Valerio 12/1, 34100 Trieste, Italy (mitidier@univ.trieste.it).

while if $0 \le \lambda < \lambda^*$, there exists a minimal regular solution U_λ of (1) and the map $\lambda \mapsto U_\lambda$ is smooth and increasing. In the unit ball B, the bifurcation picture of radial solutions is rather complete. There is always a singular solution $u_\sigma := -2\log|x|$ with corresponding parameter $\lambda_\sigma = 2(n-2)$. If $n \ge 10$, the solution branch consists only of minimal solutions and terminates at $\lambda^* = \lambda_\sigma$ in the singular solution. If $3 \le n \le 9$, then $\lambda^* > \lambda_\sigma$ and the extremal point (λ^*, U_*) is a turning point. The branch bends back and meanders infinitely many times around λ_σ while approaching the singular solution u_σ. We refer to [BV, Figure 1] for the pictures. The interested reader may see also [BE] for an account on motivations and related results.

Some interesting generalizations of (1) have been considered in the framework of second order quasi-linear operators. We refer to [GPP] for equations associated to the p-Laplace operator and to [J, JS] for the case of the k-Hessian operator.

The aim of this paper is to give a contribution to the solution of a special case of a problem formulated in [Li, section 4.2 (c)], namely, *Is it possible to obtain a description of the solution set for higher order semilinear equations associated to exponential nonlinearities?*

Recently, interest in higher order nonlinear problems due to its exciting and promising developments has become increasingly evident especially for fourth order equations [PT]. Following this trend, we shall consider in this paper the fourth order version of (1), a semilinear elliptic problem which involves the biharmonic operator, more precisely,

$$(P_\lambda) \qquad \begin{cases} \Delta^2 u = \lambda e^u & \text{in } B, \\ u = \dfrac{\partial u}{\partial \mathbf{n}} = 0 & \text{on } \partial B. \end{cases}$$

Here B denotes the unit ball in $\mathbb{R}^n$ ($n \ge 5$) centered at the origin and $\frac{\partial}{\partial \mathbf{n}}$ the differentiation with respect to the exterior unit normal, i.e., in radial direction; $\lambda \ge 0$ is a parameter. We are interested in two kinds of solutions of (P_λ), regular solutions and singular solutions; see Definition 1 in the next section. We restrict our attention to the case $n \ge 5$, where the nonlinearity is supercritical. In low dimensions $1 \le n \le 4$ the problem is subcritical and has a different behavior; see Remark 14 at the end of the following section.

Many techniques, familiar from second order equations like the maximum principle, are not available here. But since we restrict ourselves to the ball, at least a comparison principle is available; see Lemma 16 below. Moreover, in fourth order equations, one usually does not succeed in finding suitable nontrivial auxiliary functions satisfying again a differential inequality. This is a serious difficulty in proving Theorem 3 (cf. the proof of [BCMR, Theorem 3]), and it is overcome by carefully exploiting the properties of the exponential nonlinearity and the construction of minimal solutions, based upon the already mentioned comparison principle. Finally, when looking for radial solutions, one may perform a phase space analysis for the corresponding system of ODEs. Here, the phase space is no longer two-dimensional, where the topology is relatively simple and the Poincaré–Bendixson theory is available, but we have to work in a four-dimensional phase space. Some of the resulting difficulties could be overcome only with computer assistance.

This paper is organized as follows: In the next section we state some definitions and the main results contained in this work (see Theorems 3, 4, 6, 7, and 12 below). The content of sections 3 through 7 is devoted to the proofs of these theorems. Section 8 contains some results on the stability of regular solutions of (P_λ) and a list

of open problems that we consider of some interest and related to the main results of this paper. Finally, in section 9 we describe the algorithm used in the *computer assisted proof* of Theorem 7.

2. Main results. We first make precise in which sense we intend a function to solve (P_λ). For this purpose, we fix some exponent p with $p > \frac{n}{4}$ and $p \geq 2$. The definitions and results below do not depend on the special choice of p.

DEFINITION 1. *We say that $u \in L^2(B)$ is a* solution *of (P_λ) if $e^u \in L^1(B)$ and*

$$\int_B u\Delta^2 v = \lambda \int_B e^u v \qquad \text{for all } v \in W^{4,p} \cap H_0^2(B). \tag{2}$$

We say that a solution u of (P_λ) is regular *(resp.,* singular*) if $u \in L^\infty(B)$ (resp., $u \notin L^\infty(B)$).*

Clearly, according to this definition, regular and singular solutions exhaust all possible solutions. Note that by standard regularity theory for the biharmonic operator (see [ADN]), any regular solution u of (P_λ) satisfies $u \in C^\infty(\overline{B})$. Note also that by the positivity preserving property of Δ^2 in the ball [B] any solution of (P_λ) is positive; see also Lemmas 16 and 18 below for a generalized statement. This property is known to fail in general domains. For this reason, we restrict ourselves to balls also in Theorems 3 and 4; cf. also Open Problem 8 in section 8.

We also need the notion of minimal solution, as follows.

DEFINITION 2. *We call a solution U_λ of (P_λ)* minimal *if $U_\lambda \leq u_\lambda$ a.e. in B for any further solution u_λ of (P_λ).*

In order to state our results, we denote by $\lambda_1 > 0$ the first eigenvalue for the biharmonic operator with Dirichlet boundary conditions

$$\begin{cases} \Delta^2 u = \lambda_1 u & \text{in } B, \\ u = \dfrac{\partial u}{\partial \mathbf{n}} = 0 & \text{on } \partial B; \end{cases} \tag{3}$$

it is known from the mentioned positivity preserving property and Jentzsch's (or Krein–Rutman's) theorem that λ_1 is isolated and simple and that the corresponding eigenfunctions do not change sign.

We may now state the following theorem.

THEOREM 3. *There exists*

$$\lambda^* \in \left[14.72(n-1)(n-3), \frac{\lambda_1}{e}\right)$$

such that the following hold:

(i) *(P_λ) admits a minimal regular solution U_λ for all $\lambda < \lambda^*$ and no solutions if $\lambda > \lambda^*$.*

(ii) *The map $\lambda \mapsto U_\lambda(x)$ is strictly increasing for all $x \in B$. Moreover, there exists a solution U_* of (P_{λ^*}) which is the pointwise limit of U_λ as $\lambda \uparrow \lambda^*$.*

(iii) *$U_\lambda \to U_*$ in the norm topology of $H_0^2(B)$ as $\lambda \uparrow \lambda^*$.*

(iv) *The extremal solution U_* and all the minimal solutions U_λ (for $\lambda < \lambda^*$) are radially symmetric and radially decreasing.*

It is remarkable that at λ^* there is an immediate switch from existence of *regular* minimal solutions to nonexistence of *any* (even singular) solution. The only possibly singular minimal solution corresponds to $\lambda = \lambda^*$. This result is known from [BCMR] for the second order problem (1), but the *method* used there may not be carried over

to fourth order problems. Nevertheless, the *result* extends to the biharmonic case. The proof is given in Lemma 20 below.

We may also characterize the uniform convergence to 0 of U_λ as $\lambda \to 0$ by giving the precise rate of its extinction.

THEOREM 4. *For all $\lambda \in (0, \lambda^*)$ let U_λ be the minimal solution of (P_λ) and let*

$$V_\lambda(x) = \frac{\lambda}{8n(n+2)} \left[1 - |x|^2\right]^2.$$

Then $U_\lambda(x) > V_\lambda(x)$ for all $\lambda < \lambda^*$ and all $|x| < 1$, and

$$\lim_{\lambda \to 0} \frac{U_\lambda(x)}{V_\lambda(x)} = 1 \qquad \textit{uniformly with respect to } x \in B.$$

A complete result in the spirit of Gidas, Ni, and Nirenberg [GNN] does not hold for fourth order equations under Dirichlet boundary conditions. It has been recently proved by Sweers in [Sw] that for general semilinear autonomous biharmonic equations in a ball under Dirichlet boundary conditions, we may have positive radially symmetric solutions which are not radially decreasing, provided the right-hand side is *not* positive everywhere. This phenomenon may not occur in our situation; however, it is not known whether any smooth solution of (P_λ) is radially symmetric. Moreover, also in the second order case it is not known whether singular solutions are always radially symmetric. Nevertheless, Theorem 3 suggests that we pay particular attention to radially symmetric solutions. In this context, we put $r = |x|$ and consider the functions $u = u(r)$.

First of all, in the following definition we introduce a new notion of solution which seems to be the natural framework for radially symmetric solutions.

DEFINITION 5. *We say that a radial singular solution $u = u(r)$ of (P_λ) is* weakly singular *if the limit $\lim_{r \to 0} r u'(r)$ exists.*

We do not know whether every singular solution is also weakly singular. In the second order case, Joseph and Lundgren [JL] reduce (1) to a system of two ODEs and study its phase portrait in $\mathbb{R}^2$; using Bendixson's theorem, they show that singular solutions are also weakly singular. For the fourth order equation (P_λ) a similar argument should be carried out in $\mathbb{R}^4$ (see section 3) where a general result of Bendixson's type does not hold. Therefore, the equivalence between singular and weakly singular solutions seems out of reach in our context; see Open Problem 5 in section 8.

If we seek radially symmetric solutions, we rewrite problem (P_λ) as $(0 < r \le 1)$

$$(4) \qquad \begin{cases} \dfrac{d^4u}{dr^4} + \dfrac{2(n-1)}{r}\dfrac{d^3u}{dr^3} + \dfrac{(n-1)(n-3)}{r^2}\dfrac{d^2u}{dr^2} - \dfrac{(n-1)(n-3)}{r^3}\dfrac{du}{dr} = \lambda e^{u(r)}, \\ u(1) = 0, \\ \dfrac{du}{dr}\,|_{r=1} = 0. \end{cases}$$

In [GPP, JL, MP2] the second order equation (1) was reduced to a system of two autonomous ODEs. Here, we reduce (4) to a system of four equations. First, we make the change of variables

$$(5) \qquad s = \log r, \qquad v(s) = u(e^s), \qquad s \in (-\infty, 0]$$

so that (4) becomes

$$\begin{cases} \dfrac{d^4v}{ds^4} + 2(n-4)\dfrac{d^3v}{ds^3} + (n^2-10n+20)\dfrac{d^2v}{ds^2} - 2(n-2)(n-4)\dfrac{dv}{ds} = \lambda e^{4s+v(s)}, \\ v(0) = 0, \\ \dfrac{dv}{ds}\,|_{s=0} = 0; \end{cases} \tag{6}$$

then we set

$$\begin{cases} v_1(s) = v'(s) + 4, \\ v_2(s) = -v''(s) - (n-2)v'(s), \\ v_3(s) = -v'''(s) + (4-n)v''(s) + 2(n-2)v'(s), \\ v_4(s) = -\lambda e^{v(s)+4s}. \end{cases} \tag{7}$$

Finally, we obtain the following (nonlinear) differential system:

$$\begin{cases} v_1'(s) = (2-n)v_1(s) - v_2(s) + 4(n-2), \\ v_2'(s) = 2v_2(s) + v_3(s), \\ v_3'(s) = (4-n)v_3(s) + v_4(s), \\ v_4'(s) = v_1(s)v_4(s) \end{cases} \tag{8}$$

with initial conditions

$$v_1(0) = 4, \qquad v_4(0) = -\lambda. \tag{9}$$

It turns out that (8) admits only the two stationary points $P_1 = (4,0,0,0)$ and $P_2 = (0, 4n-8, 16-8n, -8(n-2)(n-4))$; see section 3.1. Then, in section 3.2, we prove the following result.

THEOREM 6. *Let $u = u(r)$ be a radial solution of (P_λ) and let*

$$V(s) = (v_1(s), v_2(s), v_3(s), v_4(s))$$

be the corresponding trajectory relative to (8). *Then*

(i) *u is regular (i.e. $u \in L^\infty(B)$) if and only if*

$$\lim_{s\to-\infty} V(s) = P_1;$$

(ii) *u is weakly singular if and only if*

$$\lim_{s\to-\infty} V(s) = P_2.$$

Our following results concern the *existence* of weakly singular solutions and a lower bound $\lambda^*_{\min}$ on the value of λ^*. For all $n = 5, \dots, 16$ we prove the existence of λ_σ such that (P_{λ_σ}) admits a weakly singular solution; we provide a lower and upper bound on the value of λ_σ. For all $n = 5, \dots, 16$ let $\lambda_\sigma^{\min}$ and $\lambda_\sigma^{\max}$ be as given in Table 1, and for all $n = 5, \dots, 10$ let $\lambda^*_{\min}$ be as given in Table 1.

THEOREM 7. *For all $n = 5, \dots, 16$ there exists $\lambda_\sigma \in [\lambda_\sigma^{\min}, \lambda_\sigma^{\max}]$ such that (P_{λ_σ}) admits a weakly singular solution U_σ. In particular, $\lambda_\sigma > 8(n-2)(n-4)$.*

For all $n = 5, \dots, 10$ the value of λ^* is larger than $\lambda^*_{\min}$.

In section 6 we use Theorem 6 to show that Theorem 7 is equivalent to some intersection properties of the unstable manifolds of P_1 and P_2 with the hyperplane

TABLE 1

n	λ_σ	λ^*	$\lambda_\sigma^{\min}$	$\lambda_\sigma^{\max}$	$\lambda_{\min}^*$
5	113.19	236.49	113.11	113.26	235.89
6	260.82	362.10	260.72	260.86	361.34
7	449.55	524.70	449.45	449.60	523.16
8	679.45	728.36	679.04	679.55	724.50
9	950.28	976.66	949.58	950.49	969.81
10	1261.79	1272.09	1260.71	1262.23	1268.48
11	1613.78	1615.77	1610.89	1615.30	
12	2006.09	2006.11	1997.53	2010.41	
13	2438.60	2438.60	2403.42	2457.15	
14	2911.21	2911.21	2843.32	2947.17	
15	3423.83	3423.83	3260.54	3514.51	
16	3976.40	3976.40	3597.37	4211.88	

$v_1 = 4$. The remaining part of the proof of Theorem 7 is divided into two parts. First, in section 6 a rigorous bound on the location of the unstable manifold close to the stationary point is obtained by analytical methods. Then the intersection of the manifold with the hyperplane and its location are proved by a computer assisted algorithm; see section 9. The following definition explains exactly what we mean by a computer assisted proof.

DEFINITION 8. *A proof is called* computer assisted *if it consists in finitely many elementary operations, but their number is so large that, although each step may be written down explicitly, it is only practical to perform such operations with a computer.*

We believe that a weakly singular solution exists in any dimension $n \geq 5$, but since our type of proof requires a finite number of steps for each value of n, we cannot prove this conjecture. We performed the computer assisted proof for $n = 5, \ldots, 16$ because the "interesting" phenomena of (P_λ) arise in these dimensions.

We expect the "singular parameter" λ_σ and the singular solution to be unique. However, also for this statement, we do not yet have a proof. See Open Problem 3 in section 8 below.

Table 1 summarizes our results: λ^* and λ_σ are the best, purely numerical, estimates for the values, up to two decimal places, while the numbers $\lambda_\sigma^{\min}$, $\lambda_\sigma^{\max}$, and $\lambda_{\min}^*$ are rigorously computed values as stated in Theorem 7.

Remark 9. We point out that both the approximate numerical computation and the computation with rigorous estimate on the error for λ^* become very difficult as n increases. For this reason the best rigorous estimate we have on $\lambda_{\min}^*$ for $n \geq 11$ is nothing but for $\lambda_\sigma^{\min}$, while the best numerical estimate we have on λ^* for $n \geq 13$ is λ_σ. These values of n may be improved with a more accurate algorithm, but we do not feel that this would lead to a qualitative improvement of the result.

From Table 1 we immediately get the following.

COROLLARY 10. *For all $n = 5, \ldots, 10$ we have $\lambda_\sigma < \lambda^*$.*

Remark 11. We have numerical evidence that $\lambda_\sigma < \lambda^*$ for $n = 11, 12$ as well, but $\lambda^* - \lambda_\sigma$ is much smaller than the rigorous estimate of the numerical error, and thus we do not have a proof. For $n \geq 13$ the values of λ_σ and λ^* are closer than the numerical error; therefore we cannot even provide a conjecture supported by numerical evidence. If one could show uniqueness of the singular parameter λ_σ and that in fact $\lambda_\sigma < \lambda^*$ in dimensions $n \leq 12$, one could conclude that here the extremal solution U_* is either regular or "strongly singular" (i.e., $\lim_{r \to 0} ru'(r)$ does not exist). For $n \geq 13$ we expect the extremal solution U_* to be weakly singular. See Open Problems 3, 4, and 5 in section 8.

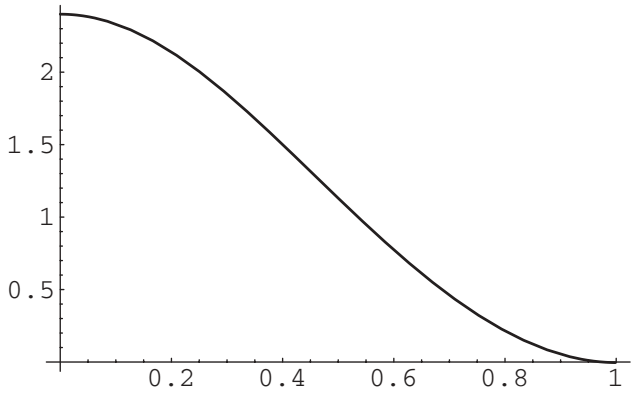

FIG. 1.

To complete the numerical inspection of the problem, we provide Figure 1, which shows the (regular) solution for $n = 5$ and $\lambda = \lambda^*$.

Theorem 6 states that any weakly singular solution $u = u(r)$ of (4) corresponds to a (weakly singular) solution $v = v(s)$ of (6) which satisfies $v(s) \approx -4s$ as $s \to -\infty$: this is because $\overline{v}(s) = -4s$ is precisely the stationary point P_2. Hence, as a further consequence of Theorem 6, we have that any weakly singular solution U_σ behaves asymptotically like $-4 \log r$ as $r \to 0$. Moreover, as may be checked by a simple calculation, the function $r \mapsto -4 \log r$ solves the equation and the first boundary condition in (P_λ) for $\lambda = 8(n-2)(n-4)$ but *not* the second boundary condition (recall also $\lambda_\sigma > 8(n-2)(n-4)$ by Theorem 7). Contrary to what happens for the second order equation (1), the explicit form of the radial weakly singular solution does not seem simple to be determined; see also Proposition 34 below. To this end, we characterize it further by means of the following theorem.

THEOREM 12. *Let U_σ be a weakly singular solution with $\lambda_\sigma > 8(n-2)(n-4)$ as it is obtained in Theorem 7 for $5 \le n \le 16$. Then*

$$U_\sigma(r) = -4 \log r + W(r),$$

where W is a bounded function satisfying

$$\lim_{r \to 0} W(r) = W_0 := \log \frac{8(n-2)(n-4)}{\lambda_\sigma} < 0$$

and (at least) one of the two following facts holds true:

(i) *$W(r) - W_0$ changes sign infinitely many times in any neighborhood of $r = 0$.*

(ii) *$W(r) \ge \max[W_0, 2r^2 - 2]$ for all $r \in (0, 1]$.*

If $n \ge 13$, case (ii) *necessarily occurs.*

Finally, the function $W(r)$ is not analytic, i.e., not a convergent power series in r^2 close to $r_0 = 0$.

Remark 13. It is quite surprising that the asymptotic behavior of weakly singular solutions of (P_λ) is the same as that of the quasi-linear equation $-\Delta_4 u = \lambda e^u$; see [GPP]. Here $-\Delta_p$ denotes the p-Laplace operator.

We conclude this section with a short remark concerning the behavior of (P_λ) in low dimensions.

Remark 14. In dimensions $1 \le n \le 4$ the problem is subcritical and has a different behavior. In particular, there are no singular solutions.

The minimal solution is constructed as in the present paper. There is a parameter $\lambda^* > 0$ such that for any $\lambda \in (0, \lambda^*)$ there is precisely one minimal stable positive solution. Taking this one as a "trivial" solution, with the help of variational techniques (which apply only in a subcritical setting) one finds a second positive "large" solution above the minimal solution and unstable. For $\lambda > \lambda^*$ there is no positive solution. Concerning the bifurcation diagram, one expects a smooth branch emanating from 0, extending until λ^*, where it bends back and approaches $\lambda = 0$, while the L^∞ norm of the solutions blows up. See also [We, Wi].

3. Characterization of regular and weakly singular radial solutions. In this section we perform a phase space analysis for the system (8), which corresponds to the radial version of (P_λ). This gives some insight into which behavior of regular and of weakly singular radial solutions may be expected in dependence on the space dimension. These results are essential for the proofs of Theorems 7 and 12. For the proofs of Theorems 3 and 4 one may skip directly to sections 4 and 5.

3.1. Analysis of the stationary points. It is easy to verify that system (8) has only two stationary points:

$$P_1 = (4, 0, 0, 0) \qquad \text{and} \qquad P_2 = (0, 4n - 8, 16 - 8n, -8(n-2)(n-4)).$$

In order to linearize (8) in a neighborhood of P_1, we must just replace $(8)_4$ with

$$v_4'(s) = 4v_4(s).$$

Then the linearized system has two distinct positive eigenvalues, $\mu_1 = 2$, $\mu_2 = 4$, and two distinct negative ones, $\mu_3 = 2 - n$, $\mu_4 = 4 - n$. We conclude that P_1 is a hyperbolic point independently of the dimension.

Eigenvectors corresponding to the positive eigenvalues μ_1, μ_2 in the neighborhood of P_1 have the form

$$\alpha_1(1, -n, 0, 0) \quad \text{and} \quad \alpha_2(-1, n + 2, 2n + 4, 2n^2 + 4n),$$

where $\alpha_1, \alpha_2 \in \mathbb{R} \setminus \{0\}$. Therefore, the tangent hyperplane to the unstable manifold of P_1 consists of those points in $\mathbb{R}^4$ whose coordinates can be represented as

$$(\alpha_1 - \alpha_2, -n\alpha_1 + (n+2)\alpha_2, (2n+4)\alpha_2, (2n^2+4n)\alpha_2) \tag{10}$$

with $\alpha_1, \alpha_2 \in \mathbb{R}$.

Similarly, the tangent hyperplane to the stable manifold of P_1 is spanned by eigenvectors corresponding to negative eigenvalues of the linearized system, that is,

$$\alpha_3(1, 0, 0, 0) \quad \text{and} \quad \alpha_4(1, -2, 2n - 4, 0),$$

where $\alpha_3, \alpha_4 \in \mathbb{R} \setminus \{0\}$.

Now consider the second critical point P_2 of (8). In its neighborhood the linear approximation of $(8)_4$ (the only nonlinear equation) takes the form

$$v_4'(s) = -8(n-2)(n-4)v_1(s).$$

Therefore, the eigenvalues of the linearized system in the neighborhood of P_2 are the solutions of the fourth order algebraic equation

$$\nu(\nu - 2)(\nu + n - 2)(\nu + n - 4) - 8(n-2)(n-4) = 0;$$

hence,

$$\nu_{1,2,3,4} = \frac{1}{2}\left(4 - n \pm \sqrt{M_1(n) \pm M_2(n)}\right),$$

where $M_1(n) = n^2 - 4n + 8 = (n-2)^2 + 4 > (n-2)^2$ and $M_2(n) = 4\sqrt{68 - 52n + 9n^2}$. Therefore,

$$\nu_1 = \frac{1}{2}\left(4 - n + \sqrt{M_1(n) + M_2(n)}\right) \qquad \text{and} \qquad \nu_2 = \frac{1}{2}\left(4 - n - \sqrt{M_1(n) + M_2(n)}\right)$$

are real numbers. It is easy to see that

$$\nu_2 < 0 < \nu_1 \qquad \text{for all } n \geq 4.$$

Moreover, for $5 \leq n \leq 12$, we have $M_1(n) - M_2(n) < 0$, while for $n \geq 13$ there holds $M_1(n) - M_2(n) > 0$. Therefore, for $5 \leq n \leq 12$ the eigenvalues

$$\nu_3 = \frac{1}{2}\left(4 - n + \sqrt{M_1(n) - M_2(n)}\right) \qquad \text{and} \qquad \nu_4 = \frac{1}{2}\left(4 - n - \sqrt{M_1(n) - M_2(n)}\right)$$

are complex conjugate with the real part

$$\text{Re } \nu_3 = \text{ Re } \nu_4 = \frac{1}{2}(4 - n) < 0,$$

while for $n \geq 13$ both ν_3 and ν_4 are real, $\nu_3 < 0$ and $\nu_4 < 0$.

This analysis implies that for all $n \geq 5$ the critical point P_2 of system (8) is also hyperbolic, but its stable manifold is three-dimensional and the unstable manifold is one-dimensional. Moreover, taking into account that for $5 \leq n \leq 12$ we have $\text{Im}\nu_3 = -\text{Im}\nu_4 \neq 0$, we deduce from the general theory of critical points (see, for example, [A2]) that for these values of n (and only for them) trajectories in the stable manifold of P_2 locally have the form of a spiral.

3.2. Proof of Theorem 6. We first consider regular solutions. It will prove to be useful to have the following meaning of $v_1, \ldots, v_4$ in terms of derivatives of u in mind:

$$\begin{cases} v_1(s) = e^s\, u'\,(e^s) + 4, \\ v_2(s) = -e^{2s} \cdot \Delta u\,(e^s)\,, \\ v_3(s) = -e^{3s}\,(\Delta u)'\,(e^s)\,, \\ v_4(s) = -\lambda e^{4s} e^{u(e^s)}. \end{cases} \tag{11}$$

If u is a regular solution of (P_λ), then $u, u', \Delta u$, and $(\Delta u)'$ stay bounded in particular for $r \searrow 0$, i.e., for $s \to -\infty$. So, we get immediately from (11) the first part of the statement.

To prove the converse, assume that

$$\lim_{s \to -\infty} (v_1(s), v_2(s), v_3(s), v_4(s)) = P_1$$

so that

$$\lim_{r \searrow 0} ru'(r) = \lim_{r \searrow 0} r^2 \Delta u(r) = \lim_{r \searrow 0} r^3 (\Delta u)'(r) = \lim_{r \searrow 0} r^4 e^{u(r)} = 0. \tag{12}$$

The first limit yields particularly, that for $r > 0$ small enough,

$$u(r) \le -\frac{1}{2}\log(r), \qquad e^{u(r)} \le r^{-1/2}.$$

Using the differential equation (P_λ) and the growth conditions (12) (observe $n > 4$), we obtain successively for r close to 0

$$(\Delta u)'(r) = O(r^{1/2}), \quad \Delta u(r) = O(1), \quad u'(r) = O(r), \quad u(r) = O(1).$$

That means that u is regular.

Next, we characterize weakly singular solutions. All the limits are intended as $s \to -\infty$; with c we denote generic constants.

Note first that if $\lim V(s) = P_2$, then the solution is weakly singular.

In order to prove the converse, we claim that

$$v'(s) \to -4. \tag{13}$$

To this end, we exclude all the other cases; recall that $\lim v'(s)$ exists by definition of weakly critical solutions.

(A) It cannot be that $\lim v'(s) = c \in (-\infty, -4)$.

For contradiction, if $\lim v'(s) = c < -4$, then by $(7)_1$ we infer

$$\lim v_1(s) = c + 4 < 0, \tag{14}$$

and by $(7)_4$ we get

$$v_4(s) \to -\infty. \tag{15}$$

Write $(8)_3$ as

$$\frac{d}{ds}[e^{(n-4)s}v_3(s)] = e^{(n-4)s}v_4(s)$$

so that by (15) we infer that the map $s \mapsto e^{(n-4)s}v_3(s)$ is decreasing in a neighborhood of $-\infty$, and therefore it admits a limit. If $e^{(n-4)s}v_3(s) \to c \ge 0$, then by $(8)_3$ and (15) we get $v_3'(s) \to -\infty$ and hence $v_3(s) \to +\infty$. If $e^{(n-4)s}v_3(s) \to c < 0$, then $v_3(s) \to -\infty$. In any case we obtain

$$|v_3(s)| \to +\infty. \tag{16}$$

A completely similar (but slightly more involved) argument shows that $(8)_2$ and (16) entail

$$|v_2(s)| \to +\infty. \tag{17}$$

Finally, $(8)_1$, (14), and (17) furnish $|v_1'(s)| \to +\infty$, which contradicts (14).

(B) It cannot be that $v'(s) \to -\infty$.

For contradiction, assume that $v'(s) \to -\infty$: then by $(7)_1$ we have

$$v_1(s) \to -\infty, \tag{18}$$

and by $(7)_4$ we get

$$v_4(s) \to -\infty; \tag{19}$$

moreover,

$$\frac{v(s)}{s} \to -\infty. \tag{20}$$

We may rewrite $(8)_3$ as

$$\frac{d}{ds}[e^{(n-4)s}v_3(s)] = e^{(n-4)s}v_4(s) = -\lambda e^{ns+v(s)} \to -\infty,$$

where the second equality is just $(7)_4$ and the infinite limit is a consequence of (20): the previous limit yields $e^{(n-4)s}v_3(s) \to +\infty$ and, in turn,

$$v_3(s) \to +\infty. \tag{21}$$

Similarly, we may rewrite $(8)_2$ as

$$\frac{d}{ds}[e^{-2s}v_2(s)] = e^{-2s}v_3(s) \to +\infty,$$

where the infinite limit is a consequence of (21): hence, we deduce that $e^{-2s}v_2(s) \to -\infty$, which, together with $(8)_2$ and (21), shows that $v_2(s) \to -\infty$. Inserting this into $(7)_2$ gives $v''(s)+(n-2)v'(s) \to +\infty$, and therefore $v'(s)+(n-2)v(s) \to -\infty$: hence,

$$\text{there exists } \sigma < 0 \quad \text{such that} \quad v'(s) + (n-2)v(s) < 0 \qquad \text{for all } s \le \sigma.$$

We rewrite this inequality as

$$\frac{d}{ds}[e^{(n-2)s}v(s)] < 0 \qquad \text{for all } s \le \sigma;$$

integrating it over $[s, \sigma]$ and taking into account that $v(\sigma) > 0$, we infer that

$$\text{there exists } K > 0 \quad \text{such that } v(s) \ge Ke^{(2-n)s} \qquad \text{for all } s \le \sigma.$$

Using (5) and returning to the function u (solution of (P_λ) and (4)), this shows that

$$\text{there exists } K > 0 \quad \text{such that } u(r) \ge \frac{K}{r^{n-2}} \qquad \text{for all } r \le e^\sigma;$$

this contradicts $e^u \in L^1(B)$.

(C) It cannot be that $\lim v'(s) = c \in (-4, 0]$.

For contradiction, if $\lim v'(s) = c \in (-4, 0]$, then by $(7)_1$ we infer

$$\lim v_1(s) = c + 4 > 0, \tag{22}$$

and by $(7)_4$ we get

$$v_4(s) \to 0. \tag{23}$$

Then from $(8)_3$ we deduce

$$v_3(s) \to 0, \tag{24}$$

because otherwise we would get a contradiction similar to that of case (A). Next, from $(8)_2$ and (24) we obtain

$$v_2(s) \to 0. \tag{25}$$

Since, by assumption, v_1 has a limit, we deduce that necessarily $v_1(s) \to 4$. This, together with (23), (24), and (25), contradicts part (i) proved above.

By (A), (B), (C), statement (13) is proved. This shows that $v_1(s) \to 0$: inserting this into $(8)_1$ gives $v_2(s) \to 4(n-2)$. Inserting the latter into $(8)_2$ yields $v_3(s) \to -8(n-2)$; finally, inserting this into $(8)_3$ gives $v_4(s) \to -8(n-2)(n-4)$. This completes the proof of (ii). □

Remark 15. If in case (ii) of Theorem 6 we do not assume that $\lim v'(s)$ exists, then we can merely show that $\liminf v'(s) \le -4 \le \limsup v'(s)$. Clearly, if one could prove that both inequalities are in fact equalities, then we would again have (13).

4. Proof of Theorem 3. We denote by $\mathcal{K}$ the cone of nonnegative L^2-functions in B,

$$\mathcal{K} = \{u \in L^2(B);\ u(x) \ge 0 \text{ for almost every } \ x \in B\},$$

and (for the sake of completeness) we prove the following weak formulation of Boggio's positivity preserving property [B], which we extensively use.

LEMMA 16. *Assume that $u \in L^2(B)$ satisfies*

$$\int_B u\Delta^2 v \ge 0 \qquad \textit{for all } v \in \mathcal{K} \cap H^4 \cap H_0^2(B);$$

then $u \in \mathcal{K}$. Moreover, one has either $u \equiv 0$ or $u > 0$ a.e. in B.

Proof. (i) Take any $\varphi \in \mathcal{K} \cap C_c^\infty(B)$ and let v_φ be the unique (classical) solution of

$$\begin{cases} \Delta^2 v_\varphi = \varphi & \text{in } B, \\ v_\varphi = \frac{\partial v_\varphi}{\partial \mathbf{n}} = 0 & \text{on } \partial B. \end{cases}$$

Then, by the classical Boggio principle [B], we infer that $v_\varphi \in \mathcal{K}$. Hence, v_φ is a possible test function for all φ so chosen, and therefore

$$\int_B u\varphi = \int_B u\Delta^2 v_\varphi \ge 0 \qquad \text{for all } \varphi \in \mathcal{K} \cap C_c^\infty(B).$$

This shows that $u \in \mathcal{K}$.

(ii) By (i) we know that $u \in \mathcal{K}$. So, assume that $u \not> 0$ a.e. in B and let ϕ denote the characteristic function of the set $\{x \in B;\ u(x) = 0\}$ so that $\phi \ge 0$, $\phi \not\equiv 0$. Let v_0 be the unique (a.e.) solution of the problem

$$\begin{cases} \Delta^2 v_0 = \phi & \text{in } B, \\ v_0 = \frac{\partial v_0}{\partial \mathbf{n}} = 0 & \text{on } \partial B. \end{cases}$$

Then

$$v_0 \in \left(\bigcap_{q \ge 1} W^{4,q}(B) \right) \subset C^3(\overline{B})$$

and by Boggio's principle [B] we have $v_0 > 0$ in B. By the biharmonic analogue of Hopf's lemma in balls (see [GS, Theorem 3.2], which also holds if $\Delta^2 v_0 \in L^p(B)$ for some $p > n/2$), we necessarily have $\Delta v_0 > 0$ on ∂B. This last inequality allows us to

state that for all $v \in C^4(\overline{B}) \cap H_0^2(B)$ there exists $t_1 \le 0 \le t_0$ such that $v + t_0 v_0 \ge 0$ and $v + t_1 v_0 \le 0$ in B. This, combined with the fact that

$$\int_B u\Delta^2 v_0 = \int_{\{u=0\}} u = 0,$$

enables us to show that both

$$0 \le \int_B u\Delta^2(v + t_0 v_0) = \int_B u\Delta^2 v \qquad \text{and} \qquad 0 \ge \int_B u\Delta^2(v + t_1 v_0) = \int_B u\Delta^2 v.$$

Hence, we have for all $v \in C^4(\overline{B}) \cap H_0^2(B)$

$$\int_B u\Delta^2 v = 0.$$

We need to show that $C^4(\overline{B}) \cap H_0^2(B)$ is dense in $H^4 \cap H_0^2(B)$. For this purpose, take any function $U \in H^4(B) \cap H_0^2(B)$ and put $f := \Delta^2 U$. We approximate f in $L^2(B)$ by $C^\infty(\overline{B})$-functions f_k and solve $\Delta^2 U_k = f_k$ in B under homogeneous Dirichlet boundary conditions. We then even have $U_k \in C^\infty(\overline{B})$, and by L^2-theory there holds $\|U_k - U\|_{H^4(B)} \to 0$ as $k \to \infty$.

By the previous statement we may now conclude that

$$\text{for all } v \in H^4 \cap H_0^2(B): \qquad \int_B u\Delta^2 v = 0.$$

Since $u \in L^2(\Omega)$, we may take as $v \in H^4 \cap H_0^2(B)$ the solution of $\Delta^2 v = u$ under homogeneous Dirichlet boundary conditions. This finally yields $u \equiv 0$. □

In particular, thanks to Lemma 16 we may establish a result in the spirit of [BCMR], as follows.

LEMMA 17. *For all $f \in L^1(B)$ such that $f \ge 0$ a.e. in B there exists a unique $u \in L^1(B)$ such that $u \ge 0$ a.e. in B and which satisfies*

$$\int_B u\Delta^2 v = \int_B fv \qquad \textit{for all } v \in C^4(\overline{B}) \cap H_0^2(B);$$

moreover, there exists $C > 0$ (independent of f) such that $\|u\|_1 \le C\|f\|_1$.

Proof. Uniqueness follows by means of the observation that L^∞-functions may be approximated by a pointwise convergent but uniformly bounded sequence of $C_c^\infty(B)$-functions. This is applied to truncations of u, and suitable test functions v are obtained from approximations of the truncations of u by solving the biharmonic Dirichlet problems.

Existence follows by truncating f and by arguing as in the proof of [BCMR, Lemma 1], the only difference being the positivity preserving property, which is standard for the Laplacian; in our case we invoke Lemma 16. □

Combining the method of proof of Lemmas 16 and 17, one also has the following.

LEMMA 18. *Assume that $u \in L^1(B)$ satisfies*

$$\int_B u\,\Delta^2 v\,dx \ge 0 \qquad \textit{for all } v \in \mathcal{K} \cap C^4(\overline{B}) \cap H_0^2(B);$$

then $u \ge 0$ a.e. in B.

As was pointed out to us by Anna Dall'Acqua (TU Delft), similar techniques and the application of Weierstraß's approximation theorem yield that also for the stronger conclusion of Lemma 16 it is enough to require $u \in L^1(B)$.

The previous lemmas enable us to make use of the super-subsolutions method, as follows.

LEMMA 19. *Let $\lambda > 0$ and assume that there exists $\bar{u} \in \mathcal{K}$ such that $e^{\bar{u}} \in L^1(B)$ and*

$$\int_B \bar{u}\Delta^2 v \geq \lambda \int_B e^{\bar{u}} v \qquad \text{for all } v \in \mathcal{K} \cap W^{4,p} \cap H_0^2(B).$$

Then there exists a solution u of (P_λ) such that $0 \leq u \leq \bar{u}$ a.e. in B.

Proof. Let $u_0 = \bar{u}$, and for all $m \in \mathbb{N}$, define inductively the function u_{m+1} as the unique solution of

$$\int_B u_{m+1}\Delta^2 v = \lambda \int_B e^{u_m} v \qquad \text{for all } v \in W^{4,p} \cap H_0^2(B). \tag{26}$$

Note that by Lemmas 16 to 18 the sequence $\{u_m\}$ is well-defined and

$$u_m \in \mathcal{K}, \quad e^{u_m} \in L^1(B), \quad 0 \leq u_{m+1}(x) \leq u_m(x) \text{ for almost every } x \in B \text{ for all } m \in \mathbb{N}.$$

Since this sequence is pointwise decreasing, there exists $u \in \mathcal{K}$ such that $e^u \in L^1(B)$ and which is the pointwise limit of $\{u_m\}$. Then, letting $m \to \infty$ in (26) and applying Lebesgue's theorem, we obtain the result. ∎

Define $\Lambda := \{\lambda \geq 0;\ (P_\lambda) \text{ admits a solution}\}$ and

$$\lambda^* := \sup \Lambda;$$

clearly $0 \in \Lambda$ and so $\Lambda \neq \emptyset$. Moreover, by the implicit function theorem we know that $\lambda^* > 0$. It follows directly from Lemma 19 that Λ is an interval.

Let $\lambda \in \Lambda$; then there exists u_λ satisfying (2). Taking into account that $e^s \geq es$ for all $s \geq 0$ with strict inequality whenever $s \neq 1$, and choosing $v = \phi_1$ (the normalized positive first eigenfunction of (3)) as a test function in (2), we get

$$\lambda_1 \int_B u_\lambda \phi_1 = \int_B u_\lambda \Delta^2 \phi_1 = \lambda \int_B e^{u_\lambda} \phi_1 > \lambda e \int_B u_\lambda \phi_1,$$

which proves that

$$\lambda < \frac{\lambda_1}{e} \qquad \text{for all } \lambda \in \Lambda. \tag{27}$$

We now prove the most delicate part of Theorem 3, namely, that for any $\lambda < \lambda^*$, there exists a *regular* solution.

LEMMA 20. *Assume that for some $\mu > 0$ there exists a (possibly singular) solution u_0 of (P_μ). Then for all $0 < \lambda < \mu$ there exists a regular solution of (P_λ).*

Proof. Let $0 < \lambda < \mu$ and consider the (unique) functions $u_1, u_2 \in L^1(B)$ satisfying, respectively,

$$\int_B u_1 \Delta^2 v = \lambda \int_B e^{u_0} v \qquad \text{for all } v \in W^{4,p} \cap H_0^2(B),$$

$$\int_B u_2 \Delta^2 v = \lambda \int_B e^{u_1} v \qquad \text{for all } v \in W^{4,p} \cap H_0^2(B). \tag{28}$$

Such functions exist by Lemma 17 and also belong to $L^2(B)$, since by Lemma 18 we have

$$u_0 > \frac{\lambda}{\mu} u_0 = u_1 \geq u_2 \qquad \text{almost everywhere in } B. \tag{29}$$

Let $\varphi(x) = (1 - |x|^2)^2$; it is readily verified that

$$\varphi \in H_0^2(B), \qquad \Delta^2 \varphi = 8n(n+2). \tag{30}$$

We also need the following elementary statement:

$$\begin{aligned} &\text{for all } \vartheta > 1 \text{ and } \delta > 0 \text{ there exists } \gamma > 0 \\ &\text{such that } e^{\vartheta s} + \gamma - (1+\delta)e^s \geq 0 \text{ for all } s \geq 0. \end{aligned} \tag{31}$$

Take $\vartheta = \mu/\lambda$, $\delta = n\lambda/4\mu$ and choose $k > 0$ in such a way that

$$e^{\frac{\mu}{\lambda} s} + \frac{8n(n+2)}{\lambda} k \geq (1+\delta)e^s \quad \text{for all } s \geq 0; \tag{32}$$

this choice is clearly allowed by (31). Thanks to (30) and (32) we find

$$\begin{aligned} \int_B (u_1 + k\varphi)\Delta^2 v = \int_B [\lambda e^{u_0} + 8n(n+2)k]v = \int_B [\lambda e^{\frac{\mu}{\lambda} u_1} + 8n(n+2)k]v \\ \geq \lambda(1+\delta) \int_B e^{u_1} v = (1+\delta) \int_B u_2 \Delta^2 v \\ \text{for all } v \in \mathcal{K} \cap W^{4,p} \cap H_0^2(B). \end{aligned}$$

Hence, by Lemma 16 we infer that $u_2 \leq \frac{u_1 + k\varphi}{1+\delta}$ in B; in particular, we get

$$e^{u_2} \leq e^{\frac{k}{1+\delta}\varphi}\, e^{\frac{\lambda}{\mu(1+\delta)} u_0},$$

from which we get at once that

$$e^{u_2} \in L^{\frac{n}{4} + \frac{\mu}{\lambda}}(B) \tag{33}$$

since $\varphi \in L^\infty(B)$ and $e^{u_0} \in L^1(B)$ (recall also our choice of δ). Finally, consider $u_3 \in L^2(B)$ such that

$$\int_B u_3 \Delta^2 v = \lambda \int_B e^{u_2} v \qquad \text{for all } v \in W^{4,p} \cap H_0^2(B).$$

By (33) and elliptic regularity [ADN], we deduce that

$$u_3 \in W^{4, \frac{n}{4} + \frac{\mu}{\lambda}}(B) \subset L^\infty(B).$$

Moreover, by (28), (29), and Lemma 16 we infer that $u_3 \leq u_2$ and hence

$$\int_B u_3 \Delta^2 v \geq \lambda \int_B e^{u_3} v \qquad \text{for all } v \in \mathcal{K} \cap W^{4,p} \cap H_0^2(B).$$

We have so found a weak bounded supersolution u_3 of (P_λ), and the statement follows from Lemma 19. □

With the help of Lemma 20 we can now show the following.

LEMMA 21. *For all $0 \le \lambda < \lambda^*$, the minimal solution U_λ exists, is regular, and is radially symmetric.*

Proof. By the preceding lemma we have the existence of a regular solution u_λ of (P_λ). This may serve as a (classical) supersolution of (P_λ), while $U_0 \equiv 0$ is a subsolution. Hence, the minimal solution U_λ of (P_λ) may be obtained as the increasing limit of the following sequence $\{U_m\}$:

$$\begin{cases} \Delta^2 U_{m+1} = \lambda e^{U_m} & \text{in } B, \\ U_{m+1} = \dfrac{\partial U_{m+1}}{\partial \mathbf{n}} = 0 & \text{on } \partial B \end{cases} \qquad (m \ge 0).$$

Since U_0 is radially symmetric, so is U_1; similarly, all the functions U_m are radially symmetric: therefore, their (pointwise) limit U_λ is also radially symmetric. □

The previous lemma allows us to show that the interval Λ is closed: we first remark that the map $\lambda \mapsto U_\lambda(x)$ is strictly increasing for all $x \in B$ (in view of Lemma 16). If $0 \le \lambda < \mu < \lambda^*$, the minimal solution U_μ of (P_μ) is a (strict) supersolution for (P_λ). Therefore

$$U_*(x) := \lim_{\lambda \to \lambda^*} U_\lambda(x) \in [0, \infty] \tag{34}$$

exists for all $x \in B$. In fact, more can be said about this limit, as follows.

LEMMA 22. *Let U_* be the function defined in* (34). *Then $U_*(x)$ is finite for almost every $x \in B$ and U_* solves (P_λ) for $\lambda = \lambda^*$. Moreover, $U_\lambda \to U_*$ in $H_0^2(B)$ as $\lambda \uparrow \lambda^*$. Finally, U_* is radially symmetric.*

Proof. By Lemma 21 we have $U_\lambda \in C^\infty(\overline{B})$, and therefore, by using the generalized Pohozaev identity [P] by Pucci and Serrin [PS] and by arguing as in the proof of [GMP, Théorème 2], we obtain that the set $\{U_\lambda;\ \lambda < \lambda^*\}$ is bounded in $H_0^2(B)$, and hence $U_\lambda \rightharpoonup U_*$ in $H_0^2(B)$, up to a subsequence (this follows by uniqueness of the pointwise limit). This shows that U_* is a.e. finite, that U_* solves (P_λ) for $\lambda = \lambda^*$, and also that $U_* e^{U_*} \in L^1(B)$. Finally, since $U_\lambda e^{U_\lambda} \le U_* e^{U_*}$, by Lebesgue's theorem we deduce that

$$\frac{1}{\lambda} \int_B |\Delta U_\lambda|^2 = \int_B U_\lambda e^{U_\lambda} \to \int_B U_* e^{U_*} = \frac{1}{\lambda^*} \int_B |\Delta U_*|^2 \qquad \text{as } \lambda \uparrow \lambda^*,$$

which, together with weak convergence, shows that $U_\lambda \to U_*$ in the norm topology of $H_0^2(B)$; since the above arguments may be repeated for any sequence in $\{U_\lambda;\ \lambda < \lambda^*\}$, the result follows without extracting subsequences.

Finally, by Lemma 21, all the minimal solutions U_λ (for $0 < \lambda < \lambda^*$) are radially symmetric. Then by (34) also U_* is radially symmetric. □

Remark 23. The proof of Lemma 22 may also be obtained by exploiting the stability of the minimal solution U_λ (see Proposition 37(i) below) and by arguing as in [BV, Remark 3.3].

Finally, we claim that

$$\lambda^* \ge 14.72(n-1)(n-3). \tag{35}$$

Indeed, this holds true by Lemma 19 since the function $\bar{u}(x) = 7.36(1-|x|)^2$ is a weak supersolution ($\bar{u} \in C^\infty(\bar{B} \setminus \{0\})$) of (P_λ) for all $\lambda \le 14.72(n-1)(n-3)$.

Proof of Theorem 3. The upper bound for λ^* follows from (27) and from Lemma 22, the latter saying that $\lambda^* \in \Lambda$. The lower bound for λ^* is proved in (35). Statement (i) follows from Lemmas 20 and 21. The map $\lambda \mapsto U_\lambda(x)$ is nondecreasing for all x by Lemma 19 and strictly increasing by Lemma 16; this proves the first part of statement (ii). The second parts of (ii) and (iii) follow from Lemma 22. Finally, the radial symmetry of U_* and of all the minimal solutions U_λ (for $\lambda < \lambda^*$) is obtained in Lemmas 22 and 21, respectively. The regular minimal solutions U_λ (for $\lambda < \lambda^*$) are strictly radially decreasing in view of [So]. Passing to the limit, we also get that U^* is radially decreasing. □

Remark 24. The above analysis does not allow us to establish whether the extremal solution U_* is regular, weakly singular, or singular. However, since it is radially symmetric, in the regular and weakly singular case, Theorem 6 describes the behavior of U_* when studied in the phase space $\mathbb{R}^4$. With our computer assisted proof, we may then find some space dimensions where the first case certainly occurs, provided that we can also show uniqueness of the weakly singular solution and the corresponding parameter λ_σ.

5. Proof of Theorem 4. We first show that

$$U_\lambda \to 0 \qquad \text{uniformly as } \lambda \to 0. \tag{36}$$

Since this is standard, we just briefly sketch its proof. By Theorem 3 we know that

$$0 < \lambda < \mu < \lambda^* \quad \Longrightarrow \quad U_\lambda(x) < U_\mu(x) \qquad \text{if } |x| < 1.$$

Then, by multiplying the equation in (P_λ) by U_λ and by integrating by parts, we obtain that $\|U_\lambda\|_{H_0^2(B)}$ remains bounded. Hence, up to a subsequence, $\{U_\lambda\}$ converges in the weak $H_0^2(B)$ topology to $U_0 \equiv 0$, which is the unique solution of (P_0). By convergence of the norms, we infer that the convergence is in the norm topology. Finally, by pointwise convergence and elliptic regularity, we infer (36).

Next, note that V_λ satisfies

$$\begin{cases} \Delta^2 V_\lambda = \lambda & \text{in } B, \\ V_\lambda = \dfrac{\partial V_\lambda}{\partial \mathbf{n}} = 0 & \text{on } \partial B. \end{cases} \tag{37}$$

Therefore, $\Delta^2 U_\lambda > \Delta^2 V_\lambda$, and the inequality $U_\lambda > V_\lambda$ follows by Lemma 16.

In order to prove the last statement of Theorem 4, note that from (36) we infer

$$\text{for all } \varepsilon > 0 \quad \text{there exists } \lambda_\varepsilon > 0 \quad \text{such that} \qquad \lambda < \lambda_\varepsilon \Longrightarrow \|U_\lambda\|_\infty < \varepsilon.$$

So, fix $\varepsilon > 0$ and let $\lambda < \lambda_\varepsilon$. Then (37) entails

$$\Delta^2 U_\lambda = \lambda e^{U_\lambda} < \lambda e^\varepsilon = e^\varepsilon \Delta^2 V_\lambda \qquad \text{in } B.$$

This shows that $U_\lambda(x) < e^\varepsilon V_\lambda(x)$ for all $x \in B$, and the result follows by arbitrariness of ε.

6. Proof of Theorem 7. The proof of Theorem 7 is obtained with computer assistance. We first describe the numerical procedure used to obtain the approximate values for λ_σ and λ^*; then we show how the algorithm can be made rigorous. We maintain here the same notation as in section 3. The computation of λ_σ is somehow simpler than the computation of λ^*, since the unstable manifold of P_2 is one-dimensional. We

choose a point $\bar{v} = P_2 + re_1$ where e_1 is an eigenvector corresponding to the unstable manifold and r is some small value. We solve system (8) with $\bar{v}$ as the initial condition and look for the intersection of the solution with the hyperplane $v_1 = 4$. The choice of a positive or negative r leads to a different result, since the manifold is made of two branches: it turns out that one branch never appears to intersect the hyperplane, while the other branch always does. If the solution intersects the hyperplane $v_1 = 4$ at some point $\hat{v} = (\hat{v}_1, \hat{v}_2, \hat{v}_3, \hat{v}_4)$ such that $\hat{v}_4 < 0$, by Theorem 6 and by (7) we have numerical evidence of a singular solution at $\lambda = -\hat{v}_4$.

In order to compute the value of λ^*, we have to study the two-dimensional unstable manifold of P_1. The direction on the tangent hyperplane can be parametrized by an angle ϑ. In order to find the largest value for λ, we use a *directional shooting* method; i.e., we choose some value ϑ (the shooting direction) and solve the equation with starting point $\bar{v} = P_1 + r(e_1 \sin\vartheta + e_2 \cos\vartheta)$, where e_1 and e_2 are the orthonormalized eigenvectors corresponding to the (tangent) unstable manifold and $r > 0$ is some small arbitrarily chosen value. If the solution intersects the hyperplane $v_1 = 4$ at some point $\hat{v} = (\hat{v}_1, \hat{v}_2, \hat{v}_3, \hat{v}_4)$ such that $\hat{v}_4 < 0$, then by Theorem 6 and by (7) and (8) we have numerical evidence of a regular solution for $\lambda = -\hat{v}_4$. By varying ϑ we can look for the maximal value of λ.

Of course these procedures do not lead to an exact value for two reasons. First, we can choose only $\bar{v}$ on the unstable manifold of the linearized equation, and although we know that we are close to the manifold of the full equation, we are not exactly on it. Second, the algorithm used to solve the differential equation provides an accurate, but not rigorous, solution. We address the problem of proving that a branch of the unstable manifold of P_2 does intersect the hyperplane $v_1 = 4$ and of computing a rigorous estimate for the values λ_σ and λ^* in the following sections.

6.1. Rigorous bounds for the manifolds. We first address the general problem of computing rigorous bounds for the location of the unstable manifold in the neighborhood of a stationary hyperbolic point of an ODE. The same technique could be applied to the stable manifold as well, but in this paper we are not interested in it.

Let $f \in C^2(\mathbb{R}^d, \mathbb{R}^d)$, $d \geq 2$. We consider the equation $\dot{x} = f(x)$ and assume that 0 is a hyperbolic stationary point. Then

$$\dot{x} = Ax + N(x), \tag{38}$$

where

$$A = \nabla f(0), \quad N(x) = O(|x|^2) \text{ as } x \to 0 \tag{39}$$

and all eigenvalues of A have nonzero real part. Let $\varphi(x,t)$ be the flow induced by (38) and let $\varphi_A(x,t)$ be the flow induced by the linear equation $\dot{x} = Ax$. Let S_0 (resp., U_0) be the span of all eigenvectors corresponding to the eigenvalues with negative (resp., positive) real part. S_0 (resp., U_0) is called the stable (resp., unstable) subspace, and it is characterized as follows: S_0 (resp., U_0) is the set of points $x \in \mathbb{R}^d$ such that $\varphi_A(x,t) \to 0$ as $t \to +\infty$ (resp., $t \to -\infty$). It is well known that the full equation also admits a stable manifold S (resp., an unstable manifold U) still defined as the set of points $x \in \mathbb{R}^d$ such that $\varphi(x,t) \to 0$ as $t \to +\infty$ (resp., $t \to -\infty$). Such a manifold is tangent at the origin to S_0 (resp., U_0). If S_0 (resp., U_0) is empty, then there exists a neighborhood of the origin which is a subset of U (resp., S). We are interested in the case when both manifolds are nontrivial, and we wish to study the intersection of the unstable manifold with some other manifold P. In order to achieve this goal,

we consider a point $\bar{x} \in U \setminus \{0\}$ and study $\varphi(\bar{x}, t)$. If we can prove that $\varphi(\bar{x}, t_0) \in P$ for some positive t_0, then we infer that $U \cap P \neq \emptyset$, and we also know the intersection point. The main problem to address is that the only point of the manifold we know precisely is the origin: the other points lie very close to U_0, at least in a neighborhood of 0, but we do not know their explicit position. We proceed as follows.

There exists an invertible matrix M such that $B := M^{-1}AM$ is block diagonal, i.e., the canonical basis $\{e_i\}$ of $\mathbb{R}^d$ is split in $S_0' \cup U_0'$, where $S_0' = \operatorname{span}\{e_1, \dots, e_m\}$ is the stable eigenspace and $U_0' = \operatorname{span}\{e_{m+1}, \dots, e_d\}$ is the unstable eigenspace. If we let $y = M^{-1}x$, the (38) can be written as

$$\dot{y} = By + M^{-1}N(My) =: g(y). \tag{40}$$

By (39), for all $\varepsilon > 0$ there exists $\beta > 0$ such that $|N(x)| \le \beta |x|^2$ for all $|x| \le \varepsilon$. Let $\alpha < 0$ be the maximum of the real parts of the eigenvalues with negative real parts, $\gamma = -\frac{\alpha}{\beta m_1^2 m_2}$, $m_1 = \|M\|$, and $m_2 = \left\|M^{-1}\right\|$. Choose $\varepsilon > 0$; let $\beta > 0$ as above; choose a vector $\hat{y} \in U_0' \setminus \{0\}$ of norm $r \le \varepsilon$ and $k > 1$. Let P_s be the orthogonal projection onto S_0'; let P_u be the orthogonal projection onto the linear space spanned by $\hat{y}$; and let

$$\Xi = \left\{ y \in \mathbb{R}^d : \frac{\gamma}{k} |P_s y| \le |P_u y|^2 \le r^2 \right\}. \tag{41}$$

We show that, under a suitable choice of $k > 1$ and $0 < r \le \varepsilon$, for all $y \in \partial\Xi$ such that $|P_u y| < r$ the flow is inward; i.e., given $\bar{y} \in \Xi$ we want the solution of the Cauchy problem $\dot{y}(t) = g(y(t))$, $y(0) = \bar{y}$ to leave Ξ only through the set $\{y \in \partial\Xi : |P_u y| = r\}$. If this happens, then for all $\hat{y} \in U_0'$ satisfying $|\hat{y}| = r$ either the unstable manifold intersects the set

$$\kappa := \kappa_{\hat{y}} := \hat{y} + \left\{ \tilde{y} \in S_0' : |\tilde{y}| \le \frac{k}{\gamma} r^2 \right\} \tag{42}$$

or it is entirely contained in Ξ. As a result, to study a branch of the unstable manifold it is sufficient to exclude the second case and consider the initial value problem for all $\bar{y} \in \kappa$.

LEMMA 25. *Choose $\varepsilon > 0$ and $k > 1$. Let α, β, γ, m_1, m_2, and Ξ be as above and let*

$$r = \min\left\{ \frac{\varepsilon\gamma}{m_1\sqrt{\gamma^2 + k^2}}, 1, \frac{\sqrt{k-1}}{k}\gamma, \frac{\gamma}{2} \right\}. \tag{43}$$

For all $\bar{y} \in \partial\Xi$ such that $0 < |P_u \bar{y}| < r$ we have

$$(g(\bar{y}), P_s \bar{y}) < 0. \tag{44}$$

Proof. Let $\hat{y} = P_s\bar{y}$, $\tilde{y} = P_u\bar{y}$, $\hat{r} = |\hat{y}|$, and $\tilde{r} = |\tilde{y}|$. Since $\tilde{r} < r \le \frac{\varepsilon\gamma}{m_1\sqrt{\gamma^2+k^2}}$, then $|M\bar{y}| \le \varepsilon$, and therefore $|N(\bar{y})| \le \beta |\bar{y}|^2$. We have

$$\begin{aligned}(B\bar{y}, \hat{y}) + (M^{-1}N(M\bar{y}), \hat{y}) &= (B\hat{y}, \hat{y}) + (N(M\bar{y}), (M^{-1})^t\hat{y}) \\ &\le \alpha\hat{r}^2 + \beta m_1^2 m_2 (\hat{r}^2 + \tilde{r}^2)\hat{r}.\end{aligned}$$

Then a simple computation shows that (44) is implied by

$$\frac{1}{2}\left(\gamma + \sqrt{\gamma^2 - 4\tilde{r}^2}\right) > \hat{r} > \frac{1}{2}\left(\gamma - \sqrt{\gamma^2 - 4\tilde{r}^2}\right). \tag{45}$$

The first inequality is satisfied because $\tilde{r} < r \le \frac{\gamma}{2}$. For all $\bar{y} \in \partial\Xi$, $0 < |P_u\bar{y}| < r$, by (41), we have $\hat{r} = \frac{k}{\gamma}\tilde{r}^2$ and

$$\frac{1}{2}\left(\gamma - \sqrt{\gamma^2 - 4\tilde{r}^2}\right) < k\frac{\tilde{r}^2}{\gamma} \qquad \text{if} \quad \tilde{r} < \frac{\sqrt{k-1}}{k}\gamma\ .$$

Since $\tilde{r} < r$, then (45) and therefore (44) hold. □

We need a condition which ensures that the invariant manifold is not entirely contained in Ξ, but it intersects κ at some point. Let α' be the minimum of the real parts of the eigenvalues of B with positive real parts.

LEMMA 26. *If r, α, α', k, and Ξ are as above and*

$$\alpha' + \frac{\alpha}{\gamma}\left(r + \frac{k^2}{\gamma^2}r^3\right) > 0, \tag{46}$$

then there exists $\delta > 0$ such that $(g(\bar{y}), P_u\bar{y}) \ge \delta|P_u\bar{y}|^2$ for all $\bar{y} \in \Xi$, and therefore the component of the flow in the direction of the unstable manifold is always increasing in Ξ, together with its first derivative.

Proof. Choose $\delta > 0$ satisfying

$$\beta m_1^2 m_2\left(r + \frac{k^2}{\gamma^2}r^3\right) \le \alpha' - \delta.$$

Fix $\bar{y} \in \Xi$ and let $\hat{y} = P_s\bar{y}$, $\tilde{y} = P_u\bar{y}$, $\hat{r} = |\hat{y}|$, and $\tilde{r} = |\tilde{y}|$. By (46) and the definition of γ, such a δ exists. We have

$$\begin{aligned}(B\bar{y}, \tilde{y}) + (M^{-1}N(M\bar{y}), \tilde{y}) &= (B\tilde{y}, \tilde{y}) + (N(M\bar{y}), (M^{-1})^t\tilde{y}) \\ &\ge \alpha'\tilde{r}^2 - \beta m_1^2 m_2\left(\hat{r}^2 + \tilde{r}^2\right)\tilde{r} \ge \delta\tilde{r}^2,\end{aligned}$$

because $\hat{r} \le \frac{k}{\gamma}\tilde{r}^2$ by the definition of Ξ. □

LEMMA 27. *Let r, α, α', k, $\hat{y}$, and κ be as above. The unstable manifold tangent to $\hat{y}$ intersects κ.*

Proof. By Lemma 26 the unstable manifold cannot be entirely contained in Ξ. By Lemmas 25 and 26 it can only exit through κ. □

In the next subsection we apply these ideas in order to prove Theorem 7.

6.2. The computer assisted proofs. We apply the general result stated in the previous subsection to system (8).

We first consider the point $P_1 = (4, 0, 0, 0)$. Let $x = v - P_1$. System (8) takes the form (38) with

$$A_1 = \begin{bmatrix} 2-n & -1 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 4-n & 1 \\ 0 & 0 & 0 & 4 \end{bmatrix},$$

$N(x) = (0, 0, 0, x_1x_4)$, and $\alpha = 4 - n$.

If we consider the linearization at $P_2 = (0, 4n-8, 16-8n, -8(n-2)(n-4))$ and set $x = v - P_2$, then system (8) can be written as (38) with

$$A_2 = \begin{bmatrix} 2-n & -1 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 4-n & 1 \\ -8(-4+n)(-2+n) & 0 & 0 & 0 \end{bmatrix},$$

and again $N(x) = (0, 0, 0, x_1 x_4)$. From section 3.1 we know that if $n = 5, \dots, 12$, the eigenvalues are $((4-n)/2 + i\sigma, (4-n)/2 - i\sigma, \lambda_1, \lambda_2)$, where $\lambda_1 < (4-n)/2 < 0$ and $\lambda_2 > 0$. It turns out that $\alpha = (4-n)/2$. If $n \geq 13$, all eigenvalues are real and

$$\alpha = 2 - \frac{1}{2}n + \frac{1}{2}\sqrt{8 - 4n + n^2 - 4\sqrt{68 - 52n + 9n^2}}.$$

We remark that since the nonlinear part is very simple, it is possible to obtain a better estimate for the coefficients β, m_1, and m_2 than the one we had in section 6.1.

In the following, let M be the matrix that diagonalizes either A_1 or A_2 and let $|M_i|$ be the (Euclidean) norm of the ith row of M.

LEMMA 28. *For all $y_1, y_2 \in \mathbb{R}^4$ the following inequality holds:*

$$(N(My_1), (M^{-1})^t y_2) \leq |M_1|\,|M_4|\left|(M^{-1})_4^t\right| |y_1|^2 |y_2|.$$

Proof. We have

$$(N(My_1), (M^{-1})^t y_2) = (My_1)_1 (My_1)_4 ((M^{-1})^t y_2)_4,$$

where we denoted by $(Av)_i$ the ith component of the vector (Av), i.e., the scalar product of the ith row of A with the vector v. The conclusion follows by the definition of $|M_i|$. ∎

By the above lemma we infer that γ may be obtained as

$$\gamma = -\frac{\alpha}{|M_1|\,|M_4|\,|(M^{-1})_4^t|} \tag{47}$$

and ε may be chosen arbitrarily.

To compute a rigorous enclosure $[\lambda_\sigma^{\min}, \lambda_\sigma^{\max}]$ for the value of λ_σ, we fix n and compute the value γ in (47). We can choose $k > 1$ and $r > 0$ satisfying (43) and (46). We have some degree of arbitrariness: we prefer a small r in order to have a small set κ, but we also like a large r in order to reach the hyperplane in fewer time steps. It is also convenient to have the smallest possible k, since it also implies a smaller set κ. We have to make an empirical choice by trying different values and selecting the best trade-off. It turns out that it is convenient to choose r first, set

$$k = \frac{\gamma^2 - \sqrt{\gamma^4 - 4r^2\gamma^2}}{2r^2}, \tag{48}$$

and check whether (46) holds. Since the unstable manifold in P_2 is one-dimensional, we have to choose between two possible directions. The numerical experiment gave us the correct direction. Once we choose r and compute k, we have the set κ as given in (42). We should compute the evolution of all points in κ and its intersection with the hyperplane $v_1 = 4$. This would require a very long computer time, but since two solutions of (38) cannot intersect, then it is enough to compute the evolution of the points in the boundary of κ, provided we can prove that the trajectories of all points in the interior of κ also reach the hyperplane $v_1 = 4$. This can be checked by the following lemma.

LEMMA 29. *Set*

$$\kappa' := \kappa'_{\hat{y}} := \hat{y} + \left\{ \tilde{y} \in S_0 : |\tilde{y}| = \frac{k}{\gamma} r^2 \right\}. \tag{49}$$

Assume that the trajectories of all points in κ' intersect the hyperplane $v_1 = 4$ and do not intersect the hyperplane $(2-n)v_1 - v_2 + 4(n-2) = 0$. Let $\hat{\kappa}$ be the intersection of all such trajectories with $v_1 = 4$.

Then the trajectories of all points in κ also intersect the hyperplane $v_1 = 4$, and the intersection takes place in the region bounded by $\hat{\kappa}$.

Proof. Since $v_1' = (2-n)v_1 - v_2 + 4(n-2)$, then v_1' is positive and bounded away from zero for all points of the trajectories starting from κ'. Then, by the uniqueness and continuous dependence on the initial condition of the Cauchy problem, it follows that the union $\tau(\kappa')$ of such trajectories is a "tube" in $\mathbb{R}^4$ and the trajectories of all points in $\kappa \setminus \kappa'$ cannot exit $\tau(\kappa')$. Then v_1' is also positive and bounded away from zero for all points starting in κ, and the trajectory of every point in κ reaches $v_1 = 4$ in a finite time. □

Our strategy is as follows: We compute the intersection of the flow starting from all points in κ' with the hyperplane $v_1 = 4$. If all the trajectories intersect the hyperplane, we have a proof that the singular solution exists; furthermore the envelope in the v_4-direction of all intersections yields the desired λ-interval. Note that the set κ' is the image of S^2 through an invertible affine map, and therefore we need an efficient discretization of a sphere.

LEMMA 30. *For all $n = 5, \dots, 16$, let $r = .001$, let k be as in (48), and, let $\kappa' = \kappa'_{\hat{y}}$ as in (49). For a suitable choice of the direction $\hat{y}$ in the one-dimensional unstable manifold U_0, the following conclusions hold:*

1. *The flow starting in κ' intersects the hyperplane $v_1 = 4$.*
2. *The absolute value of the first coordinate of the intersection point is in the interval set $[\lambda_\sigma^{\min}, \lambda_\sigma^{\max}]$ defined in Table 1.*
3. *The flow starting in κ' and ending on the hyperplane $v_1 = 4$ does not intersect the hyperplane $(2-n)v_1 - v_2 + 4(n-2) = 0$.*

The proof is by computer assistance, as described in section 9.

In order to compute a rigorous lower bound for λ^*, we consider the trajectories of points in the unstable manifold of P_1 and compute the intersection with the hyperplane $v_1 = 4$. Since the manifold is two-dimensional, we have to decide the direction to follow: we use the numerical results presented above to compute the direction that gives the highest possible value for λ. We define κ as above, and we wish to prove that all trajectories starting from κ intersect the hyperplane $v_1 = 4$. We also need to estimate the location of such intersections. It would save some computer time to restrict the computation to the boundary of κ as in the proof of Lemma 30, but we cannot proceed as in Lemma 29 because P_1 lies on the hyperplane $v_1 = 4$ and therefore v_1 cannot be monotone. Furthermore, since the unstable manifold has now dimension 2, we do not have the topological argument (the tube) used before. On the other hand, in this case we only have to consider a region which is the affine image of a disk; therefore it is feasible to compute the trajectory for all point in the disk.

LEMMA 31. *For all $n = 5, \dots, 10$, let $r = .001$ if $n \le 9$ and $r = .0001$ if $n = 10$; let $\hat{y} = P_1 + r(e_1 \sin\vartheta_n + e_2 \cos\vartheta_n)$, where e_1 and e_2 are the eigenvectors of A_1 with unit norm and positive first component corresponding, respectively, to the eigenvalues 2 and 4 and $\vartheta_5 = 6.2829856$, $\vartheta_6 = 6.28298854$, $\vartheta_7 = 6.2829901$, $\vartheta_8 = 6.2829918$, $\vartheta_9 = 6.2829914$, $\vartheta_{10} = 6.28316589$; let k be as in (48) and let κ be as in (42).*

1. *The flow starting at all points of κ intersects the hyperplane $v_1 = 4$.*
2. *The absolute value of the first coordinate of the intersection point is larger than the $\lambda^*_{\min}$ displayed in Table 1.*

We point out that this statement shows only that there exists a regular solution

for some value of λ obtained as the intersection of a one-dimensional submanifold of the unstable manifold with the hyperplane $v_1 = 4$. Since we cannot exclude that there exists a solution for a larger value of λ, we only have a lower bound for λ^*.

The proof of Theorem 7 follows by Lemmas 27–31.

7. Proof of Theorem 12. In this section we use both the PDE notation Δ^2 and the ODE notation with primes denoting differentiation (with respect to r or s, depending on the context).

We assume that U_σ is any radial weakly singular solution of (P_{λ_σ}) with

$$\lambda_\sigma > 8(n-2)(n-4). \tag{50}$$

In particular, we deal with those solutions obtained in Theorem 7; see also Table 1. Then, by Theorem 6(ii), we know that

$$U_\sigma(r) = -4\log r + o(|\log r|) \qquad \text{as } r \to 0.$$

Therefore, we define the function

$$W(r) := U_\sigma(r) + 4\log r$$

and study its behavior. After some calculations, we find that it weakly solves the equation

$$\begin{cases} \Delta^2 W = \dfrac{1}{|x|^4}[\lambda_\sigma e^W - 8(n-2)(n-4)] & \text{in } B, \\ W = 0 & \text{on } \partial B, \\ \dfrac{\partial W}{\partial \mathbf{n}} = 4 & \text{on } \partial B. \end{cases} \tag{51}$$

The proof of Theorem 12 follows from the next two lemmas and Proposition 34 at the end of this section.

LEMMA 32. *Assume* (50) *and assume that* $W \in C^4(0,1]$ *weakly solves* (51) *(*$W = W(r)$*); then*

$$\lim_{r\to 0} W(r) = \log \frac{8(n-2)(n-4)}{\lambda_\sigma} = W_0 < 0. \tag{52}$$

Moreover, at least one of the two following facts holds true:

(i) *The function* $W(r) - W_0$ *changes sign infinitely many times in any neighborhood of* $r = 0$.

(ii) $W(r) \geq \max[W_0, 2r^2 - 2]$ *for all* $r \in (0,1]$.

Proof. The negativity of W_0 follows from (50), while (52) is a consequence of Theorem 6.

Assume that case (i) in the statement does not occur; we first claim that

$$W(r) \geq W_0 \qquad \text{for all } r \in (0,1]. \tag{53}$$

For contradiction, assume that (53) does not hold; then there exists $\bar R \in (0,1)$ such that $W(\bar R) < W_0$ and two cases may occur, as follows.

First case. There exists $R \in (0,1)$ such that $W'(R) = 0$ and $W_0 \leq W(r) < W(R)$ for all $r \in (0,R)$. In this case, let $H(r) = W(r) - W(R)$ so that $H(r) < 0$ for all $r \in (0,R)$; on the other hand, H weakly solves the problem

$$\begin{cases} \Delta^2 H = \Delta^2 W \geq 0 & \text{in } B_R, \\ H = \dfrac{\partial H}{\partial \mathbf{n}} = 0 & \text{on } \partial B_R, \end{cases}$$

so that by Lemma 16, one gets $H(r) \geq 0$ for all $r \in (0, R)$, a contradiction.

Second case. There exists $R \in (0,1)$ such that $W'(R) = 0$ and $W_0 \geq W(r) > W(R)$ for all $r \in (0, R)$. In this case, $H(r) = W(r) - W(R)$ satisfies both $H(r) > 0$ for all $r \in (0, R)$ and

$$\begin{cases} \Delta^2 H = \Delta^2 W \leq 0 & \text{in } B_R, \\ H = \dfrac{\partial H}{\partial \mathbf{n}} = 0 & \text{on } \partial B_R, \end{cases}$$

giving again a contradiction.

We have so proved (53): hence, if we define the function $\phi(r) = W(r) + 2 - 2r^2$, we infer that $\phi = \phi(|x|)$ weakly satisfies

$$\begin{cases} \Delta^2 \phi = \Delta^2 W \geq 0 & \text{in } B, \\ \phi = \dfrac{\partial \phi}{\partial \mathbf{n}} = 0 & \text{on } \partial B; \end{cases}$$

this yields $\phi(r) \geq 0$, namely, $W(r) \geq 2r^2 - 2$ for all $r \in (0, 1]$.

We have so proved that if (i) does not occur, then (ii) holds true, that is, the statement follows. □

In high dimensions the previous alternative breaks down, and we can describe the behavior of weakly singular solutions.

LEMMA 33. *If $n \geq 13$, then case* (i) *of Lemma* 32 *cannot occur.*

Proof. Let $W = W(r)$, let W_0 be as in Lemma 32, and consider the function

$$Z(s) = W(e^s) - W_0, \qquad s \in (-\infty, 0).$$

Then, since W satisfies (51), we deduce that

$$L_4 Z + p(s) Z = 0, \qquad s \in (-\infty, 0), \tag{54}$$

where $L_4 Z = Z'''' + 2(n-4)Z''' + (n^2 - 10n + 20)Z'' - 2(n-2)(n-4)Z'$ and

$$p(s) = -8(n-2)(n-4)\frac{e^{Z(s)} - 1}{Z(s)}.$$

Note that $p(s)$ is well-defined for all $s < 0$ and that, by (52), $p(s) \to -8(n-2)(n-4)$ as $s \to -\infty$. In particular, for all $\varepsilon > 0$ there exists $s_\varepsilon < 0$ such that

$$p(s) \geq -[8(n-2)(n-4) + \varepsilon] \quad \text{for all } s \leq s_\varepsilon. \tag{55}$$

Since $n \geq 13$, for sufficiently small ε, the linear equation

$$L_4 Z - [8(n-2)(n-4) + \varepsilon] Z = 0 \tag{56}$$

admits four linearly independent solutions of "exponential type," namely, $Z_i(s) = e^{\nu_i s}$ for some $\nu_i \in \mathbb{R}$ $(i = 1, \ldots, 4)$; see also the discussion in section 3.1. Hence, (56) is nonoscillatory in $(-\infty, 0)$ according to the definition in [E]. Therefore, by (55) and [E, Corollary 1], also (54) is nonoscillatory in $(-\infty, 0)$ and the statement follows. □

Let us conclude this section with the observation that an explicit form of the weakly singular solution U_σ seems not so easy to be obtained.

PROPOSITION 34. *Assume that the function W is a solution of* (51) *as considered in Lemma* 32. *Then the function $W = W(r)$ is not analytic in r close to* 0.

Proof. For contradiction, let $a_k = W^{(2k)}(0)/(2k)!$ and assume that

$$W(r) = \sum_{k=0}^{\infty} a_k r^{2k}$$

is a convergent power series for r close to 0. Since W is regular, the right-hand side of the equation in (51) is bounded as $r \to 0$, and we necessarily have

$$a_0 = \log \frac{8(n-2)(n-4)}{\lambda_\sigma}\,, \qquad a_1 = \frac{W''(0)}{2} = 0. \tag{57}$$

Then

$$W^{(k)}(r) = \frac{W''''(0)}{(4-k)!}\, r^{4-k} + O\left(r^{5-k}\right) \quad \text{as } r \to 0 \qquad (k = 1, 2, 3),$$

and hence

$$\frac{n(n+2)}{3} W''''(0) = \Delta^2 W|_{r=0} = \frac{\lambda_\sigma e^{a_0} W''''(0)}{24} = \frac{(n-2)(n-4)}{3} W''''(0),$$

where we have used (51) and (57). This shows that $W''''(0) = 0$ and $a_2 = 0$.

We now proceed by induction. Assume that for some $k \geq 2$ we have shown that $a_1 = \cdots = a_k = 0$; we claim that $a_{k+1} = 0$. Once we show this, we achieve a contradiction and the statement follows. Note that $\lambda_\sigma e^W - 8(n-2)(n-4) = 8(n-2)(n-4)[e^{W-a_0} - 1]$ and, by induction assumption,

$$\frac{1}{r^4}\left(e^{W-a_0} - 1\right) = a_{k+1} r^{2k-2} + O\left(r^{2k}\right).$$

Therefore, from (51) we get

$$\left(\frac{d}{dr}\right)^{2k-2} \Delta^2 W\Big|_{r=0} = 8(2k-2)!(n-2)(n-4)a_{k+1}. \tag{58}$$

On the other hand, recalling the radial form of Δ^2 (see the left-hand side of (4)) and taking into account that (as $r \to 0$)

$$W'(r) \sim \frac{W^{(2k+2)}(0)}{(2k+1)!}\, r^{2k+1}\,, \qquad W''(r) \sim \frac{W^{(2k+2)}(0)}{(2k)!}\, r^{2k},$$
$$W'''(r) \sim \frac{W^{(2k+2)}(0)}{(2k-1)!}\, r^{2k-1}\,, \qquad W''''(r) \sim \frac{W^{2k+2}(0)}{(2k-2)!}\, r^{2k-2},$$

we also deduce that

$$\left(\frac{d}{dr}\right)^{2k-2} \Delta^2 W|_{r=0} = 2k(2k+2)(n+2k)(n+2k-2) \cdot (2k-2)!\, a_{k+1}.$$

Combining this with (58), we get

$$a_{k+1}\left\{2k(2k+2)(n+2k)(n+2k-2) - 8(n-2)(n-4)\right\} = 0.$$

Since the term in brackets is strictly positive, this yields $a_{k+1} = 0$. ∎

8. Further results and open problems. First, we discuss the stability of the linearizations around regular solutions of problem (P_λ). For this purpose we observe that the minimal solution depends continuously on λ.

PROPOSITION 35. *As before, let U_λ denote the minimal solution of (P_λ). Then $[0,\lambda^*) \ni \lambda \mapsto U_\lambda \in C^{4,\alpha}(\overline{B})$ is continuous from the left. Moreover, if $\lambda_0 \in [0,\lambda^*)$ is such that the first eigenvalue of the linearization $L_{U_{\lambda_0}} := \Delta^2 - \lambda_0 \exp(U_{\lambda_0})$ is strictly positive, then $\lambda \mapsto U_\lambda$ is also continuous in $\lambda = \lambda_0$.*

Proof. Let $\lambda_k \nearrow \lambda_0$. Since $U_{\lambda_k} \le U_{\lambda_0}$ and since the $(U_{\lambda_k})_k$ are monotonically increasing, we get $\tilde{U} := \lim_{k\to\infty} U_{\lambda_k}$, first in any L^q-space, then by elliptic theory in $W^{4,q}$, and finally in $C^{4,\alpha}(\overline{B})$. Hence, $\tilde{U}$ also solves (P_{λ_0}), and $0 < \tilde{U} \le U_{\lambda_0}$. We conclude that $\tilde{U} = U_{\lambda_0}$ by minimality of U_{λ_0}.

The second statement follows from the implicit function theorem and again the monotonicity of U_λ in λ. □

The next statement extends some results of [CR] to the biharmonic case; see Proposition 2.15 there. In order to show the sign condition of eigenfunctions, we use a decomposition method with respect to pairs of dual cones.

PROPOSITION 36. *Let u be a regular solution for (P_λ), where $\lambda \in (0,\lambda^*]$. Let the first eigenvalue μ_1 of the linearization $L_u := \Delta^2 - \lambda e^u$ under Dirichlet boundary conditions be nonnegative: $\mu_1 \ge 0$. Then every eigenfunction of $L_u\varphi = \mu_1\varphi$ is of fixed sign. Moreover, if $v \in C^4(\bar{B})$ solves $\Delta^2 v \ge \lambda e^v$ in B and $v = \frac{\partial v}{\partial \mathbf{n}} = 0$ on ∂B, then it follows that $v \ge u$. Finally, if $\mu_1 = 0$, then we even have $v = u$.*

Proof. In order to show that the first eigenfunction φ of L_u is of fixed sign, we need to explain a decomposition technique with respect to dual cones, which was found in the abstract setting by Moreau [Mo] and adapted to biharmonic Dirichlet problems in [GG]. As usual we equip $H_0^2(B)$ with the scalar product

$$(u,w)_{H_0^2} := \int_B \Delta u\, \Delta w\, dx.$$

Here, let

$$\mathcal{K} = \left\{ u \in H_0^2(B);\ u \ge 0 \ \text{ a.e. in } B \right\},$$

denote the convex closed cone of nonnegative H_0^2-functions and

$$\mathcal{K}' = \left\{ u \in H_0^2(B); \quad \text{for all } w \in \mathcal{K}: \quad (u,w)_{H_0^2} \le 0 \right\}$$

its dual cone in H_0^2 of weak subsolutions of the clamped plate equation. By Lemma 16 we see that $\mathcal{K}' \subset -\mathcal{K}$. For any $w \in \mathcal{K}'$ we even have that either $w \equiv 0$ or $w < 0$ in B.

Assume now by contradiction that φ is not of fixed sign. Then, according to [Mo], we may decompose

$$\varphi = \varphi_1 + \varphi_2$$

with $\varphi_1 \in \mathcal{K}$, $\varphi_2 \in \mathcal{K}'$, and $\varphi_1 \perp \varphi_2$ in $H_0^2(B)$. By assumption we have that $\varphi_1 \ge 0$, $\varphi_1 \not\equiv 0$, and $\varphi_2 < 0$. But then

$$\begin{aligned} 0 \le \mu_1 &= \inf_{w \in H_0^2(B)\setminus\{0\}} \frac{\int_B \left((\Delta w)^2 - \lambda \exp(u) w^2 \right) dx}{\int_B w^2\, dx} \\ &\le \frac{\int_B \left((\Delta(\varphi_1 - \varphi_2))^2 - \lambda \exp(u)(\varphi_1 - \varphi_2)^2 \right) dx}{\int_B (\varphi_1 - \varphi_2)^2\, dx} \end{aligned}$$

$$< \frac{\int_B \left((\Delta(\varphi_1+\varphi_2))^2 - \lambda \exp(u)(\varphi_1+\varphi_2)^2 \right) dx}{\int_B (\varphi_1+\varphi_2)^2\,dx}$$
$$= \frac{\int_B \left((\Delta\varphi)^2 - \lambda \exp(u)\varphi^2 \right) dx}{\int_B \varphi^2\,dx} = \mu_1,$$

a contradiction. Hence, φ is of fixed sign, say $\varphi \geq 0$, and in a second step we may conclude from the equation and the strict positivity of the biharmonic Green function (in the ball) that $\varphi > 0$.

We consider now u and v as in the statement. For $\tau \in [0,1]$ we look at

$$\begin{aligned} &\Delta^2 \left(u + \tau(v-u)\right) - \lambda \exp\left(u + \tau(v-u)\right) \\ &\geq \Delta^2 \left(u + \tau(v-u)\right) - \lambda \left(\tau \exp(v) + (1-\tau)\exp(u)\right) \geq 0. \end{aligned} \tag{59}$$

Since (59) equals 0 for $\tau = 0$, its first derivative at $\tau = 0$ must be nonnegative:

$$\Delta^2(v-u) - \lambda e^u (v-u) =: f \geq 0. \tag{60}$$

If $\mu_1 > 0$, a decomposition trick as above applied to the functional $w \mapsto \int_B ((\Delta w)^2 - \lambda e^u w^2 - fw)dx$ shows that $v \geq u$.

If $\mu_1 = 0$, we test (60) with the positive first eigenfunction φ and get

$$\Delta^2(v-u) - \lambda e^u (v-u) = 0.$$

That means that also the first derivative of (59) with respect to $\tau = 0$ vanishes, so that the second derivative needs to be nonnegative:

$$-\lambda e^u (v-u)^2 \geq 0.$$

But this immediately yields $v = u$. □

Concerning the stability behavior of the linearizations around regular solutions, we have the following.

PROPOSITION 37. *Let $\lambda > 0$, let u be a regular solution of (P_λ), let $L_u = \Delta^2 - \lambda e^u$ be the linearized operator at u, and let $\mu_1 = \mu_1(L_u)$ be the smallest eigenvalue of L_u; then*

(i) *if $\lambda < \lambda^*$ and u is the minimal solution, then $\mu_1 > 0$;*

(ii) *if $\lambda < \lambda^*$ and u is not the minimal solution, then $\mu_1 < 0$;*

(iii) *if $\lambda = \lambda^*$ and the extremal solution $u = U_*$ is regular, then $\mu_1 = 0$.*

Finally, if U_λ denotes the minimal (regular) solution of (P_λ) and $\mu_1(\lambda) = \mu_1(L_{U_\lambda})$, then the map $\lambda \mapsto \mu_1(\lambda)$ is decreasing.

Proof. (i) The monotonicity of $\mu_1(\lambda)$ follows immediately from the variational characterization

$$\mu_1(\lambda) = \inf_{w \in H_0^2(B)\setminus\{0\}} \frac{\int_B (\Delta w)^2\,dx - \int_B \exp(U_\lambda) w^2\,dx}{\int_B w^2\,dx}$$

and from the monotonicity of U_λ with respect to λ. By Proposition 35 we see that the function $\lambda \mapsto \mu_1(\lambda)$ is continuous from the left on $(0,\lambda^*)$ and even on $(0,\lambda^*]$, provided the extremal solution U_* is regular.

Assume by contradiction that there exists a $\tilde\lambda \in (0,\lambda^*)$ with $\mu_1(\tilde\lambda) \leq 0$. We put

$$\lambda_0 := \sup\{\lambda \geq 0 :\ \mu_1(\lambda) > 0\} \leq \tilde\lambda < \lambda^*.$$

According to the mentioned continuity from the left, we have $\mu_1(\lambda_0) \geq 0$. If we assume $\mu_1(\lambda_0) > 0$, then the second part of Proposition 35 would give $\mu_1(\lambda) > 0$ also for some $\lambda > \lambda_0$, a contradiction. Consequently we have $\mu_1(\lambda_0) = 0$. Let $u = U_{\lambda_0} > 0$ be the corresponding minimal solution:

$$\Delta^2 u = \lambda_0 e^u \text{ in } B, \quad u = \nabla u = 0 \text{ on } \partial B.$$

Consider any $\lambda \in (\lambda_0, \lambda^*)$ with minimal solution $v = U_\lambda > 0$:

$$\Delta^2 v = \lambda e^v \text{ in } B, \quad v = \nabla v = 0 \text{ on } \partial B.$$

Since $\lambda > \lambda_0$, Proposition 36 applies and yields $v = u$ and hence $\lambda = \lambda_0$, a contradiction.

(ii) Let U_λ be the minimal solution for (P_λ) so that $u \geq U_\lambda$. If the linearization around u had nonnegative first eigenvalue, then Proposition 36 would also yield $u \leq U_\lambda$ so that u and U_λ necessarily coincide, a contradiction.

(iii) Assume that the extremal solution $u = U_*$ is regular. By continuity, we have $\mu_1 \geq 0$. If $\mu_1 > 0$, the implicit function theorem would also yield solutions for some $\lambda > \lambda^*$. This is a contradiction, so that $\mu_1 = 0$. □

Open Problem 1. Does (ii) of Proposition 37 extend to weak solutions u as formulated in [BV, Theorem 3.1]?

We now turn to the extremal solution U_*. We first suggest the following open problem.

Open Problem 2. Do we have uniqueness of weak solutions for (P_{λ^*})? By Proposition 37(iii), and arguing as in Lemma 2.6 in [BV], one obtains that if the extremal solution is *regular*, then it is unique even in a weak sense. However, without the regularity assumption on U_*, the proof seems much more difficult; we refer to [Ma] for the corresponding result related to the second order problem (1). In particular, the proof of a result in the spirit of [Ma, Lemma 2.1] requires a new trick, probably of the same kind as the one we used to prove Lemma 20.

Perhaps, the precise characterization of *all* singular solutions U_σ and the corresponding "singular" parameters λ_σ is the most interesting and difficult problem we have to leave open in the present paper.

Open Problem 3. Are the singular parameter and the weakly singular solution unique? In order to construct a weakly singular radial solution, according to Theorem 6, one has to follow the unstable branch arising from P_2. One can do so in two (opposite) exit directions. In one direction we actually find at most (and presumably precisely) one solution by the result of Soranzo [So]: the solution of the PDE has to be strictly decreasing. We emphasize that this result extends to the class of *weakly singular* radial solutions. For the ODE system (8) this means that any "singular" trajectory may intersect the hyperplane $v_1 = 4$ only once and cannot come back to it. But we do not have a proof that the unstable branch leaving P_2 in the other direction will *not* intersect the hyperplane $v_1 = 4$ even if numerical experiments suggest so.

Next, we recall that in [GGM] it was shown that for any open bounded domain $\Omega \subset \mathbb{R}^n$ there exist $C_1, C_2 > 0$ such that the following improved Hardy inequality holds:

(61)
$$\int_\Omega |\Delta u|^2 \, dx \geq \frac{n^2(n-4)^2}{16} \int_\Omega \frac{u^2}{|x|^4} \, dx + C_1 \int_\Omega \frac{u^2}{|x|^2} \, dx + C_2 \int_\Omega u^2 \, dx \quad \text{for all } u \in H_0^2(\Omega).$$

A similar inequality was used in [BV] in order to establish the space dimensions in which the extremal solution for (1) is regular or singular. For (P_λ) this seems more intriguing: it is not clear which is the role of each of the remainder terms in (61). Furthermore, as we have seen in Theorem 12 and Proposition 34, the singular solution is difficult to describe. However, we have a partial result relating Hardy's inequality with extremal solutions: clearly, this statement is weaker than Corollary 10 if $n \le 10$.

PROPOSITION 38. *Let λ_σ and U_σ be as in Theorem 7 and assume that $\lambda_\sigma = \lambda^*$. Then if $n \le 12$, case* (ii) *in Theorem 12 cannot occur.*

Proof. By Proposition 37(i), by Theorem 3(ii)–(iii), and by using the notation of Theorem 12, we infer that

$$\int_B |\Delta\phi|^2 \ge \lambda^* \int_B e^{U_*}\phi^2 = \lambda^* \int_B \frac{e^W}{|x|^4}\phi^2 \qquad \text{for all } \phi \in H_0^2(B). \tag{62}$$

For contradiction, if (ii) in Theorem 12 holds, then

$$\lambda^* \int_B \frac{e^W}{|x|^4}\phi^2 \ge 8(n-2)(n-4)\int_B \frac{\phi^2}{|x|^4} \qquad \text{for all } \phi \in H_0^2(B).$$

Since $8(n-2)(n-4) > \frac{n^2(n-4)^2}{16}$ whenever $n \le 12$, the last inequality, together with (62), would improve the best constant in Hardy's inequality, a contradiction. ∎

Proposition 38 and Corollary 10 suggest the following question and conjecture.

Open Problem 4. Which are all the space dimensions $n \ge 5$ for which $\lambda_\sigma < \lambda^*$? We conjecture that the answer is $n \le 12$. In view of Corollary 10 we know that among these dimensions n, there are at least $5 \le n \le 10$. Moreover, Theorem 12 and Proposition 38 prove "half" of this conjecture when $n = 11, 12$. Maybe the proof relies on the interpretation of the two remainder terms in (61).

Open Problem 5. Show that any radial singular solution is also weakly singular, according to Definition 5. In particular, this would strengthen the statement of Theorem 6.

If the previous three open problems could be solved in the affirmative, then we could also conclude that the extremal solution U_* is singular if and only if $n \ge 13$.

We conclude this paper with some further problems. The next one is not yet completely solved even in the second order case.

Open Problem 6. Do there exist singular *nonradial* solutions to (P_λ) for some $\lambda > 0$? We conjecture that the answer is positive; see also Problem 7 in [BV].

Figure 2 displays the numerically computed value of $-v_4$ of the intersection of a portion of the unstable manifold of P_1 with the hyperplane $v_1 = 4$ in the case $n = 5$.

More precisely, $-v_4$ is displayed as a function of $x := -\log(-\vartheta)$. One may observe the estimated value of λ^* as the maximum value reached by $-v_4$; furthermore, as $\vartheta \to 0^-$ the value of $-v_4$ appears to asymptotically reach λ_σ oscillating around it. This leads us to the following problem.

Open Problem 7. Assume $n \le 12$. Prove that for every $N \in \mathbb{N}$ there exists $\varepsilon = \varepsilon(N) > 0$ such that for $\lambda \in [\lambda_\sigma - \varepsilon, \lambda_\sigma + \varepsilon]$ there exist at least N distinct regular radial solutions. For the second order problem the same statement holds true; see [GPP, Theorem 15].

Open Problem 8. How can one proceed in arbitrary smooth domains where it is known that comparison principles like Lemma 16 become false? How can one construct and *characterize* the *minimal* solution? Does one have similar bifurcation diagrams, where the solutions, however, can no longer be expected to be positive

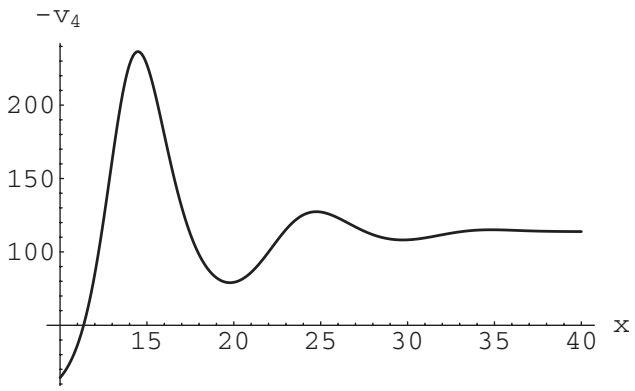

FIG. 2.

everywhere? Or does the lack of comparison principles lead to a completely different behavior, at least in geometrically very complicated domains?

9. Appendix. Computation techniques. We describe here the algorithm used in the computer assisted proofs. In order to prove Theorem 7 we need a rigorous estimate of the intersection of a branch of the unstable manifold with the hyperplane $v_1 = 4$. Since we do not know the exact location of any point of the manifold, except for the stationary point, we compute the trajectory of the whole set κ' as described in section 6. Since no analytical solution of the equation is available, we estimate the trajectories of all points of the set and compute the intersections with the hyperplane $v_1 = 4$ with rigorous error bounds. In order to compute the image of an infinite set of points, we partition it into boxes with small enough sides, which we call *interval sets*, and we compute their trajectories using interval arithmetics. More precisely, we start with a Taylor approximation of order 10; i.e., we estimate the trajectory of an interval by using the Taylor expansion of order 10 and estimate the error by the Lagrange remainder. If h is the time step, we compute a rough but rigorous enclosure D of the trajectory at times $[0, h]$, which is an interval set D such that the solution of the equation lies in D for all times between 0 and h. By Lagrange theorem we estimate the error we make neglecting the remaining terms of the Taylor expansion by computing $x^{(11)}(D)\frac{h^{11}}{11!}$. We compute $x^{(11)}(D)$ (which is an interval enclosing all possible values assumed by the 11th derivative of the trajectory, therefore enclosing the Lagrange remainder) using a recursive algorithm for the time derivatives of the solutions (see section I.8 in [HNW]). We point out that it takes a finite amount of s-time to go from any point in the set κ' (in Lemma 30) or κ (in Lemma 31) to the hyperplane $v_1 = 4$. The actual number and size of the intervals that we used as a partition of the sets κ' and κ can be read directly from the *Mathematica* notebook, together with the time step we used for the integration. We feel that it is pointless to display here the long list of numbers which represents such partitions, but since such a list is an essential part of the proof, we make it available in the *Mathematica* notebook.

The interval arithmetics algorithms address the problem of computing the trajectory of an interval and of keeping track of the errors in an elegant and rigorous way, but they introduce another problem. Indeed, even in the simplest dynamical system, the procedure described above leads to a very rough estimate of trajectories, due to the *wrapping effect* which makes the bounds on the error grow exponentially fast. The wrapping effect is one of the main problems one faces when trying to do

rigorous numerics for ODEs.

We describe it with one example: Consider a square centered at the origin $x = [-\delta, \delta]^2$ and the matrix that represents the rotation in $\mathbb{R}^2$ by an angle α,

$$R(\alpha) = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix}.$$

Assume for simplicity that $0 < |\alpha| < \pi/2$. If we apply R_α to x and wish to represent the result by another interval (i.e., another rectangle with sides parallel to the coordinate axes), we see that we need $(\cos(\alpha) + |\sin(\alpha)|)[-\delta, \delta]^2$; therefore, although R_α is an isometry, its computer realization has a growth factor $\cos(\alpha) + |\sin(\alpha)| > 1$. When solving the system of equations of the harmonic oscillator

$$\dot{x} = -y, \quad \dot{y} = x, \tag{63}$$

the 2π-shift along the trajectory is an identity map, but when we compute it numerically in interval arithmetics, say with time step $h = 2\pi/N$, we have to compose N times the map induced by $R(h)$. An easy computation shows this computation yields a growth factor $e^{2\pi} \approx 535$ as $h \to 0$.

We substantially reduce the wrapping effect by using the Lohner algorithm. A complete description of interval arithmetics and of the Lohner algorithm is beyond the scope of this paper; we refer to section 6 in [AZ] and the references cited therein for an exhaustive treatment of the topic. More specifically, see [MZ] concerning interval arithmetics and [Lo] for the Lohner algorithm. For the purpose of this description it suffices to consider the Lohner algorithm as a finite number of interval arithmetic operations based on the Taylor expansion which, given (8), an (interval) initial condition $V_0 \subset \mathbb{R}^4$, and a time step h, returns an interval $V_1 \subset \mathbb{R}^4$ such that for all points $v_0 \in V_0$ the solution $v(s)$ of the Cauchy problem with initial condition $v(0) = v_0$ satisfies $v(h) \in V_1$. In other words, the Lohner algorithm provides a rigorous enclosure of the solution at time h of a given Cauchy problem by performing a finite number of operations. The fact that the operations involved are in finite number and purely arithmetical (they are basically sums and multiplications, which can be performed with computer representable numbers with rigorous control on the round-off) makes it suitable for implementation with a computer.

We must determine the intersection of the trajectory with the hyperplane $v_1 = 4$. Since we are computing the trajectory of an interval, it takes a finite (nonzero) amount of "time" to cross the hyperplane; therefore we necessarily introduce another error when estimating the intersection point and have to give a rigorous bound for this error as well. We proceed as follows. We numerically compute the time s_1 required for the flow to reach the intersection. We compute with the Lohner algorithm the solution V_1 of the problem at time s_1. We check if the first component $(V_1)_1$ of V_1 is contained in $(-\infty, 4]$. If $(V_1)_1 \subset (-\infty, 4]$, then no points in V_1 have crossed the hyperplane. If $(V_1)_1 \not\subset (-\infty, 4]$, we choose (arbitrarily) a smaller value of s_1 and repeat the step. Then we roughly compute the time s_2 required for the set V_1 to cross the hyperplane. With the Lohner algorithm we compute the solution V_2 of the problem at time s_2. We check that all points in V_2 have crossed the hyperplane, i.e., $(V_2)_1 \subset [4, +\infty)$. If not, we choose a larger value for s_2 and repeat the step. We are interested only in the value of the fourth component of the solution: since at all points of our interest $v_4' < 0$ (because $v_4' = v_1 v_4$), it suffices to compute the hull of the interval value of v_4 before and after the crossing of the hyperplane. We now have a rigorous proof that the intersection takes place at some $v_4 \in [\min(V_2)_4, \max(V_1)_4]$, and this last interval

(with the left bound rounded down and the right bound rounded up) is the value we display in Table 1. For the λ^* computation we display only $\max(V_1)_4$ rounded down, since the other side of the interval does not have any meaning.

In order to check the third statement in Lemma 30, it is not enough to check that the evolution of all points is in A as defined in subsection 6.2. Indeed, if the time step is large, it may happen that some trajectory leaves A and reenters it in a single integration step. We have therefore to check at every time step that the whole rough enclosure D as defined above is in A and that the part of the set A which is contained in the flow tube has a trivial topology, i.e., it does not have holes. The round-off errors are taken care directly by suitable C++ procedures. Such errors may vary by changing computers and/or operating systems, but since they are usually very small when compared to the wrapping effect, we expect that the proofs can be easily reproduced on any recent computer obtaining very similar bounds.

To perform the proofs, we implemented a version of the whole algorithms in a combination of *Mathematica* 4.0 and C++ (gcc version 2.95.1) under the Linux operating system. More precisely, *Mathematica* was used to handle all the data and to perform a few algorithms which are less demanding for the CPU, but more complicated to implement. Furthermore *Mathematica* was used to make all numerical experiments and to draw the pictures. On the other hand C++ was used for the heavy interval arithmetic computations, where it offered much higher speed and more controllable accuracy. The connection between the two languages was obtained by MathLink. The verification of the whole proof takes a few days of CPU time on a machine equipped with an Athlon XP1700 processor. The computer programs which are part of the proofs can be obtained from the authors upon request, while the interval algorithms are provided by [CAPD].

Acknowledgments. We are grateful to Anna Dall'Acqua (TU Delft) and Elvise Berchio (Università di Torino) for their careful reading and for helpful remarks. We are also grateful to the referees for their careful review of the paper and for some very useful suggestions on the presentation.

REFERENCES

[ADN] S. Agmon, A. Douglis, and L. Nirenberg, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions.* I, Comm. Pure Appl. Math., 12 (1959), pp. 623–727.

[A1] H. Amann, *Fixed point equations and nonlinear eigenvalue problems in ordered Banach spaces*, SIAM Rev., 18 (1976), pp. 620–709.

[A2] H. Amann, *Ordinary Differential Equations. An Introduction to Nonlinear Analysis*, de Gruyter, Berlin, 1990.

[AZ] G. Arioli and P. Zgliczyński, *Symbolic dynamics for the Hénon–Heiles Hamiltonian on the critical level*, J. Differential Equations, 171 (2001), pp. 173–202.

[BE] J. Bebernes and D. Eberly, *Mathematical Problems from Combustion Theory*, Appl. Math. Sci. 83, Springer, New York, 1989.

[B] T. Boggio, *Sulle funzioni di Green d'ordine m*, Rend. Circ. Mat. Palermo, 20 (1905), pp. 97–135.

[BCMR] H. Brezis, T. Cazenave, Y. Martel, and A. Ramiandrisoa, *Blow up for $u_t - \Delta u = g(u)$ revisited*, Adv. Differential Equations, 1 (1996), pp. 73–90.

[BV] H. Brezis and J. L. Vazquez, *Blow-up solutions of some nonlinear elliptic problems*, Rev. Mat. Univ. Complut. Madrid, 10 (1997), pp. 443–468.

[C] S. Chandrasekhar, *An Introduction to the Study of Stellar Structure*, Dover, New York, 1967.

[CAPD] *Computer Assisted Proofs in Dynamics*, http://capd.wsb-nlu.edu.pl.

[CR] M. C. Crandall and P. H. Rabinowitz, *Some continuation and variational methods for positive solutions of nonlinear elliptic eigenvalue problems*, Arch. Ration. Mech. Anal., 58 (1975), pp. 207–218.

[E] U. Elias, *Nonoscillation and eventual disconjugacy*, Proc. Amer. Math. Soc., 66 (1977), pp. 269–275.

[GMP] T. Gallouet, F. Mignot, and J. P. Puel, *Quelques résultats sur le problème* $-\Delta u = \lambda e^u$, C. R. Acad. Sci. Paris Sér. I Math., 307 (1988), pp. 289–292.

[GPP] J. García Azorero, I. Peral Alonso, and J. P. Puel, *Quasilinear problems with exponential growth in the reaction term*, Nonlinear Anal., 22 (1994), pp. 481–498.

[GG] F. Gazzola and H.-Ch. Grunau, *Critical dimensions and higher order Sobolev inequalities with remainder terms*, NoDEA Nonlinear Differential Equations Appl., 8 (2001), pp. 35–44.

[GGM] F. Gazzola, H.-Ch. Grunau, and E. Mitidieri, *Hardy inequalities with optimal constants and remainder terms*, Trans. Amer. Math. Soc., 356 (2004), pp. 2149–2168.

[G] I. M. Gel'fand, *Some problems in the theory of quasilinear equations*, Amer. Math. Soc. Transl. (2), 29 (1963), pp. 295–381; translated from the Russian, Uspekhi Mat. Nauk, 14 (1959), pp. 87–158.

[GNN] B. Gidas, W. M. Ni, and L. Nirenberg, *Symmetry and related properties via the maximum principle*, Comm. Math. Phys., 68 (1979), pp. 209–243.

[GS] H.-Ch. Grunau and G. Sweers, *Positivity properties of elliptic boundary value problems of higher order*, Nonlinear Anal., 30 (1997), pp. 5251–5258.

[HNW] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations* I: *Nonstiff problems*, 2nd ed., Springer, New York, 2000.

[J] J. Jacobsen, *A Liouville–Gelfand equation for k-Hessian operators*, Rocky Mountain J. Math., 34 (2004), pp. 665–684.

[JS] J. Jacobsen and K. Schmitt, *The Liouville–Bratu–Gelfand problem for radial operators*, J. Differential Equations, 184 (2002), pp. 283–298.

[JL] D. Joseph and T. S. Lundgren, *Quasilinear Dirichlet problems driven by positive sources*, Arch. Ration. Mech. Anal., 49 (1973), pp. 241–269.

[Li] P.-L. Lions, *On the existence of positive solutions of semilinear elliptic equations*, SIAM Rev., 24 (1982), pp. 441–467.

[Lo] R. J. Lohner, *Computation of guaranteed enclosures for the solutions of ordinary initial and boundary value problems*, in Computational Ordinary Differential Equations, J. R. Cash and I. Gladwell, eds., Clarendon Press, Oxford, 1992, pp. 425–435.

[Ma] Y. Martel, *Uniqueness of weak extremal solutions for nonlinear elliptic problems*, Houston J. Math., 23 (1997), pp. 161–168.

[MP1] F. Mignot and J. P. Puel, *Sur une classe de problèmes non linéaires avec nonlinéarité positive, croissante, convexe*, Comm. Partial Differential Equations, 5 (1980), pp. 791–836.

[MP2] F. Mignot and J. P. Puel, *Solution radiale singulière de* $-\Delta u = \lambda e^u$, C. R. Acad. Sci. Paris Sér. I Math., 307 (1988), pp. 379–382.

[Mo] J. J. Moreau, *Décomposition orthogonale d'un espace hilbertien selon deux cônes mutuellement polaires*, C. R. Acad. Sci. Paris, 255 (1962), pp. 238–240.

[MZ] M. Mrozek and P. Zgliczyński, *Set arithmetic and the enclosing problem in dynamics*, Ann. Polon. Math., 74 (2000), pp. 237–259.

[PT] L. A. Peletier and W. C. Troy, *Spatial patterns. Higher order models in physics and mechanics*, Progr. Nonlinear Differential Equations Appl. 45, Birkhäuser, Boston, Boston, MA, 2001.

[P] S. I. Pohožaev, *Eigenfunctions of the equation* $\Delta u + \lambda f(u) = 0$, Soviet Math. Dokl., 6 (1965), pp. 1408–1411.

[PS] P. Pucci and J. Serrin, *A general variational identity*, Indiana Univ. Math. J., 35 (1986), pp. 681–703.

[So] R. Soranzo, *A priori estimates and existence of positive solutions of a superlinear polyharmonic equation*, Dynam. Systems Appl., 3 (1994), pp. 465–487.

[Sw] G. Sweers, *No Gidas–Ni–Nirenberg type result for biharmonic problems*, Math. Nachr., 246/247 (2002), pp. 202–206.

[We] J. Wei, *Asymptotic behavior of a nonlinear fourth order eigenvalue problem*, Comm. Partial Differential Equations, 21 (1996), pp. 1451–1467.

[Wi] Ch. Wieners, *Numerische Existenzbeweise für Schwache Lösungen Nichtlinearer Elliptischer Randwertaufgaben (Numerical proofs of existence of weak solutions of nonlinear elliptic boundary value problems)*, Ph.D. thesis, University of Cologne, 1994.

SIAM J. MATH. ANAL.
Vol. 36, No. 4, pp. 1259–1281

ON A STOKES-LIKE SYSTEM FOR MIXTURES OF FLUIDS*

JENS FREHSE†, SONJA GOJ‡, AND JOSEF MÁLEK§

Abstract. We consider a simplified model that describes steady flows of a miscible mixture of fluids. The corresponding system of equations is studied in the whole space $\mathbb{R}^3$. The densities ρ_i of the species and their velocity fields $u^{(i)}$ are prescribed at infinity: $\rho_i|_\infty = \rho_{i\infty} > 0\,, \quad u^{(i)}|_\infty = 0$. We prove the existence of weak solutions to the system under consideration.

Key words. mixtures of fluids, Stokes-like equations, compressible fluid, weak solution

AMS subject classifications. 35Q30, 76N10

DOI. 10.1137/S0036141003433425

1. Introduction.

1.1. Description of the model. There are various approaches to model the behavior of mixtures. In this paper we deal with a continuum mechanics model describing flows of the mixture of N compressible fluids (we call such a material *N-component* or *multicomponent fluid*) at constant temperature. Based on the assumption that the mixture is sufficiently dense, it is reasonable to assume that at each point of the space occupied by the mixture there are particles belonging to each component.

Following the theory introduced in [24] that generalizes the earlier pioneering works by Fick [10], Darcy [2], and later on by Truesdell [26], isothermal flows of such a material can be fully captured by $2N$ functions $(\rho_i, u^{(i)}), i = 1, \dots, N$, representing the densities ρ_i and the velocities $u^{(i)}$ of the ith component of the mixture, and solving the equations

$$(\rho_i)_t + \operatorname{div}\left(\rho_i u^{(i)}\right) = 0\,, \tag{1.1}$$

$$\left(\rho_i u^{(i)}\right)_t + \operatorname{div}\left(\rho_i u^{(i)} \otimes u^{(i)}\right) - \operatorname{div} T^{(i)} = \rho_i f^{(i)} + J^{(i)}\,, \tag{1.2}$$

where $f^{(i)}$ is a given external force, $T^{(i)}$ represents the partial Cauchy stress, and $J^{(i)}$ is an interaction term (the momentum source); all quantities are associated to the ith constituent. Note that (1.1) and (1.2) follow from the *balance of mass* (no chemical reactions) and the *balance of linear momentum* for each constituent.

Just for simplicity we "restrict" ourselves to a two-component fluid, i.e., $N = 2$. Then the principle of action and reaction implies that

$$J^{(1)} = -J^{(2)}\,. \tag{1.3}$$

*Received by the editors August 17, 2003; accepted for publication (in revised form) May 14, 2004; published electronically February 3, 2005. This work was supported by the SFB 611 at the University of Bonn and the European HYKE network (contract HPRN–CT–2002–00282).

http://www.siam.org/journals/sima/36-4/43342.html

†Institute for Applied Mathematics, University of Bonn, Beringstr. 6, 53115 Bonn, Germany (erdbeere@iam.uni-bonn.de). This author was also supported by a cooperation program of Bonn and Prague University.

‡Institute for Applied Mathematics, University of Bonn, Beringstr. 6, 53115 Bonn, Germany (goj@mailcip.iam.uni-bonn.de).

§Mathematical Institute, Charles University, Sokolovská 83, 18675 Prague 8, Czech Republic (malek@karlin.mff.cuni.cz). This author was also supported by the project GACR 201/00/0768 and a cooperation program of Bonn and Prague University.

Let further ψ_i denote the Helmholtz potential associated to the ith constituent. Then the *balance of entropy* for the total mixture and *the second law of thermodynamics* imply (recall that the temperature is supposed to be constant) that

$$\sum_{i=1}^{2} T^{(i)} \cdot \nabla u^{(i)} + J^{(1)} \cdot \left(u^{(2)} - u^{(1)}\right) - \sum_{i=1}^{2} \left[\rho_i (\psi_i)_t + \rho_i u_k^{(i)} (\psi_i)_{x_k}\right] \geq 0\,, \tag{1.4}$$

where we also used (1.3). Inequality (1.4) helps to identify the structure of the constitutive equations for $T^{(i)}$ and $J^{(i)}$.

We start with the assumption that the energy storage mechanism is the same for each constituent, i.e., $\psi_1 = \psi_2$, and for $i = 1, 2$ we have

$$\psi_i = \Psi(\rho_1 + \rho_2) \quad \text{or more generally} \quad \psi_i = \tilde{\Psi}\left(\frac{\rho_1}{\rho_{1,ref}} + \frac{\rho_2}{\rho_{2,ref}}\right), \tag{1.5}$$

where $\rho_{i,ref}$, $i = 1, 2$, are positive reference densities. Considering the former for simplicity, inserting it into (1.4), using also (1.1), and setting

$$P_i(\rho) = P_i(\,(\rho_1, \rho_2)^T\,) = \rho_i(\rho_1 + \rho_2)\,\Psi'(\rho_1 + \rho_2)\,, \tag{1.6}$$

we eventually arrive at

$$\sum_{i=1}^{2} \left[T^{(i)} + P_i(\rho) I\right] \cdot \nabla u^{(i)} + \left[J^{(1)} + \Psi'(\rho_1 + \rho_2)\,(\rho_1 \nabla \rho_2 - \rho_2 \nabla \rho_1)\right] \cdot (u^{(2)} - u^{(1)}) \geq 0\,, \tag{1.7}$$

where I denotes the identity tensor. Put

$$\sigma^{(i)} := T^{(i)} + P_i(\rho) I \quad \text{and} \quad G := J^{(1)} + \Psi'(\rho_1 + \rho_2)\,(\rho_1 \nabla \rho_2 - \rho_2 \nabla \rho_1)\,. \tag{1.8}$$

Then, denoting $D(w) := \frac{1}{2}\left(\nabla w + (\nabla w)^T\right)$ for $w\colon \mathbb{R}^3 \to \mathbb{R}^3$, we observe that setting

$$\begin{aligned} \sigma^{(i)} &= \mu_{i1} D(u^{(1)}) + \mu_{i2} D(u^{(2)}) + \nu_{i1} \operatorname{div} u^{(1)} I + \nu_{i2} \operatorname{div} u^{(2)} I\,, \\ G &= a(\rho_1, \rho_2, |u^{(1)} - u^{(2)}|)(u^{(2)} - u^{(1)}) \end{aligned}$$

and requiring[1] that for a certain $c_0 > 0$

$$\sum_{i=1}^{2} \sigma^{(i)} \cdot \nabla u^{(i)} \geq c_0 |\nabla u|^2 \qquad \text{and} \qquad a(\rho_1, \rho_2, |u^{(1)} - u^{(2)}|) \geq 0\,, \tag{1.9}$$

(1.7) is automatically fulfilled. This means that the system (1.1)–(1.2) with

$$T^{(i)} = -P_i(\rho) I + \sigma^{(i)}, \tag{1.10}$$

$$J^{(1)} = a(\rho_1, \rho_2, |u^{(1)} - u^{(2)}|)(u^{(2)} - u^{(1)}) - \Psi'(\rho_1 + \rho_2)\,(\rho_1 \nabla \rho_2 - \rho_2 \nabla \rho_1) \tag{1.11}$$

[1] In terms of the viscosities, $(1.9)_1$ is equivalent to

$$\begin{gathered} \mu_{11} > 0\,, \quad \mu_{22} > 0\,, \quad 2\mu_{11} + \nu_{11} > 0\,, \quad 2\mu_{22} + \nu_{22} > 0\,, \\ 4\mu_{11}\mu_{22} - (\mu_{12} + \mu_{21})^2 > 0\,, \\ 4(2\mu_{11} + \nu_{11})(2\mu_{22} + \nu_{22}) - (2\mu_{12} + \nu_{12} + 2\mu_{21} + \nu_{21})^2 > 0\,. \end{gathered}$$

is *thermomechanically consistent*. It means the basic energy estimates are in place.

We, however, consider an approximation of this system neglecting the second term in (1.11). In other words, we assume that the momentum source due to $\rho_1\nabla\rho_2-\rho_2\nabla\rho_1$ is of much lower order than the effects due to the difference between the velocities of the constituents. Numerical simulations of flows in special geometries have also shown that there is no significant difference in the resulting flows regardless of whether the second term in (1.11) is considered or neglected (see also [25]). One may also argue that it is difficult to identify the second term in (1.11) experimentally. Thus, in what follows the interaction terms take the form[2]

$$J^{(1)} = -J^{(2)} = a(\rho_1, \rho_2, |u^{(1)} - u^{(2)}|)\,(u^{(2)} - u^{(1)})\,. \tag{1.12}$$

However, once we accept (1.12) instead of (1.11), the basic energy identity is lost. We will return to this issue in subsection 1.4.

We also restrict ourselves to steady flows, i.e., $(\rho_i)_t = 0$ and $(\rho_i u^{(i)})_t = 0$ in (1.1) and (1.2).

Finally, we neglect the convective term $\operatorname{div}(\rho_i u^{(i)} \otimes u^{(i)})$ in (1.2), which may be justified either geometrically (there are flows in special geometries where $\operatorname{div}(\rho u\otimes u) = 0$) or via a proper scaling that leads to a nondimensional form which allows us to neglect the convective term for slow flows. Since the full system is very complex, our motivation to neglect the convective term has been rather technical: to start with the investigation of a simpler yet interesting system first. Thus, in analogy to the mathematical theory for incompressible fluids, we start with the *Stokes-like system for a mixture of fluids*.

To summarize, we end up with the system ($i = 1, 2$)

$$\operatorname{div}\left(\rho_i u^{(i)}\right) = 0\,, \tag{1.13}$$

$$L^{(i)}u = -\nabla P_i(\rho) + \rho_i f^{(i)} + J^{(i)}, \tag{1.14}$$

where $J^{(i)}$ is of the form (1.12), $L^{(i)}u$ is defined through

$$L^{(i)}u = -\operatorname{div}\sigma^{(i)} = -\sum_{k=1}^{2}(\mu_{ik}\Delta u^{(k)} + (\mu_{ik} + \nu_{ik})\nabla \operatorname{div} u^{(k)})\,, \tag{1.15}$$

and the constant coefficients μ_{ik} and ν_{ik} and the function a fulfill (1.9).

The structure of $P_i(\rho)$ comes from (1.6). To give an example, assuming that $\Psi = (\rho_1 + \rho_2)^{\gamma-1}$ in (1.5) with $\gamma > 1$, we obtain

$$P_i(\rho) = c_i\rho_i\,(\rho_1 + \rho_2)^{\gamma-1} \qquad \text{with} \quad \gamma > 1 \qquad (c_i > 0)\,. \tag{1.16}$$

In this paper we study (1.13)–(1.14) in the whole space $\mathbb{R}^3$, assuming that for given positive $\rho_{i\infty}$, $i = 1, 2$, the following holds:

$$u^{(i)} \to 0 \quad \text{and} \quad \rho_i \to \rho_{i\infty} \qquad \text{as} \quad |x| \to \infty \quad (i = 1, 2)\,. \tag{1.17}$$

We are interested in proving the existence of solutions to (1.13)–(1.14) and (1.17), completed with, say, (1.12), (1.15) with $(1.11)_1$, and (1.16). The precise assumptions on $P_i(\rho)$ and $J^{(i)}$ are given in subsection 1.2. The precise formulation of the main result is postponed to subsection 1.3.

[2]The main reason we prefer (1.12) rather than (1.11) comes of course from the analysis of the system: We have not been able to find sufficient information in order to pass to the limit in the nonlinear terms containing $\nabla\rho_i$.

1.2. Effective viscous flux and assumptions on P_i and $J^{(i)}$. Applying div to (1.14) we obtain

$$(-\Delta)\left(\sum_{k=1}^{2}(2\mu_{ik}+\nu_{ik})\operatorname{div}u^{(k)}-P_i(\rho)\right)=\operatorname{div}\left(\rho_i f^{(i)}+J^{(i)}\right),$$

which can be rewritten as

$$\begin{pmatrix} 2\mu_{11}+\nu_{11} & 2\mu_{12}+\nu_{12} \\ 2\mu_{21}+\nu_{21} & 2\mu_{22}+\nu_{22} \end{pmatrix}\begin{pmatrix} \operatorname{div}u^{(1)} \\ \operatorname{div}u^{(2)} \end{pmatrix}-\begin{pmatrix} P_1(\rho) \\ P_2(\rho) \end{pmatrix}=\begin{pmatrix} F^{(1)} \\ F^{(2)} \end{pmatrix}, \tag{1.18}$$

where

$$F^{(i)}=(-\Delta)^{-1}\operatorname{div}\left(\rho_i f^{(i)}+J^{(i)}\right). \tag{1.19}$$

Alternatively, (1.18) can be viewed as

$$\begin{pmatrix} \operatorname{div}u^{(1)} \\ \operatorname{div}u^{(2)} \end{pmatrix}=\begin{pmatrix} 2\mu_{11}+\nu_{11} & 2\mu_{12}+\nu_{12} \\ 2\mu_{21}+\nu_{21} & 2\mu_{22}+\nu_{22} \end{pmatrix}^{-1}\left[\begin{pmatrix} P_1(\rho) \\ P_2(\rho) \end{pmatrix}+\begin{pmatrix} F^{(1)} \\ F^{(2)} \end{pmatrix}\right]. \tag{1.20}$$

Since the matrix on the right-hand side of (1.20) is not necessarily symmetric, one may choose β_0 such that A_0 defined through

$$A_0=\begin{pmatrix} \beta_0 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 2\mu_{11}+\nu_{11} & 2\mu_{12}+\nu_{12} \\ 2\mu_{21}+\nu_{21} & 2\mu_{22}+\nu_{22} \end{pmatrix}^{-1} \tag{1.21}$$

is symmetric, and positive definite if $2\mu_{ik}+\nu_{ik}>0$. With this notation (1.20) results in the form

$$\begin{pmatrix} \beta_0\operatorname{div}u^{(1)} \\ \operatorname{div}u^{(2)} \end{pmatrix}=A_0\left[\begin{pmatrix} P_1(\rho) \\ P_2(\rho) \end{pmatrix}+\begin{pmatrix} F^{(1)} \\ F^{(2)} \end{pmatrix}\right]. \tag{1.22}$$

This equation will play a key role in our analysis. Analogously to the theory of a single compressible fluid we call (1.22) the *equation for the effective viscous flux*.

Now, we can formulate the assumptions on $P_i(\rho)$. We suppose that there are $\gamma>1$ and $\beta_0\neq 0$ such that $P(\rho)=(P_1(\rho),P_2(\rho))^T$ fulfills

(i) *the monotonicity condition*:
there is $\lambda_0>0$ such that for all $\rho=(\rho_1,\rho_2)^T$ and $\hat\rho=(\hat\rho_1,\hat\rho_2)^T$, $\rho_i,\hat\rho_i\geq 0$,

$$(\rho-\hat\rho)^T A_0\left(P(\rho)-P(\hat\rho)\right)\geq\lambda_0\left(|\rho|^{\gamma-1}+|\hat\rho|^{\gamma-1}\right)|\rho-\hat\rho|^2; \tag{1.23}$$

(ii) *the coerciveness condition*:
there are $\lambda_1 > 0, \lambda_2 \in \mathbb{R}$ such that for all $\rho = (\rho_1, \rho_2)^T, \rho_i \geq 0$,

$$(\rho^\gamma - \rho_\infty^\gamma)^T A_0 \left(P(\rho) - P(\rho_\infty)\right) \geq \lambda_1 |\rho - \rho_\infty|^{2\gamma} - \lambda_2 |\rho - \rho_\infty|^2 ; \tag{1.24}$$

(iii) *the growth condition*:
there is $K > 0$ such that for all $\rho = (\rho_1, \rho_2)^T$ and $\hat{\rho} = (\hat{\rho}_1, \hat{\rho}_2)^T$

$$|P(\rho) - P(\hat{\rho})| \leq K \left(|\rho|^{\gamma-1} + |\hat{\rho}|^{\gamma-1}\right) |\rho - \hat{\rho}| . \tag{1.25}$$

In the appendix of [12], we show that $P(\rho)$ of the form (1.16) fulfills the coerciveness condition (1.24). Similarly, one can show that $P(\rho)$ of the form (1.16) fulfills (1.23) under reasonable conditions[3] on γ. Note also that (1.25) is satisfied by (1.16) as well.

Next, we give the assumptions on the interaction terms. We suppose that $J^{(1)}$ and $J^{(2)}$ satisfy (1.12) and

$$\begin{aligned} &\text{there are } \theta \in (0,1) \text{ and } K > 0 \text{ such that} \\ &|J^{(1)}| = |J^{(2)}| = |a(x, \rho, |u^{(1)} - u^{(2)}|)(u^{(2)} - u^{(1)})| \leq K(1 + |u^{(1)} - u^{(2)}|)^\theta \end{aligned} \tag{1.26}$$

and there is $R_0 \gg 1$ such that

$$a \in C(\mathbb{R}^3 \times \mathbb{R}_0^+ \times \mathbb{R}_0^+) \text{ and } a(x, \cdot, \cdot) = 0 \text{ if } |x| \geq R_0 . \tag{1.27}$$

Sublinear growth of $J^{(i)}$, $i = 1, 2$, seems to be crucial in our method in order to deduce estimates for ρ, while the assumption concerning the compact support of a is useful in order to avoid technical difficulties concerning the asymptotic behavior of $u^{(1)}, u^{(2)}$ as $|x| \to \infty$.

We complete this part with two inequalities that follow from (1.23) and (1.25). First of all, taking $\hat{\rho} = \rho_\infty$ in (1.25) and using some elementary considerations, we obtain that there is a $K = K(\rho_\infty) > 0$ such that for all $\rho = (\rho_1, \rho_2)^T$

$$|P(\rho) - P(\rho_\infty)| \leq K(\rho_\infty) \left(|\rho - \rho_\infty|^\gamma + |\rho - \rho_\infty|\right) . \tag{1.28}$$

Also, there is $\hat{\lambda}_0$ such that for all $\rho = (\rho_1, \rho_2)^T$

$$(\rho - \rho_\infty)^T A_0 \left(P(\rho) - P(\rho_\infty)\right) \geq \hat{\lambda}_0 \left(|\rho - \rho_\infty|^{\gamma+1} + |\rho - \rho_\infty|^2\right) , \tag{1.29}$$

which is a consequence of (1.23) and the algebraic inequality

$$(g + h)^{\gamma-1} \geq \frac{1}{2^{\gamma-1}} g^{\gamma-1} - h^{\gamma-1} \qquad (g, h \geq 0) \tag{1.30}$$

---

[3] These conditions will be addressed in [14]. Let us mention two special cases where (1.23) holds.
(i) $\mu_{ij} > 0$ ($\Longrightarrow$ $A_{ij} < 0$, $i \neq j$, where $A_0 = (A_{ij})$) and for all $r \in \mathbb{R}$ with $(B_{ij}) = A_0 \left(\begin{smallmatrix} c_1 & 0 \\ 0 & c_2 \end{smallmatrix}\right)$:

$$\begin{aligned} &(B_{11} + B_{12}(\gamma - 1))r^2 + (B_{21} + \gamma B_{12} + B_{22}(\gamma - 1))r + \gamma B_{22} \geq \lambda_0(1 + r^2) , \\ &(B_{22} + A_{21}(\gamma - 1))r^2 + (B_{12} + \gamma B_{21} + B_{11}(\gamma - 1))r + \gamma B_{11} \geq \lambda_0(1 + r^2) . \end{aligned}$$

These conditions are fulfilled if γ is not too large, say, $\gamma \leq 2$, and the diagonal terms of A_0 dominate the nondiagonal ones.
(ii) $A_{11} = A_{22}$, $A_{12} = A_{21}$, $A_{12}/A_{11} = \chi \leq 0$, and $\gamma(1 + \chi(\gamma - 1)) - 1/4(\chi(\gamma + 1) + \gamma - 1)^2 > 0$.

that follows from $g^{\gamma-1} \le (g+h-h)^{\gamma-1} \le 2^{\gamma-1}\Big((g+h)^{\gamma-1}+h^{\gamma-1}\Big)$. To see that (1.29) holds, we take $\hat\rho = \rho_\infty$ in (1.23) and compute

$$\begin{aligned}(\rho-\rho_\infty)^T A_0\left(P(\rho)-P(\rho_\infty)\right) &\ge \lambda_0\left(|\rho-\rho_\infty+\rho_\infty|^{\gamma-1}+|\rho_\infty|^{\gamma-1}\right)|\rho-\rho_\infty|^2\\ &\ge \frac{\lambda_0}{2}\left(|\rho-\rho_\infty+\rho_\infty|^{\gamma-1}+|\rho_\infty|^{\gamma-1}\right)|\rho-\rho_\infty|^2+\frac{\lambda_0}{2}|\rho_\infty|^{\gamma-1}|\rho-\rho_\infty|^2\\ &\overset{(1.30)}{\ge} \frac{\lambda_0}{2}\left(\frac{1}{2^{\gamma-1}}|\rho-\rho_\infty|^{\gamma-1}\right)|\rho-\rho_\infty|^2+\frac{\lambda_0}{2}|\rho_\infty|^{\gamma-1}|\rho-\rho_\infty|^2,\end{aligned}$$

which implies (1.29) with $\hat\lambda_0 = \min\{\lambda_0/2^\gamma, \lambda_0|\rho_\infty|^{\gamma-1}/2\}$.

1.3. Formulation of the result. First, we fix the notation of the function spaces we will use. The symbols $L^q(\mathbb{R}^3)$, or L^q in short, $1 \le q < \infty$, denote the standard Lebesgue spaces of scalar-valued measurable functions. The space norm is denoted by $\|u\|_{L^q}^q = \int_{\mathbb{R}^3}|u(x)|^q\,dx$. If $X(\mathbb{R}^3)$ is a space of (scalar) functions, then $X(\mathbb{R}^3;\mathbb{R}^3)$ denotes the space of all $w = (w_1,w_2,w_3)^T\colon \mathbb{R}^3 \to \mathbb{R}^3$ such that $w_j \in X(\mathbb{R}^3)$, $j=1,2,3$.

We will further use the space $H_0^1 := H_0^1(\mathbb{R}^3;\mathbb{R}^3)$ defined as the closure of the space of smooth functions with compact support in $\mathbb{R}^3$ with respect to the norm $\|u\|_{H_0^1} = \|\nabla u\|_{L^2}$. Recall that $H_0^1 \hookrightarrow L^6$, but H_0^1 is not embedded into L^2.

We formulate our result.

THEOREM 1.1. *Let $\rho_\infty = (\rho_{1\infty},\rho_{2\infty})^T$, $\rho_{1\infty},\rho_{2\infty} > 0$, be given. Let*

$$f^{(i)} \in L^\infty(\mathbb{R}^3;\mathbb{R}^3) \text{ and have compact support.} \tag{1.31}$$

Assume that ν_{ik},μ_{ik} forming the operators $L^{(i)}$ according to (1.15) *fulfill the condition* $(1.9)_1$. *Assume further that $P(\rho) = (P_1(\rho),P_2(\rho))^T$ satisfies* (1.23)–(1.25) *with $\gamma > 1$. Finally, assume that $J^{(i)}$, $i=1,2$, of the form* (1.12) *fulfill* (1.26)–(1.27). *Then there is a weak solution (ρ,u), $\rho = (\rho_1,\rho_2)^T$, $u = (u^{(1),T},u^{(2),T})^T$, such that for $i=1,2$*

$$\rho_i \ge 0\,, \tag{1.32}$$

$$\rho_i - \rho_{i\infty} \in L^2(\mathbb{R}^3)\cap L^{2\gamma}(\mathbb{R}^3)\,, \tag{1.33}$$

$$u^{(i)} \in H_0^1(\mathbb{R}^3;\mathbb{R}^3)\,, \tag{1.34}$$

solving the equations (1.13)–(1.14) *in a distributional sense.*

The proof of Theorem 1.1 is performed in the consequent sections via the following scale of approximations.

For $\alpha > 0$, $\beta > 0$, $\sigma > 0$, and $s_0 > \max\{4,2\gamma\}$, we consider the system ($i=1,2$)

$$-\sigma\Delta\rho_i + \alpha\left(1+|\rho_i-\rho_{i\infty}|^{s_0-2}\right)(\rho_i-\rho_{i\infty}) + \operatorname{div}\left(\rho_i u^{(i)}\right) = 0, \tag{1.35}$$

$$L^{(i)}u = -\nabla\left(w_\beta(P_i(\rho)-P_i(\rho_\infty))\right) + \rho_i f^{(i)} + (-1)^{i+1}J^{(1)} \quad \text{with } w_\beta = (1+\beta|x|)^{-4}. \tag{1.36}$$

Clearly, if we set $\sigma = 0$, $\alpha = 0$, and $\beta = 0$ in (1.35) and (1.36), we obtain (1.13)–(1.14).

In section 2 we prove the existence of weak solutions to (1.35)–(1.36) in $\mathbb{R}^3$. In section 3, we derive the equation for the effective viscous flux related to this approximation. Using this tool, we can then derive estimates for the densities and the velocities that are uniform with respect to β. This allows us to show not only that there is a weak solution to (1.35) and (1.36) with w_β replaced by 1, but also that this solution satisfies certain types of inequalities, from which estimates uniform with respect to α can be derived. The derivation of these estimates is performed in section 4, together with the passage to the limit as $\alpha \to 0$. Finally, in section 5, we let $\sigma \to 0$.

1.4. Main difficulties. First of all, we would like to point out a remarkable difference between the analysis of the system (1.13)–(1.14) describing the steady flow of two miscible fluids on the one hand and the analysis of the analogous model for a one-constituent compressible fluid on the other hand. This difference manifests itself in the first step, i.e., in the derivation of a priori estimates. To be more precise, consider the equations for one single fluid that are analogous to (1.13)–(1.14):

$$\operatorname{div}(\varrho v) = 0 \quad \text{in } \mathbb{R}^3, \tag{1.37}$$

$$-\mu\Delta v - (\lambda+\mu)\nabla \operatorname{div} v = -\nabla \varrho^\gamma + \varrho f \quad \text{in } \mathbb{R}^3, \tag{1.38}$$

$$v \to 0 \quad \text{and} \quad \varrho \to \varrho_\infty \quad \text{as } |x| \to \infty\,, \tag{1.39}$$

with $\varrho\colon \mathbb{R}^3 \to \mathbb{R}$, $v\colon \mathbb{R}^3 \to \mathbb{R}^3$, and a given $\varrho_\infty > 0$. Multiplying (1.38) by v and integrating by parts, we observe that the left-hand side controls the L^2-norm of the velocity gradient, and the first term on the right-hand side leads to

$$-\int \nabla \varrho^\gamma v\,dx = -\frac{\gamma}{\gamma-1}\int \varrho v\cdot\nabla\varrho^{\gamma-1}\,dx = -\frac{\gamma}{\gamma-1}\int \operatorname{div}(\varrho v)\,\varrho^{\gamma-1}\,dx = 0\,, \tag{1.40}$$

thanks to (1.37). Thus, energy estimates are available once we control the term $\int \varrho f v\,dx$, which is in some sense a lower order term. We will not complete this step.

Rather we wish to emphasize that the cancellation property (1.40) cannot be expected for the model for mixtures (1.13)–(1.14). If we imitate the above outlined procedure and multiply the ith equation in (1.14), where $P_i(\rho)$ is of the form (1.16), by $u^{(i)}$ and sum the resulting equations over $i = 1, 2$, we obtain again with the aid of $(1.9)_1$ the control of the L^2-norm of the velocity gradient on the left-hand side; however, the terms containing the partial pressures on the right-hand side taking the form

$$\sum_{i=1}^{2} c_i \int_{\mathbb{R}^3} (\rho_1+\rho_2)^{\gamma-1}\rho_i \operatorname{div} u^{(i)}\,dx \tag{1.41}$$

are not vanishing, as one can easily check. This means that we have to proceed in a different way.

Our approach is based on the equation for the effective viscous flux (1.22), known from the theory for standard compressible fluid models (see [19, 23, 7, 8]). We multiply this equation gradually by $\rho_i - \rho_{i\infty}$ or $\rho_i^\gamma - \rho_{i\infty}^\gamma$, integrate, and use also the properties of $P_i(\rho)$ and (1.1). Consequently, we succeed in proving estimates on the (approximate) densities.

In this procedure, there arise boundary terms which we are not able to treat up to now for the standard Dirichlet (no-slip) or Neumann-type (no-stick) boundary conditions. Thus, we deal with the problem in $\mathbb{R}^3$, where, however, some additional technical difficulties occur.

Finally, we have to handle the difficulties due to the nonlinearity of the pressure term $P(\rho)$. For this purpose we apply the techniques which were developed to treat the compressible Navier–Stokes equations (see [19, 7]).

1.5. Related results. To our knowledge such systems as (1.13)–(1.14) have not been considered so far for space dimension $d \geq 2$. (The one-dimensional case has been treated, for example, in [17] and [27].) This is in contrast to the successful mathematical theory of "immiscible" mixtures of fluids (cf. [20, 21, 4, 16]), the theory of the density-dependent incompressible Navier–Stokes equations ([18, 1]), and the

compressible Navier–Stokes equations in the steady case (cf. [22, 23, 19]) and in the evolutionary case (cf. [19, 7, 9, 5, 3]).

The result of Theorem 1.1 was announced in [11], where we sketch the limit from σ-approximations to the original problem (cf. section 5 in this paper). In [11], we also show the additional regularity property of the solutions $\rho \in L^q_{loc}$, $\nabla u \in L^q_{loc}$ for all $q < \infty$.

Let us remark that the methods of the present paper can be adapted to treat the stationary case on a bounded domain with Dirichlet boundary conditions in the presence of the convective term, provided that an H^1-estimate for u is available. This is, e.g., possible if $P_i(\rho) \sim |\rho|^{\frac{1}{2}}$, which holds for some examples of mixtures. The quasi-stationary case is studied in [15], where even first derivatives of ρ_i can be estimated. Finally, in [13], the authors show that in the absence of $f^{(i)}$ and interaction terms the only solution to (1.13)–(1.14) is $u^{(i)} = 0$ and $\rho_i = \rho_{i\infty}$. This is a uniqueness type result.

2. Solvability of the $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\sigma})$-approximations. The goal of this section is to prove the following statement.

PROPOSITION 2.1. *For each $\alpha > 0$, $\beta > 0$, and $\sigma > 0$, under the assumptions of Theorem* 1.1*, there exists a weak solution $(\rho, u) = (\rho^{\beta,\alpha,\sigma}, u^{\beta,\alpha,\sigma})$, $\rho = (\rho_1, \rho_2)^T$, $u = (u^{(1),T}, u^{(2),T})^T$ solving* (1.35) *and* (1.36) *such that for $i = 1, 2$*

$$\rho_i \geq 0\,, \quad \rho_i - \rho_{i\infty} \in L^2(\mathbb{R}^3) \cap L^{s_0}(\mathbb{R}^3) \cap H^1(\mathbb{R}^3) \qquad \text{and} \qquad u^{(i)} \in H^1_0(\mathbb{R}^3; \mathbb{R}^3).$$

Proof. We consider the system (1.35)–(1.36) in the ball $B_R = \{x \in \mathbb{R}^3; |x| \leq R\}$, where we impose the following boundary conditions:

$$u^{(i)} = 0 \quad \text{and} \quad \nabla \rho_i \cdot n = 0 \text{ on } \partial B_R \qquad (n \text{ being the outer normal to } \partial B_R). \tag{2.1}$$

We focus on only two steps. First, we find energy estimates to (1.35)–(1.36) in B_R that are uniform with respect to R. Note that these estimates are then not only sufficient for the existence of a solution for fixed R, but they also suffice to pass to the limit as $R \to \infty$. We skip these standard arguments based on the compact embeddings. In the second step, we outline that ρ_i are nonnegative.

Step 1. *Energy inequalities.* First, we test (1.36) by $u^{(i)}$ and sum over $i = 1, 2$ (in the next computation, in contrast to the rest of the paper, we use the summation convention over repeated indices i). It leads to (for $q > 2$)

$$\begin{aligned}
c_0 \int_{B_R} |\nabla u|^2 \, dx &\leq \int_{B_R} L^{(i)} u \cdot u^{(i)} \, dx \\
&= \int_{B_R} w_\beta (P_i(\rho) - P_i(\rho_\infty)) \operatorname{div} u^{(i)} \, dx + \int_{B_R} (\rho_i - \rho_{i\infty}) f^{(i)} \cdot u^{(i)} \, dx \\
&\quad + \rho_{i\infty} \int_{B_R} f^{(i)} \cdot u^{(i)} \, dx + (-1)^{i+1} \int_{B_R} a(\rho, |u^{(1)} - u^{(2)}|)(u^{(2)} - u^{(1)}) \cdot u^{(i)} \, dx \\
&\leq \varepsilon \int_{B_R} |\nabla u|^2 \, dx + \delta \int_{B_R} |P(\rho) - P(\rho_\infty)|^q \, dx + K_\delta \int_{B_R} w_\beta^{\frac{2q}{q-2}} \, dx \\
&\quad + 2\varepsilon \int_{B_R} |u|^6 \, dx + \varepsilon \int_{B_R} |\rho - \rho_\infty|^2 \, dx + K_\varepsilon \int_{B_R} |f|^3 \, dx \\
&\quad + K_\varepsilon \int_{B_R} |f|^{6/5} \, dx + \int_{B_{R_0}} (1 + |u^{(1)} - u^{(2)}|)^\theta |u| \, dx\,,
\end{aligned} \tag{2.2}$$

where we used (1.26) and (1.27). Since $0 < \theta < 1$, the last term is bounded (using also Young's inequality) by

$$C \int_{B_{R_0}} \left(|u| + |u|^{1+\theta} \right) dx \le \varepsilon \int_{B_R} |\nabla u|^2 \, dx + K(R_0) \qquad \text{for all } R \ge R_0. \tag{2.3}$$

Note that if θ were greater than or equal to 1, the last estimate would not be available. The sublinear growth of $J^{(1)}$ seems to be important.

Next, since $f \in L^\infty(\mathbb{R}^3)$ and has compact support and $w_\beta^{\frac{2q}{q-2}}$ is integrable over $\mathbb{R}^3$ for any $\beta > 0$, it follows directly from (2.2) and (2.3) and then from (1.28) that

$$\begin{aligned} \int_{B_R} |\nabla u|^2 \, dx &\le \tilde{\varepsilon} \int_{B_R} |P(\rho) - P(\rho_\infty)|^q \, dx + \tilde{\varepsilon} \int_{B_R} |\rho - \rho_\infty|^2 \, dx + K \\ &\le \tilde{\varepsilon} K(\rho_\infty) \int_{B_R} |\rho - \rho_\infty|^{q\gamma} + |\rho - \rho_\infty|^q \, dx + \tilde{\varepsilon} \int_{B_R} |\rho - \rho_\infty|^2 \, dx + K, \end{aligned} \tag{2.4}$$

where $\tilde{\varepsilon}$ is chosen as small as needed and $K = K(\beta^{-1}, \delta^{-1}, R_0)$.

Next, we test (1.35) by $\rho_i - \rho_{i\infty}$. Using inequality (2.4), we obtain

$$\begin{aligned} &\sigma \int_{B_R} |\nabla \rho_i|^2 \, dx + \alpha \int_{B_R} |\rho_i - \rho_{i\infty}|^2 + |\rho_i - \rho_{i\infty}|^{s_0} \, dx = \int_{B_R} \rho_i u^{(i)} \nabla (\rho_i - \rho_{i\infty}) \, dx \\ &= -\frac{1}{2} \int_{B_R} \operatorname{div} u^{(i)} (\rho_i - \rho_{i\infty})^2 \, dx - \rho_{i\infty} \int_{B_R} \operatorname{div} u^{(i)} (\rho_i - \rho_{i\infty}) \, dx \\ &\le \frac{\alpha}{8} \int_{B_R} |\rho_i - \rho_{i\infty}|^4 \, dx + K_\alpha \int_{B_R} |\nabla u|^2 \, dx + \frac{\alpha}{8} \int_{B_R} |\rho_i - \rho_{i\infty}|^2 \, dx \\ &\le \frac{\alpha}{8} \int_{B_R} |\rho_i - \rho_{i\infty}|^4 \, dx + \frac{\alpha}{8} \int_{B_R} |\rho_i - \rho_{i\infty}|^2 \, dx + \tilde{\varepsilon} K(\rho_\infty) \int_{B_R} |\rho - \rho_\infty|^{q\gamma} \, dx \, . \end{aligned}$$

Observing that for $s_0 > 2\gamma$ there is always a $q > 2$ such that $s_0 > q\gamma$, we can (via interpolation) subtract all the terms containing ρ to finally obtain

$$\sigma \int_{B_R} |\nabla \rho_i|^2 \, dx + \alpha \int_{B_R} |\rho_i - \rho_{i\infty}|^2 \, dx + \alpha \int_{B_R} |\rho_i - \rho_{i\infty}|^{s_0} \, dx \le K \, . \tag{2.5}$$

From (2.4) it then follows that

$$\int_{B_R} |\nabla u|^2 \, dx \le K \qquad \text{with } K = K(\alpha^{-1}, \beta^{-1}, R_0) \text{ but independent of } R \, . \tag{2.6}$$

Step 2. *Nonnegativity of the densities.* The proof is performed via standard weak maximum principles. To be more specific, we multiply (1.35) (for fixed $i = 1$ or 2) by a function G solving

$$\begin{aligned} -\sigma \Delta G - u_k^{(i)} \frac{\partial G}{\partial x_k} + \alpha G + \alpha |\rho_i - \rho_{i\infty}|^{s_0 - 2} G = \chi_{\{\rho_i < 0\}} \text{ in } B_R \, , \\ \nabla G \cdot n = 0 \quad \text{ on } \partial B_R \, . \end{aligned}$$

Before doing so, one may prefer to mollify the coefficients $u_k^{(i)}$ and $|\rho_i - \rho_{i\infty}|^{s_0 - 2}$ to construct G smooth enough. Integrating the resulting equation over B_R and using

integration by parts together with the boundary conditions for ρ_i, G and $u^{(i)}$ lead finally to the conclusion that

$$\int_{B_R} \rho_i \chi_{\{\rho_i<0\}}\, dx \geq 0\,,$$

which implies $\rho_i^- = 0$ a.e. in B_R. Thus $\rho_i \geq 0$ a.e. in B_R.

The sketch of the proof of Proposition 2.1 is finished. □

3. Solvability of the (α, σ)-approximations.

3.1. The equation for the effective viscous flux. Let $(\rho, u) = (\rho^{\beta,\alpha,\sigma}, u^{\beta,\alpha,\sigma})$ be solutions of (1.35)–(1.36) constructed above. The aim is to show that (ρ, u) also satisfy (a.e. in $\mathbb{R}^3$) the equation for the effective viscous flux

$$A_0\left(w_\beta\left(P(\rho)-P(\rho_\infty)\right)\right) - \begin{pmatrix} \beta_0 \operatorname{div} u^{(1)} \\ \operatorname{div} u^{(2)} \end{pmatrix} = A_0 \operatorname{div} \Delta^{-1} \begin{pmatrix} \rho_1 f^{(1)} + J^{(1)} \\ \rho_2 f^{(2)} - J^{(1)} \end{pmatrix}. \tag{3.1}$$

Thus, we multiply (1.36) by $\nabla\varphi^{(i)}$, where $\varphi^{(i)}$ is a smooth function with compact support in $\mathbb{R}^3$. In matrix notation, we obtain

$$\begin{aligned} &\int \begin{pmatrix} \Delta\varphi^{(1)} \\ \Delta\varphi^{(2)} \end{pmatrix}^T \begin{pmatrix} 2\mu_{11}+\nu_{11} & 2\mu_{12}+\nu_{12} \\ 2\mu_{21}+\nu_{21} & 2\mu_{22}+\nu_{22} \end{pmatrix} \begin{pmatrix} \operatorname{div} u^{(1)} \\ \operatorname{div} u^{(2)} \end{pmatrix} dx \\ &\quad = \int \begin{pmatrix} \Delta\varphi^{(1)} \\ \Delta\varphi^{(2)} \end{pmatrix}^T \begin{pmatrix} w_\beta\left(P_1(\rho)-P_1(\rho_\infty)\right) - \operatorname{div}\Delta^{-1}\left(\rho_1 f^{(1)} + J^{(1)}\right) \\ w_\beta\left(P_2(\rho)-P_2(\rho_\infty)\right) - \operatorname{div}\Delta^{-1}\left(\rho_2 f^{(2)} - J^{(1)}\right) \end{pmatrix} dx\,, \end{aligned}$$

where we used the identity

$$\int (\rho_i f^{(i)} + J^{(i)})\nabla\varphi^{(i)}\, dx = -\int \operatorname{div}\Delta^{-1}(\rho_i f^{(i)} + J^{(i)})\Delta\varphi^{(i)}\, dx, \tag{3.2}$$

which is valid certainly for smooth functions $\varphi^{(i)}$ with compact support, and by density arguments also for functions such that $\Delta\varphi^{(i)} \in L^2$. To see this, we first denote $h_i = \Delta^{-1}(\rho_i f^{(i)} + J^{(i)})$ so that $\Delta h_i = \rho_i f^{(i)} + (-1)^{i+1} a(\cdot, \rho, |u^{(1)} - u^{(2)}|)(u^{(2)} - u^{(1)})$ in $\mathbb{R}^3$. Then, using (1.26)–(1.27),

$$\begin{aligned} \left|\int \operatorname{div} h_i \Delta\varphi^{(i)}\, dx\right| &\leq \|\Delta\varphi^{(i)}\|_{L^2} \|\operatorname{div} h_i\|_{L^2} \leq c\|\Delta\varphi^{(i)}\|_{L^2}\|\rho_i f^{(i)}\|_{(H_0^1)^*} \\ &\leq c\|\Delta\varphi^{(i)}\|_{L^2}\Big(\|(\rho_i-\rho_{i\infty})f^{(i)}\|_{L^{6/5}} + \rho_{i\infty}\|f^{(i)}\|_{L^{6/5}} \\ &\qquad + \|1+|u^{(1)}-u^{(2)}|\,\|^{2\theta}_{L^{\frac{12\theta}{5}}(B_{R_0})}\Big) \\ &\leq c\|\Delta\varphi^{(i)}\|_{L^2}\left(\|\rho_i-\rho_{i\infty}\|_{L^{s_0}}\|f^{(i)}\|_{L^{\frac{6s_0}{5s_0-6}}} + \|\nabla u\|^{2\theta}_{L^2} + K_f\right) \\ &\leq K\|\Delta\varphi^{(i)}\|_{L^2}\,, \end{aligned} \tag{3.3}$$

as follows from the facts that $\rho_i - \rho_{i\infty} \in L^{s_0}$, $u \in H_0^1$, and $f^{(i)} \in L^\infty$ and a have compact supports. Thus, the validity of (3.2) for $\varphi^{(i)}$ with $\Delta\varphi^{(i)} \in L^2$ is verified.

Choose $\varphi = A_0\tilde{\psi}$, where $\varphi = \left(\varphi^{(1),T}, \varphi^{(2),T}\right)^T$, $\tilde{\psi} = (\tilde{\psi}^{(1),T}, \tilde{\psi}^{(2),T})^T$, and A_0 is introduced in (1.21). Then we obtain

$$\int \begin{pmatrix} \Delta\tilde{\psi}^{(1)} \\ \Delta\tilde{\psi}^{(2)} \end{pmatrix}^T \begin{pmatrix} \beta_0 \operatorname{div} u^{(1)} \\ \operatorname{div} u^{(2)} \end{pmatrix} dx$$
$$= \int \begin{pmatrix} \Delta\tilde{\psi}^{(1)} \\ \Delta\tilde{\psi}^{(2)} \end{pmatrix}^T A_0 \begin{pmatrix} w_\beta \left(P_1(\rho) - P_1(\rho_\infty)\right) - \operatorname{div} \Delta^{-1}\left(\rho_1 f^{(1)} + J^{(1)}\right) \\ w_\beta \left(P_2(\rho) - P_2(\rho_\infty)\right) - \operatorname{div} \Delta^{-1}\left(\rho_2 f^{(2)} - J^{(1)}\right) \end{pmatrix} dx\,.$$

For any $z \in L^2(\mathbb{R}^3)$ solve $-\Delta\tilde{\psi}^{(i)} = z$ in L^2, $i = 1, 2$. Then (3.1) follows.

3.2. Limit $\boldsymbol{\beta \to 0}$. We prove the following statement.

PROPOSITION 3.1. *Let $\alpha > 0$ and $\sigma > 0$ be fixed. Let all assumptions of Theorem* 1.1 *be fulfilled. Then there exists a solution $(\rho, u) = (\rho^{\alpha,\sigma}, u^{\alpha,\sigma}), \rho = (\rho_1, \rho_2)^T, u = (u^{(1),T}, u^{(2),T})^T$, such that (for $i = 1, 2$ and $\varepsilon > 0$ small)*

$$\rho_i \ge 0,\ \rho_i - \rho_{i\infty} \in L^{1+\varepsilon}(\mathbb{R}^3) \cap L^{s_0}(\mathbb{R}^3)\,, \tag{3.4}$$
$$u^{(i)} \in H^1_0(\mathbb{R}^3; \mathbb{R}^3) \tag{3.5}$$

solving weakly (1.35), (1.36), *and* (3.1) *with $\beta = 0$ ($\Longrightarrow w_\beta = 1$).*

In addition, (ρ, u) satisfies

$$\sum_{i=1}^2 \sigma\hat{\beta}_i \int \frac{|\nabla\rho_i|^2}{\rho_i}\,dx + \sum_{i=1}^2 \frac{\alpha\hat{\beta}_i}{2} \int (\rho_i - \rho_{i\infty})(\log\rho_i - \log\rho_{i\infty})\,dx$$
$$+ \sum_{i=1}^2 \frac{\alpha\hat{\beta}_i}{2} \int |\rho_i - \rho_{i\infty}|^{s_0-2}(\rho_i - \rho_{i\infty})(\log\rho_i - \log\rho_{i\infty})\,dx$$
$$+ \int A_0\left(P(\rho) - P(\rho_\infty)\right)\cdot(\rho - \rho_\infty)\,dx \le \int A_0 F\cdot(\rho - \rho_\infty) =: Y_F\,, \tag{3.6}$$

where F is introduced in (1.19), *$\hat{\beta}_1 = \beta_0$ (cf.* (1.21)*), and $\hat{\beta}_2 = 1$.*

Proof. We split the proof into two steps. First, we derive (3.6) for $\left(\rho^{\beta,\alpha,\sigma}, u^{\beta,\alpha,\sigma}\right)$. This means that w_β occurs in the last term of the left-hand side of (3.6). Let us call this inequality $(3.6)_\beta$. To be able to pass to the limit as $\beta \to 0$ we need to find the estimates that are uniform with respect to β. This is done in the second step.

Step 1. *Derivation of* $(3.6)_\beta$. We take the L^2 scalar product of (3.1) with $\tau(\rho - \rho_\infty)$, where τ is the usual cut-off function with $\tau = 1$ on the ball B_R and $\tau = 0$ outside of B_{2R}. We have verified in (3.3) that $F^{(i)}$ is bounded in L^2, which implies that $A_0 F\cdot(\rho - \rho_\infty) \in L^1$. Thus, it is easy to pass to the limit as $R \to \infty$ in all terms except for the terms with $\operatorname{div} u^{(i)}$ that require a detailed investigation. Clearly,

$$-\left(\operatorname{div} u^{(i)}, \tau(\rho_i - \rho_{i\infty})\right) = \left(u^{(i)}, \nabla\tau(\rho_i - \rho_{i\infty})\right) + \left(u^{(i)}, \tau\nabla\rho_i\right) =: H_1 + H_2\,, \tag{3.7}$$

where

$$H_2 = \left((\rho_i + \delta)u^{(i)}\tau, \nabla\log(\rho_i + \delta)\right)$$
$$= -\left(\operatorname{div}\left(\rho_i u^{(i)}\right)\tau, \log(\rho_i + \delta) - \log(\rho_{i\infty} + \delta)\right)$$
$$-\delta\left(\operatorname{div} u^{(i)}\tau, \log(\rho_i + \delta) - \log(\rho_{i\infty} + \delta)\right)$$
$$-\left((\rho_i + \delta)u^{(i)}\nabla\tau, \log(\rho_i + \delta) - \log(\rho_{i\infty} + \delta)\right) =: H_3 + H_4 + H_5\,.$$

First we observe that

$$|H_1| \le \|u^{(i)}\|_{L^6(B_{2R}\setminus B_R)} \|\nabla\tau\|_{L^3} \|\rho_i - \rho_{i\infty}\|_{L^2(B_{2R}\setminus B_R)} \to 0 \text{ as } R\to\infty \tag{3.8}$$

due to the fact that $u \in L^6(\mathbb{R}^3;\mathbb{R}^3)$, $\rho_i - \rho_{i\infty} \in L^2(\mathbb{R}^3)$, and $\|\nabla\tau\|_{L^3} \le C$. Also, since $|\log(\rho_i+\delta) - \log(\rho_{i\infty}+\delta)| \le \delta^{-1}(\rho_i - \rho_{i\infty})$, similar arguments ($\rho_i - \rho_{i\infty} \in L^4(\mathbb{R}^3)$) imply that

$$|H_5| \le \delta^{-1} \int (\rho_i - \rho_{i\infty})^2 |u^{(i)}||\nabla\tau|\, dx + (\rho_{i\infty}+\delta)\delta^{-1}|H_1| \to 0 \text{ as } R\to\infty\,. \tag{3.9}$$

Since $H_4 = -\int \operatorname{div} u^{(i)} \tau(\delta \int_0^1 \frac{1}{\delta + t\rho_i + (1-t)\rho_{i\infty}}\, dt)(\rho_i - \rho_{i\infty})\, dx$ and $\operatorname{div} u^{(i)} \cdot (\rho_i - \rho_{i\infty})$ belongs to L^1, and also

$$\delta \int_0^1 \frac{1}{\delta + t\rho_i + (1-t)\rho_{i\infty}}\, dt \le \delta \int_0^1 \frac{1}{\delta + (1-t)\rho_{i\infty}}\, dt = \frac{2\delta\log\delta}{\rho_{i\infty}} \to 0 \text{ as } \delta\to 0\,,$$

we observe that

$$H_4 \text{ is uniformly bounded with respect to } R \quad \text{and } H_4 \to 0 \text{ as } \delta\to 0\,. \tag{3.10}$$

Finally, inserting (1.35) into H_3 we obtain

$$\begin{aligned} H_3 &= \sigma \int \nabla\rho_i \cdot \nabla\tau \left(\log(\rho_i+\delta) - \log(\rho_{i\infty}+\delta)\right)\, dx \\ &\quad + \sigma \int \frac{|\nabla\rho_i|^2}{\rho_i+\delta}\tau\, dx + \alpha \int (\rho_i - \rho_{i\infty})\tau \left(\log(\rho_i+\delta) - \log(\rho_{i\infty}+\delta)\right)\, dx \\ &\quad + \alpha \int |\rho_i - \rho_{i\infty}|^{s_0-2}(\rho_i - \rho_{i\infty})\tau \left(\log(\rho_i+\delta) - \log(\rho_{i\infty}+\delta)\right)\, dx\,. \end{aligned} \tag{3.11}$$

As $R\to\infty$, the first term on the right-hand side of (3.11) vanishes, arguing as in (3.8) or (3.9). In the remaining terms the cut-off function τ tends to 1. Thus, letting first $R\to\infty$ and then $\delta\to 0$ we obtain inequality $(3.6)_\beta$, using (3.7)–(3.11) and Fatou's lemma.

Step 2. *Estimates uniform with respect to* β. Using the same arguments as in (3.3) to estimate Y_F we conclude that

$$|Y_F| \le \frac{\alpha\hat\beta}{4}\|\rho - \rho_\infty\|_{L^2}^2 + \frac{\alpha\hat\beta}{4}\|\rho - \rho_\infty\|_{L^{s_0}}^{s_0} + K_{R_0}\|\nabla u\|_{L^2}^{2\theta} + K_f\,. \tag{3.12}$$

Thus, the terms including the norms of densities can be absorbed into the left-hand side of $(3.6)_\beta$.

Next, multiplying (1.36) by $u^{(i)}$ (and summing over $i=1,2$) gives similarly[4] as in (2.2) with the help of inequality (1.28) and Hölder's inequality

$$c_0\|\nabla u\|_{L^2}^2 \le K\left(\|\rho_i - \rho_{i\infty}\|_{L^{s_0}}^{s_0} + \|\rho_i - \rho_{i\infty}\|_{L^2}^2\right) + K\,, \tag{3.13}$$

which implies

$$\begin{aligned} \|\nabla u\|_{L^2}^{2\theta} &\le K\left(\|\rho_i - \rho_{i\infty}\|_{L^{s_0}}^{s_0} + \|\rho_i - \rho_{i\infty}\|_{L^2}^2\right)^\theta + K \\ &\le \varepsilon\left(\|\rho_i - \rho_{i\infty}\|_{L^{s_0}}^{s_0} + \|\rho_i - \rho_{i\infty}\|_{L^2}^2\right) + K\,. \end{aligned} \tag{3.14}$$

[4]This time we have to avoid the dependence of the estimates on β. We therefore use the fact that $|w_\beta| \le 1$.

Inserting (3.14) into (3.12) and then into $(3.6)_\beta$, we obtain (first from $(3.6)_\beta$ and then from (3.13))

$$\sum_{i=1}^{2}\left(\int\frac{|\nabla\rho_i|^2}{\rho_i}\,dx+\|\rho-\rho_\infty\|_{L^2}^2+\|\rho-\rho_\infty\|_{L^{s_0}}^{s_0}\right)+\|\nabla u\|_{L^2}^2\le K=K(\alpha^{-1},\sigma^{-1}),$$

where K is independent of β. Since $s_0>4$, this estimate implies that

$$\|\nabla\rho\|_{L^{8/5}}\le K.$$

Letting $\beta\to 0$, it is straightforward to find a subsequence of $(\rho^{\beta,\alpha,\sigma},u^{\beta,\alpha,\sigma})$ weakly converging to $(\rho^{\alpha,\sigma},u^{\alpha,\sigma})$ in $W^{1,8/5}\times H_0^1$, and converging strongly in $(L^{s_0-\varepsilon}\cap L^2)\times L^6$ locally and a.e. in $\mathbb{R}^3$. This is certainly enough to show that $(\rho^{\alpha,\sigma},u^{\alpha,\sigma})$ is a weak solution to (1.35) and (1.36) with $w_\beta=w_0=1$.

What remains is to consider the limit $\beta\to 0$ in $(3.6)_\beta$. Here, we incorporate (in addition to the above mentioned a.e. convergence) lower semicontinuity arguments and Fatou's and Vitali's lemmas in order to obtain (3.6). □

4. Solvability of the σ-approximations. In this section, we prove three assertions. In Proposition 4.1, we find estimates for $(\rho,u)=(\rho^{\alpha,\sigma},u^{\alpha,\sigma})$ that are uniform in *both* α and σ. Then, in Proposition 4.2, we derive estimates involving the gradient of the density. In these estimates, which are uniform with respect to α only, the parameter ε appearing therein is such that $1+\varepsilon\le\gamma$. If we were allowed to take $\varepsilon=1$ in these estimates (which means that γ has to be greater than 2), we could simplify the approximating procedure by setting $\sigma=\alpha$ (however, we want to avoid any bound on γ).

Based on the estimates from Propositions 4.1 and 4.2 we are able to show, in Proposition 4.3, that a suitable weak limit (ρ^σ,u^σ) of $\{(\rho^{\alpha,\sigma},u^{\alpha,\sigma})\}$ is a solution of the σ-approximations. Even more, we show that (ρ^σ,u^σ) fulfills a corresponding equation for the effective viscous flux, and due to weak lower semicontinuity of the norms, the estimates presented in Proposition 4.1 are also valid for (ρ^σ,u^σ) uniformly with respect to σ. These estimates then play a starting role for further consideration regarding the compactness of ρ^σ, which is presented in section 5.

PROPOSITION 4.1. *Let $(\rho^{\alpha,\sigma},u^{\alpha,\sigma})$ be weak solutions to* (1.35)–(1.36) *with $\beta=0$. Then there exists a constant $K>0$ independent of α and σ such that*

$$\|\rho^{\alpha,\sigma}-\rho_\infty\|_{L^2}^2+\|\rho^{\alpha,\sigma}-\rho_\infty\|_{L^{2\gamma}}^{2\gamma}+\|\nabla u^{\alpha,\sigma}\|_{L^2}^2\le K<\infty\,. \tag{4.1}$$

Proof. Step 1. *Derivation of the inequality*

$$\|\rho^{\alpha,\sigma}-\rho_\infty\|_{L^2}^2+\|\rho^{\alpha,\sigma}-\rho_\infty\|_{L^{\gamma+1}}^{\gamma+1}\le K+K\|\nabla u^{\alpha,\sigma}\|_{L^2}^{2\theta}\,. \tag{4.2}$$

Since $(\rho,u)=(\rho^{\alpha,\sigma},u^{\alpha,\sigma})$ fulfills (3.6), neglecting the first three nonnegative terms and using inequality (1.29), we obtain

$$\hat\lambda_0\left(\|\rho-\rho_\infty\|_{L^{\gamma+1}}^{\gamma+1}+\|\rho-\rho_\infty\|_{L^2}^2\right)\le|Y_F|\,.$$

Then $|Y_F|$ is estimated as in (3.12) replacing $\alpha\hat\beta$ by $\hat\lambda_0$ and the L^{s_0}-norm by the $L^{\gamma+1}$-norm. Doing so, we obtain (4.2) (K depends neither on α nor on σ).

Step 2. *Derivation of the inequality*

$$\|\rho^{\alpha,\sigma}-\rho_\infty\|_{L^{2\gamma}}^{2\gamma}\le K+K\|\nabla u^{\alpha,\sigma}\|_{L^2}^{2\theta}\,. \tag{4.3}$$

According to Proposition 3.1, the equation for the effective viscous flux (3.1) with $w_\beta = 1$ holds for $(\rho, u) = (\rho^{\alpha,\sigma}, u^{\alpha,\sigma})$. Multiplying the ith equation by $\tau^2(\rho_i^\gamma - \rho_{i\infty}^\gamma)$, where τ is the standard cut-off function ($\tau = 1$ on B_R and $\tau = 0$ outside of B_{2R}, $|\nabla\tau| \le CR^{-1}$), and integrating it, we end up with[5]

$$\begin{aligned}\left((A_0(P(\rho) - P(\rho_\infty)))_i, (\rho_i^\gamma - \rho_{i\infty}^\gamma)\tau^2\right) &= \hat\beta_i \left(\operatorname{div} u^{(i)}, \tau^2(\rho_i^\gamma - \rho_{i\infty}^\gamma)\right) - Z_F,\\ \text{where } Z_F &:= \left(a_0^{(i)} F^{(i)}, (\rho_i^\gamma - \rho_{i\infty}^\gamma)\tau^2\right).\end{aligned} \tag{4.4}$$

In Step 4 below, we will be able to show that (for α, σ fixed)

$$\limsup_{R\to\infty} \left(\operatorname{div} u^{(i)}, \tau^2(\rho_i^\gamma - \rho_{i\infty}^\gamma)\right) \le 0\,. \tag{4.5}$$

Assuming that (4.5) is valid, it follows from (4.4)–(4.5) and inequality (1.24) that

$$\|\rho - \rho_\infty\|_{L^{2\gamma}}^{2\gamma} \le K\|\rho - \rho_\infty\|_{L^2}^2 + \limsup_{R\to\infty} |Z_F|\,. \tag{4.6}$$

Since, by the mean value theorem ($t \in (0,1)$),

$$|\rho_i^\gamma - \rho_{i\infty}^\gamma| = \gamma|(\rho_{i\infty} + t(\rho_i - \rho_{i\infty}))^{\gamma-1}(\rho_i - \rho_{i\infty})| \le K\left(|\rho_i - \rho_{i\infty}| + |\rho_i - \rho_{i\infty}|^\gamma\right), \tag{4.7}$$

we conclude from (4.6) and (4.7) that

$$\|\rho_i - \rho_{i\infty}\|_{L^{2\gamma}}^{2\gamma} + \|\rho_i^\gamma - \rho_{i\infty}^\gamma\|_{L^2}^2 \le K\left(\|\rho_i - \rho_{i\infty}\|_{L^2}^2 + \limsup_{R\to\infty} |Z_F|\right). \tag{4.8}$$

Next, we estimate $|Z_F|$ using similar arguments as in (3.3) and obtain

$$|Z_F| \le \frac{1}{2}\|\rho_i^\gamma - \rho_{i\infty}^\gamma\|_{L^2}^2 + K\|\nabla u^{(i)}\|_{L^2}^{2\theta} + K\|\rho_i - \rho_{i\infty}\|_{L^2}^2 + K_f\,. \tag{4.9}$$

Taking (4.8)–(4.9) into account together with (4.2) leads to (4.3).

Step 3. *Derivation of* (4.1). We multiply (1.36) (with $\beta = 0$) by $u^{(i)}$. Using standard manipulations we obtain (cf. (3.13))

$$c_0\|\nabla u^{\alpha,\sigma}\|_{L^2}^2 \le K\sum_{i=1}^{2}\left(\|\rho_i^{\alpha,\sigma} - \rho_{i\infty}\|_{L^{2\gamma}}^{2\gamma} + \|\rho_i^{\alpha,\sigma} - \rho_{i\infty}\|_{L^2}^2\right). \tag{4.10}$$

The required energy estimates are thus obtained from (4.2), (4.3), and (4.10).

Step 4. *A proof of* (4.5). We start with observing that

$$\begin{aligned}\left(\operatorname{div} u^{(i)}, \tau^2(\rho_i^\gamma - \rho_{i\infty}^\gamma)\right) &= -\left(u^{(i)}, \nabla\tau^2(\rho_i^\gamma - \rho_{i\infty}^\gamma)\right) - \gamma\left(u^{(i)}, \tau^2\rho_i^{\gamma-1}\nabla\rho_i\right)\\ &= -\left(u^{(i)}, \nabla\tau^2(\rho_i^\gamma - \rho_{i\infty}^\gamma)\right) - \gamma\left(u^{(i)}, \tau^2\nabla\rho_i(\rho_i + \delta)^{\gamma-1}\right)\\ &\quad + \gamma\left(u^{(i)}, \tau^2\nabla\rho_i\Big((\rho_i + \delta)^{\gamma-1} - \rho_i^{\gamma-1}\Big)\right) =: F_1 + F_2 + F_3\,.\end{aligned} \tag{4.11}$$

[5]Recall that $\hat\beta_1 = \beta_0$, $\hat\beta_2 = 1$, and $a_0^{(i)}$ is the ith row of A_0.

Recalling that (cf. (3.4)) $\rho_i - \rho_{i\infty} \in L^{1+\varepsilon} \cap L^{s_0}$ for $\varepsilon > 0$ small and $s_0 > 2\gamma$, we conclude, with the help of (4.7), that (for α, $\sigma > 0$ fixed)

$$|F_1| \le \|\nabla\tau\|_{L^3}\|u^{(i)}\|_{L^6(B_{2R}\setminus B_R)}\|\rho_i^\gamma - \rho_{i\infty}^\gamma\|_{L^2(B_{2R}\setminus B_R)} \to 0 \quad \text{as } R\to\infty\,. \tag{4.12}$$

Also, it is not difficult to see that $F_3 \to 0$ as $\delta \to 0$.

Next, defining $T(\rho_i) := (\rho_i+\delta)^{\gamma-1} - (\rho_{i\infty}+\delta)^{\gamma-1}$, we have

$$\begin{aligned} F_2 &= -\frac{\gamma}{\gamma-1}\left((\rho_i+\delta)u^{(i)}, \tau^2\nabla\left((\rho_i+\delta)^{\gamma-1} - (\rho_{i\infty}+\delta)^{\gamma-1}\right)\right) \\ &= \frac{\gamma}{\gamma-1}\left(\operatorname{div}\left((\rho_i+\delta)u^{(i)}\right), \tau^2 T(\rho_i)\right) + \frac{\gamma}{\gamma-1}\left((\rho_i+\delta)u^{(i)}, \nabla\tau^2 T(\rho_i)\right) := F_4 + F_5\,. \end{aligned}$$

Inserting (1.35) into F_4, we obtain

$$F_4 = \frac{\gamma}{\gamma-1}\big(\sigma\Delta\rho_i, \tau^2 T(\rho_i)\big) - \alpha\big(|\rho_i-\rho_{i\infty}|^{s_0-2}(\rho_i-\rho_{i\infty}) + (\rho_i-\rho_{i\infty}), \tau^2 T(\rho_i)\big)\,.$$

Since the last term is nonpositive (it is a monotone operator), we can neglect it. We integrate the first term by parts. Dropping the term with $-\sigma\int|\nabla\rho_i|^2(\rho_i+\delta)^{\gamma-2}\,dx$, we obtain

$$\begin{aligned} F_4 &\le -\frac{\sigma\gamma}{\gamma-1}\left(\nabla\tau^2\nabla\rho_i, (\rho_i+\delta)^{\gamma-1} - (\rho_{i\infty}+\delta)^{\gamma-1}\right) \\ &= \gamma\sigma\left((\rho_i+\delta)^\gamma - (\rho_{i\infty}+\delta)^\gamma, \Delta\tau^2\right) \end{aligned}$$

and the last term vanishes as $R\to\infty$ since $\|\Delta\tau\|_{L^2}\to 0$ and $(\rho_i+\delta)^\gamma - (\rho_{i\infty}+\delta)^\gamma$ is an L^2-integrable function (α, σ fixed).

Concerning F_5, we first write $\rho_i+\delta = (\rho_i-\rho_{i\infty}) + (\rho_{i\infty}+\delta)$ and consider two cases: $\gamma \ge 2$ and $\gamma \in (1,2)$. If $\gamma \ge 2$, then (4.7) with $\gamma-1$ instead of γ implies

$$\begin{aligned} |F_5| &\le K\gamma\int_{B_{2R}\setminus B_R} |u^{(i)}|\,|\nabla\tau|\Big(|\rho-\rho_\infty|^2 + |\rho-\rho_\infty|^\gamma + |\rho-\rho_\infty|^{\gamma-1} + |\rho_i-\rho_{i\infty}|\Big)\,dx \\ &\le K\|\nabla\tau\|_{L^3}\|\nabla u\|_{L^2(B_{2R})\setminus B_R)}\Big(\|\rho_i-\rho_{i\infty}\|_{L^4} + \|\rho_i-\rho_{i\infty}\|_{L^{2\gamma}} \\ &\qquad + \|\rho_i-\rho_{i\infty}\|_{L^{2(\gamma-1)}} + \|\rho_i-\rho_{i\infty}\|_{L^2}\Big) \to 0 \text{ as } R\to\infty\,. \end{aligned}$$

If $\gamma\in(1,2)$, then $(\rho+\delta)^{\gamma-1} - (\rho_{i\infty}+\delta)^{\gamma-1} \le (\rho_{i\infty})^{\gamma-2}(\rho_i-\rho_{i\infty})$. Then, however, the same arguments complete the proof of the fact that $|F_5|\to 0$ as $\mathbb{R}\to\infty$. ∎

PROPOSITION 4.2. *Let $(\rho^{\alpha,\sigma}, u^{\alpha,\sigma})$ be weak solutions to* (1.35)–(1.36) *with $w_\beta = 1$. For $\varepsilon > 0$ small enough (i.e., $1+\varepsilon \le \gamma$) there is a K, independent of α and σ, such that*

$$\sigma\int|\nabla\rho_i^{\alpha,\sigma}|^2\rho_i^{\varepsilon-1}\,dx + \alpha\int|\rho_i-\rho_{i\infty}|^{s_0-1+\varepsilon}\,dx + \alpha\int|\rho_i-\rho_{i\infty}|^{1+\varepsilon}\,dx \le K\,. \tag{4.13}$$

Proof. We multiply (1.35) by $\varphi^\varepsilon(\rho_i) = \left((\rho_i+\delta)^\varepsilon - (\rho_{i\infty}+\delta)^\varepsilon\right)\tau^2$, where τ is as above. After integrating over $\mathbb{R}^3$ we obtain

$$\begin{aligned} &\varepsilon\sigma\int|\nabla\rho_i|^2(\rho_i+\delta)^{\varepsilon-1}\tau^2\,dx + \sigma\int\left((\rho_i+\delta)^\varepsilon - (\rho_{i\infty}+\delta)^\varepsilon\right)\nabla\rho_i\cdot\nabla\tau^2\,dx \\ &\quad + \int\operatorname{div}\left(\rho_i u^{(i)}\right)\left((\rho_i+\delta)^\varepsilon - (\rho_{i\infty}+\delta)^\varepsilon\right)\tau^2\,dx \\ &\quad + \alpha\int(\rho_i-\rho_{i\infty})\left((\rho_i+\delta)^\varepsilon - (\rho_{i\infty}+\delta)^\varepsilon\right)\tau^2\,dx \\ &\quad + \alpha\int|\rho_i-\rho_{i\infty}|^{s_0-2}(\rho_i-\rho_{i\infty})\left((\rho_i+\delta)^\varepsilon - (\rho_{i\infty}+\delta)^\varepsilon\right)\tau^2\,dx = 0\,. \end{aligned} \tag{4.14}$$

Note that the first, fourth, and fifth terms in (4.14) are nonnegative; in fact, when letting $R \to \infty$ and $\delta \to 0$, these terms provide information on the left-hand side of (4.13). It then remains to estimate the second and third terms in (4.14)—we denote them S and T. Starting with the former, we observe that

$$\begin{aligned} S &= \sigma \int \left(\frac{1}{1+\varepsilon} \nabla(\rho_i+\delta)^{1+\varepsilon} - (\rho_{i\infty}+\delta)^{\varepsilon} \nabla(\rho_i - \rho_{i\infty}) \right) \nabla\tau^2 \, dx \\ &= -\frac{\sigma}{1+\varepsilon} \int \left((\rho_i+\delta)^{1+\varepsilon} - (\rho_{i\infty}+\delta)^{1+\varepsilon} - (1+\varepsilon)(\rho_{i\infty}+\delta)^{\varepsilon}(\rho_i - \rho_{i\infty}) \right) \Delta\tau^2 \, dx \,. \end{aligned}$$

Then the Taylor expansion for $(\delta + x)^{1+\varepsilon}$ and (4.1) imply

$$|S| \le \frac{\sigma\varepsilon}{2\delta} \int |\rho_i - \rho_{i\infty}|^2 \Delta\tau^2 \, dx \le \frac{\sigma\varepsilon}{2\delta R^2} \|\rho_i - \rho_{i\infty}\|_{L^2}^2 \to 0 \quad \text{as } R \to \infty \,.$$

Next,

$$\begin{aligned} -T &= \int \rho_i u^{(i)} \nabla(\rho_i+\delta)^{\varepsilon} \tau^2 \, dx + \int (\rho_i - \rho_{i\infty}) u^{(i)} \big((\rho_i+\delta)^{\varepsilon} - (\rho_{i\infty}+\delta)^{\varepsilon} \big) \nabla\tau^2 \, dx \\ &\quad + \rho_{i\infty} \int u^{(i)} \big((\rho_i+\delta)^{\varepsilon} - (\rho_{i\infty}+\delta)^{\varepsilon} \big) \nabla\tau^2 \, dx =: T_1 + T_2 + T_3 \,. \end{aligned}$$

Recall that

$$\begin{aligned} (\rho_i+\delta)^{\varepsilon} - (\rho_{i\infty}+\delta)^{\varepsilon} &= \varepsilon \int_0^1 (\delta + \rho_{i\infty} + s(\rho_i - \rho_{i\infty}))^{\varepsilon-1} \, ds (\rho_i - \rho_{i\infty}) \\ &\le \varepsilon \int_0^1 (\delta + s(\rho_i - \rho_{i\infty}))^{\varepsilon-1} \, ds (\rho_i - \rho_{i\infty}) \,. \end{aligned} \tag{4.15}$$

Since the right-hand side of (4.15) is bounded by $2\varepsilon(\delta + |\rho_i - \rho_{i\infty}|)^{\varepsilon}$, we see that

$$|T_2| \le K \Big(|T_3| + \|\rho_i - \rho_{i\infty}\|_{L^{2(1+\varepsilon)}(B_{2R}\setminus B_R)}^{1+\varepsilon} \|u^{(i)}\|_{L^6(B_{2R}\setminus B_R)} \|\nabla\tau\|_{L^3} \Big) \,.$$

The right-hand side of (4.15) is also bounded by $\varepsilon\delta^{-1}(\rho_i - \rho_{i\infty})$. This helps to conclude that

$$|T_3| \le \varepsilon\delta^{-1} \rho_{i\infty} \|u^{(i)}\|_{L^6(B_{2R}\setminus B_R)} \|\rho_i - \rho_{i\infty}\|_{L^2(B_{2R}\setminus B_R)} \|\nabla\tau\|_{L^3} \,.$$

Since the L^3-norm of $\nabla\tau$ is bounded, we see that for α and σ fixed and $R \to \infty$ both T_3 and T_2 vanish. (In fact, there is no restriction on ε as we could still use the fact that $\rho_i - \rho_{i\infty}$ belongs to $L^{s_0} \cap L^2$.)

Finally,

$$T_1 = \int \left(\frac{\nabla\big((\rho_i+\delta)^{1+\varepsilon} - (\rho_{i\infty}+\delta)^{1+\varepsilon} \big)}{1+\varepsilon} - \delta \nabla \big((\rho_i+\delta)^{\varepsilon} - (\rho_{i\infty}+\delta)^{\varepsilon} \big) \right) u^{(i)} \tau^2 \, dx \,.$$

The integration by parts leads to four terms. Those with $\nabla\tau^2$ are treated as T_2 and T_3 above. The terms with $\operatorname{div} u^{(i)}$ are estimated in an analogous way; we present the proof of the most difficult one. We have ($t^* \in (0,1)$)

$$\begin{aligned} &\int \big((\rho_i+\delta)^{1+\varepsilon} - (\rho_{i\infty}+\delta)^{1+\varepsilon} \big) \operatorname{div} u^{(i)} \tau^2 \, dx \\ &\qquad = (1+\varepsilon) \int (\delta + \rho_{i\infty} + t^*(\rho_i - \rho_{i\infty}))^{\varepsilon} (\rho_i - \rho_{i\infty}) \operatorname{div} u^{(i)} \tau^2 \, dx \\ &\qquad \le (1+\varepsilon) K \big(\|\rho_i - \rho_{i\infty}\|_{L^2} + \|\rho_i - \rho_{i\infty}\|_{L^{2(1+\varepsilon)}}^{1+\varepsilon} \big) \|\operatorname{div} u^{(i)}\|_{L^2} \,, \end{aligned}$$

which is bounded due to (4.1) provided that $2(1+\varepsilon) \le 2\gamma$.

The proof of Proposition 4.2 is complete. ∎

PROPOSITION 4.3. *Let $\sigma > 0$ be fixed and let all assumptions of Theorem 1.1 be fulfilled. Then there exists a solution $(\rho, u) = (\rho^\sigma, u^\sigma)$ such that (for $i = 1, 2$)*

$$\rho_i \geq 0, \qquad \rho_i^\sigma - \rho_{i\infty} \in L^2(\mathbb{R}^3) \cap L^{2\gamma}(\mathbb{R}^3)\,, \qquad \text{and} \qquad u_\sigma^{(i)} \in H_0^1(\mathbb{R}^3; \mathbb{R}^3),$$

solving weakly

$$-\sigma \Delta \rho_i^\sigma + \operatorname{div}\left(\rho_i^\sigma u_\sigma^{(i)}\right) = 0 \ \text{in } \mathbb{R}^3\,, \tag{4.16}$$

$$L^{(i)} u_\sigma = -\nabla P_i(\rho_i^\sigma) + \rho_i^\sigma f^{(i)} + J_\sigma^{(i)} \ \text{in } \mathbb{R}^3, \tag{4.17}$$

satisfying the equation for the effective viscous flux

$$A_0\left(P(\rho^\sigma) - P(\rho_\infty)\right) - \begin{pmatrix} \beta_0 \operatorname{div} u_\sigma^{(1)} \\ \operatorname{div} u_\sigma^{(2)} \end{pmatrix} = A_0 \operatorname{div} \Delta^{-1} \begin{pmatrix} \rho_1^\sigma f^{(1)} + J_\sigma^{(1)} \\ \rho_2^\sigma f^{(2)} - J_\sigma^{(1)} \end{pmatrix} \tag{4.18}$$

and the estimates

$$\|\rho_i^\sigma - \rho_{i\infty}\|_{L^{2\gamma}}^{2\gamma} + \|\rho_i^\sigma - \rho_{i\infty}\|_{L^2}^2 + \|\nabla u_\sigma^{(i)}\|_{L^2}^2 \leq K, \tag{4.19}$$

where K is independent of σ.

Proof. Letting $\alpha \to 0$, it follows from (4.1) that there is $(\rho, u) = (\rho^\sigma, u^\sigma)$ such that modulo subsequences ($i = 1, 2$)

$$\rho_i^{\alpha,\sigma} - \rho_{i\infty} \rightharpoonup \rho_i^\sigma - \rho_{i\infty} \ \text{ weakly in } L^2(\mathbb{R}^3) \cap L^{2\gamma}(\mathbb{R}^3)\,, \tag{4.20}$$

$$\nabla u_{\alpha,\sigma}^{(i)} \rightharpoonup \nabla u_\sigma^{(i)} \qquad \text{weakly in } L^2(\mathbb{R}^3)\,, \tag{4.21}$$

and due to the weak lower semicontinuity of the norms, (ρ^σ, u^σ) fulfills (4.19).

Even more, for $q \in [1, 6)$,

$$u_{\alpha,\sigma}^{(i)} \to u_\sigma^{(i)} \text{ strongly in } L_{loc}^q \qquad \text{and} \qquad u_{\alpha,\sigma}^{(i)} \to u_\sigma^{(i)} \text{ a.e. in } \mathbb{R}^3. \tag{4.22}$$

We will show below that (4.1) and (4.13) imply

$$\|\nabla \rho^{\alpha,\sigma}\|_{L_{loc}^{2r}} \leq K \qquad \text{with } r = \frac{2\gamma}{2\gamma + 1 - \varepsilon}\,. \tag{4.23}$$

Consequently, for $s \in [1, r^*)$,

$$\rho^{\alpha,\sigma} \to \rho^\sigma \text{ strongly in } L_{loc}^s(\mathbb{R}^3) \qquad \text{and} \qquad \rho^{\alpha,\sigma} \to \rho^\sigma \text{ a.e. in } \mathbb{R}^3. \tag{4.24}$$

Having (4.20)–(4.24), (4.1), and (4.8) in hand, it is straightforward to pass to the limit, as $\alpha \to 0$, in the weak formulations of (1.35), (1.36) and (3.1) with $w_\beta = 1$ and to obtain (4.16)–(4.19).

Thus, the proof of Proposition 4.3 is complete once we verify (4.23). This, however, follows from

$$\int |\nabla \rho^{\alpha,\sigma}|^{2r}\, dx = \int \left(\frac{|\nabla \rho^{\alpha,\sigma}|^2}{(\rho^{\alpha,\sigma})^{1-\varepsilon}}\right)^r (\rho^{\alpha,\sigma})^{(1-\varepsilon)r}\, dx \overset{(4.13)}{\leq} K \left(\int (\rho^{\alpha,\sigma})^{(1-\varepsilon)\frac{r}{r-1}}\, dx\right)^{1-r}$$

if we put $(1-\varepsilon)\frac{r}{r-1} = 2\gamma$, which implies $r = \frac{2\gamma}{2\gamma+1-\varepsilon}$, and use (4.1). ∎

5. Solvability of the original problem. Due to the estimates (4.19), which are uniform with respect to σ, we have, as $\sigma \to 0$,

$$u_\sigma^{(i)} \rightharpoonup u^{(i)} \text{ weakly in } H^1(\mathbb{R}^3;\mathbb{R}^3) \text{ and } \rho_i^\sigma \rightharpoonup \rho_i \text{ weakly in } L^{2\gamma}_{loc}(\mathbb{R}^3) \tag{5.1}$$

and, owing to the compact embedding,

$$u_\sigma^{(i)} \to u^{(i)} \text{ strongly in } L^2_{loc}(\mathbb{R}^3) \text{ and } \rho_i^\sigma \to \rho_i \text{ strongly in } (H^1_{loc}(\mathbb{R}^3))^*\,, \tag{5.2}$$

and we can pass to the limit in the weak formulations of (4.16) and (4.17). The only difficult terms are those including $P_i(\rho)$. Since ρ_i^σ is uniformly bounded in $L^{2\gamma}$, the weak limit $P_i(\rho^\sigma) \rightharpoonup P_i(\rho)$ will follow if we show that (modulo subsequences)

$$\rho_i^\sigma \to \rho_i \quad \text{a.e. in } \mathbb{R}^3\,. \tag{5.3}$$

Let us emphasize that as soon as (5.3) is proved, the proof of Theorem 1.1 is complete. The rest of this section is focused on proving (5.3).

Proof. Let $\tau = \tau_R$ be our usual localization function: $\tau = 1$ on the ball B_R and $\tau = 0$ on $\mathbb{R}^3 \setminus B_{2R}$. We multiply the ith component of (4.18) by $(\rho_i^\sigma - \rho_i)\tau$ and study the limits of particular terms as $\sigma \to 0$. First of all, we observe that

$$(\operatorname{div}\Delta^{-1}(\rho_i^\sigma f^{(i)} + J_\sigma^{(i)}), (\rho_i^\sigma - \rho_i)\tau) \to 0 \quad \text{as} \quad \sigma \to 0\,, \tag{5.4}$$

which follows from (5.2) and the fact that $(\rho_i^\sigma - \rho_{i\infty})f^{(i)} + \rho_{i\infty}f^{(i)} + J_\sigma^{(i)}$ is bounded in $L^2(\mathbb{R}^3;\mathbb{R}^3)$ (implying that $\operatorname{div}\Delta^{-1}(\rho_i f^{(i)} + J_\sigma^{(i)})$ is bounded in H^1_0). Thanks to (5.4), we have

$$\begin{aligned}
&\lim_{\sigma\to 0}\int A_0\,(P(\rho^\sigma) - P(\rho_\infty))\cdot(\rho^\sigma - \rho)\tau\,dx \\
&\qquad = \lim_{\sigma\to 0}\sum_{i=1}^{2}\hat\beta_i\int \rho_i^\sigma \operatorname{div} u_\sigma^{(i)}\tau\,dx - \sum_{i=1}^{2}\hat\beta_i\int \rho_i \operatorname{div} u^{(i)}\,dx \\
&\qquad = \beta_0 \lim_{\sigma\to 0} D_1^\sigma - \beta_0 D_1 + \lim_{\sigma\to 0} D_2^\sigma - D_2.
\end{aligned} \tag{5.5}$$

Next,

$$\begin{aligned}
D_i^\sigma &= \int \rho_i^\sigma \operatorname{div} u_\sigma^{(i)}\tau\,dx = -\int u_\sigma^{(i)}\nabla\rho_i^\sigma\tau\,dx - \int u_\sigma^{(i)}\rho_i^\sigma\nabla\tau\,dx \\
&= -\int (\rho_i^\sigma+\delta)u_\sigma^{(i)}\nabla\log(\rho_i^\sigma+\delta)\tau\,dx - \int u_\sigma^{(i)}\rho_i^\sigma\nabla\tau\,dx \\
&= \int \operatorname{div}\left((\rho_i^\sigma+\delta)u_\sigma^{(i)}\right)\log(\rho_i^\sigma+\delta)\tau\,dx \\
&\quad + \int (\rho_i^\sigma+\delta)u_\sigma^{(i)}\log(\rho_i^\sigma+\delta)\nabla\tau\,dx - \int u_\sigma^{(i)}\rho_i^\sigma\nabla\tau\,dx =: D_{i1}^\sigma + D_{i2}^\sigma - D_{i3}^\sigma.
\end{aligned}$$

Using (4.16), we obtain

$$D_i^\sigma = \sigma\int \Delta\rho_i^\sigma\log(\rho_i^\sigma+\delta)\tau\,dx + \delta\int \operatorname{div} u_\sigma^{(i)}\log(\rho_i^\sigma+\delta)\tau\,dx + D_{i2}^\sigma - D_{i3}^\sigma\,.$$

and with the help of (4.19) ($F(\xi)$ is the primitive function to $\xi\log\xi$)

$$\begin{aligned}
\sigma\int \Delta\rho_i^\sigma\log(\rho_i^\sigma+\delta)\tau\,dx &\le -\sigma\int \nabla\rho_i^\sigma\log(\rho_i^\sigma+\delta)\nabla\tau\,dx \\
&= \sigma\int F(\rho_i^\sigma+\delta)\Delta\tau\,dx \to 0 \text{ as } \sigma\to 0.
\end{aligned}$$

Furthermore, since $\log(\rho_i^\sigma+\delta)\le\log\delta+\rho_i^\sigma$, we have

$$\begin{aligned}\delta\int \operatorname{div} u_\sigma^{(i)}\log(\rho_i^\sigma+\delta)\tau\,dx &\le \delta\int|\rho_i^\sigma|\,|\operatorname{div} u_\sigma^{(i)}|\tau\,dx+\delta|\log\delta|\int\left|\operatorname{div} u_\sigma^{(i)}\right|\tau\,dx\\ &\le \delta|\log\delta|KR^{\frac32}+K\delta\le\delta|\log\delta|KR^{\frac32}.\end{aligned}$$

Applying (5.2) to $\lim_{\sigma\to0}D_{i3}^\sigma$ we conclude that

$$\lim_{\sigma\to0}D_i^\sigma\le\delta|\log\delta|KR^{\frac32}+\lim_{\sigma\to0}D_{i2}^\sigma-\int\rho_iu^{(i)}\nabla\tau\,dx\,.$$

Since

$$\begin{aligned}\lim_{\sigma\to0}D_{i2}^\sigma&=\lim_{\sigma\to0}\int(\rho_i^\sigma+\delta)u_\sigma^{(i)}\log(\rho_i^\sigma+\delta)\nabla\tau\,dx\\ &=\lim_{\sigma\to0}\int\rho_i^\sigma u_\sigma^{(i)}\log(\rho_i^\sigma+\delta)\nabla\tau\,dx+\text{a term smaller than }\delta|\log\delta|KR^{\frac32}\end{aligned}$$

and

$$\begin{aligned}\int\rho_i^\sigma u_\sigma^{(i)}\log(\rho_i^\sigma+\delta)\nabla\tau\,dx&=\int\rho_i^\sigma u_\sigma^{(i)}\log(\rho_{i\infty}+\delta)\nabla\tau\,dx\\ &\quad+\int\rho_i^\sigma u_\sigma^{(i)}\left(\log(\rho_i^\sigma+\delta)-\log(\rho_{i\infty}+\delta)\right)\nabla\tau\,dx\\ &=:D_{i4}^\sigma+D_{i5}^\sigma\,,\end{aligned}$$

we observe first that

$$D_{i4}^\sigma\to\int\rho_iu^{(i)}\log(\rho_{i\infty}+\delta)\nabla\tau\,dx\quad\text{ as }\sigma\to0\,,$$

and then concentrate on D_{i5}^σ. We choose $\varepsilon_0=\frac{\rho_{i\infty}}{2}$ and we split the integral into one over the set $\{\rho_i^\sigma\le\varepsilon_0\}$, denoted $D_{i5}^\sigma(\rho_i^\sigma\le\varepsilon_0)$, and one over the complement $\{\rho_i^\sigma>\varepsilon_0\}$, denoted $D_{i5}^\sigma(\rho_i^\sigma>\varepsilon_0)$. Since $\int|\rho_i^\sigma-\rho_{i\infty}|^2\,dx\le K$, we have $(\rho_{i\infty}/2)^2|\{\rho_i^\sigma\le\varepsilon_0\}|\le K$, which means that the measure of the set where $\rho_i^\sigma\le\varepsilon_0$ is finite. Consequently,

$$\begin{aligned}D_{i5}^\sigma(\rho_i^\sigma\le\varepsilon_0)&\le K_{\varepsilon_0}\int\limits_{\rho_i^\sigma\le\varepsilon_0}|u_\sigma^{(i)}|\,|\nabla\tau|\,dx\\ &\le K_{\varepsilon_0}\|u_\sigma^{(i)}\|_6\,\|\nabla\tau\|_\infty\,|\{\rho_i^\sigma\le\varepsilon_0\}|^{\frac56}\le K_{\varepsilon_0}R^{-1}\,.\end{aligned}$$

The other part is estimated as follows:

$$\begin{aligned}D_{i5}^\sigma(\rho_i^\sigma>\varepsilon_0)&\le\int\limits_{\rho_i^\sigma>\varepsilon_0}|\rho_i^\sigma-\rho_{i\infty}|\,|u_\sigma^{(i)}|\,|\log(\rho_i^\sigma+\delta)-\log(\rho_{i\infty}+\delta)|\,|\nabla\tau|\,dx\\ &\quad+\rho_{i\infty}\int\limits_{\rho_i^\sigma>\varepsilon_0}|u_\sigma^{(i)}|\,|\log(\rho_i^\sigma+\delta)-\log(\rho_{i\infty}+\delta)|\,|\nabla\tau|\,dx\\ &=:E_{0\sigma}+E_{1\sigma}\,.\end{aligned}$$

Since $\log(\xi+\delta)$ is a K_{ε_0}-Lipschitz function on $\{\xi \geq \varepsilon_0\}$, we have

$$E_{1\sigma} \leq K_{\varepsilon_0}\rho_{i\infty}\int_{B_{2R}\setminus B_R} |u_\sigma^{(i)}|\,|\rho_i^\sigma - \rho_{i\infty}|\;|\nabla\tau|\,dx$$

and hence, since $\|\rho_i^\sigma - \rho_{i\infty}\|_{L^2} \leq K$, via Hölder's inequality and with (5.2)

$$\lim_{\sigma\to 0} E_{1\sigma} \leq K_{\varepsilon_0} K\left(\int_{B_{2R}\setminus B_R} |\nabla\tau|^2\,|u^{(i)}|^2\,dx\right)^{\frac{1}{2}}.$$

The other term $E_{0\sigma}$ is split again, with $L := 2\rho_\infty$ this time:

$$\int\limits_{\rho_i^\sigma>\varepsilon_0}\cdots = \int\limits_{L>\rho_i^\sigma>\varepsilon_0}\cdots + \int\limits_{\rho_i^\sigma\geq L}\cdots =: E_{2\sigma} + E_{3\sigma}\,.$$

Since $\int |\rho_i^\sigma - \rho_{i\infty}|^2\,dx \leq K$ uniformly, $|\{\rho_i^\sigma \geq L\}| \leq K$, and we have

$$\int\limits_{\rho_i^\sigma\geq L} |\rho_i^\sigma|^2\,dx \leq K + \int\limits_{\rho_i^\sigma\geq L} |\rho_{i\infty}|^2\,dx \leq K.$$

Similarly, for all $s>1$ we have $\int_{\rho_i^\sigma\geq L} |\log(\rho_i^\sigma+\delta)|^s\,dx \leq K_s$. Thus, we conclude from Hölder's inequality (recalling $\|u_\sigma^{(i)}\|_6 \leq K$) that $|E_{3\sigma}| \leq K\|\nabla\tau\|_\infty \leq KR^{-1}$.

To estimate $E_{2\sigma}$, we use $|\log(\rho_i^\sigma+\delta)| \leq K_{\varepsilon_0,L}$ on $\{L \geq \rho_i^\sigma \geq \varepsilon_0\}$. Thus

$$E_{2\sigma} \leq LK_{\varepsilon_0,L}\int_{\mathbb{R}^3} |u_\sigma^{(i)}|\,|\rho_i^\sigma - \rho_{i\infty}|\,|\nabla\tau|\,dx \leq LK_{\varepsilon_0}K\left(\int_{B_{2R}\setminus B_R} |u_\sigma^{(i)}|^2|\nabla\tau|^2\,dx\right)^{\frac{1}{2}},$$

and thanks to (5.2) we can pass to limit as $\sigma\to 0$.

Collecting our estimates, we obtain

$$\begin{aligned}\lim_{\sigma\to 0} D_i^\sigma \leq{}& \delta|\log\delta|KR^{\frac{3}{2}} - \int \rho_i u^{(i)}\nabla\tau\,dx + \int \rho_i u^{(i)}\log(\rho_{i\infty}+\delta)\nabla\tau\,dx\\ &+ K_{\varepsilon_0}R^{-1} + LK_{\varepsilon_0}K\left(\int_{B_{2R}\setminus B_R} |u^{(i)}|^2\,|\nabla\tau|^2\,dx\right)^{\frac{1}{2}}.\end{aligned}$$

We now analyze the terms D_i introduced in (5.5):

$$D_i = \int \rho_i \operatorname{div} u^{(i)}\tau\,dx = \int (\rho_i * \omega_h)\operatorname{div} u^{(i)}\tau\,dx + \varepsilon_h\,,$$

where ω_h is a mollifier and $\varepsilon_h \to 0$ as the parameter h tends to 0. We have

$$\begin{aligned}
D_i &= -\int \nabla(\rho_i * \omega_h) u^{(i)} \tau \, dx - \int (\rho_i * \omega_h) u^{(i)} \nabla\tau \, dx + \varepsilon_h \\
&= -\int (\rho_i * \omega_h + \delta) u^{(i)} \nabla \log(\rho_i * \omega_h + \delta) \tau \, dx - \int (\rho_i * \omega_h) u^{(i)} \nabla\tau \, dx + \varepsilon_h \\
&= \int \operatorname{div}\left[(\rho_i * \omega_h + \delta) u^{(i)}\right] \log(\rho_i * \omega_h + \delta) \tau \, dx \\
&\quad + \int (\rho_i * \omega_h) u^{(i)} \log(\rho_i * \omega_h + \delta) \nabla\tau \, dx + \delta \int u^{(i)} \log(\rho_i * \omega_h + \delta) \nabla\tau \, dx \\
&\quad - \int (\rho_i * \omega_h) u^{(i)} \nabla\tau \, dx + \varepsilon_h \\
&=: D_{i1} + D_{i2} + \delta \int u^{(i)} \log(\rho_i * \omega_h + \delta) \nabla\tau \, dx - D_{i3} + \varepsilon_h \,,
\end{aligned}$$

where $\delta \int u^{(i)} \log(\rho_i * \omega_h + \delta) \nabla\tau \, dx \le \delta \log \delta \, K$. The terms $\lim_{h\to 0} D_{i3}$ will cancel with the terms $\lim_{\sigma\to 0} D^\sigma_{i3}$. The terms D_{i1} are treated in a way known from the theory of compressible fluids (cf. [19, 7, 9]), namely, with the aid of the lemma of DiPerna and Lions (cf. [6]) we know that

$$\tau \left(\operatorname{div}\left[(\rho_i * \omega_h) u^{(i)}\right] - \operatorname{div}\left[\omega_h * (\rho_i u^{(i)})\right] \right) \rightharpoonup 0 \text{ weakly in } L^{\frac{2\gamma}{\gamma+1}} \quad (h \to 0) \,. \tag{5.6}$$

(In fact, we use $\rho_i \in L^{2\gamma}_{loc}(\mathbb{R}^3), \nabla u^{(i)} \in L^2(\mathbb{R}^3)$, write down the explicit definition of the two mollified terms, and estimate $\nabla\omega_h(r)\rho_i(x+r)\,(u(x) - u(x+r))$ by using Poincaré's inequality for u, which gives the factor $r \sim h$ and the estimate $|\nabla\omega_h(r)| \le \frac{K}{h}$.)

Since $\operatorname{div}\left(\rho_i u^{(i)}\right) = 0$ weakly, we conclude from (5.6) that

$$\tau \operatorname{div}\left[(\rho_i * \omega_h) u^{(i)}\right] \rightharpoonup 0 \text{ weakly in } L^{\frac{2\gamma}{\gamma+1}} \quad (h \to 0) \quad (\delta, R \text{ fixed}) \,.$$

Since $\log(\rho_i * \omega_h + \delta)\tau$ is uniformly bounded in L^q for all q when δ and R are fixed, we conclude that

$$\begin{aligned}
D_{i1} &= \int \operatorname{div}\left[(\rho_i * \omega_h + \delta) u^{(i)}\right] \log(\rho_i * \omega_h + \delta) \tau \, dx \\
&= \varepsilon^1_h + \delta \int \operatorname{div} u^{(i)} \log(\rho_i * \omega_h + \delta) \tau \, dx \to \delta \int \operatorname{div} u^{(i)} \log(\rho_i + \delta) \tau \, dx \quad (h \to 0) \,.
\end{aligned}$$

It remains to treat the term D_{i2}. As $D_{i2} \to \int \rho_i u^{(i)} \log(\rho_i + \delta) \nabla\tau \, dx$ for $h \to 0$, we have

$$\begin{aligned}
\lim_{h\to 0} D_{i2} &= \log(\rho_{i\infty} + \delta) \int \rho_i u^{(i)} \nabla\tau \, dx + \int \rho_i u^{(i)} \left(\log(\rho_i + \delta) - \log(\rho_{i\infty} + \delta)\right) \nabla\tau \, dx \\
&=: D_{i4} + D_{i5} \,.
\end{aligned}$$

The term D_{i4} cancels with the corresponding term $\lim_{\sigma\to 0} D^\sigma_{i4}$ and the term D_{i5} is treated in the same way as D^σ_{i5} above; one obtains

$$D_{i5} \le K R^{-1} + K \int_{B_{2R} \setminus B_R} |u^{(i)}|^2 \, |\nabla\tau|^2 \, dx \,.$$

Bringing all these estimates together, we arrive at the inequality

$$\lim_{\sigma\to 0} D_i^\sigma - D_i \le K\delta|\log\delta|R^{\frac{3}{2}} + KR^{-1} + K\int_{B_{2R}\setminus B_R} |u^{(i)}|^2\,|\nabla\tau|^2\,dx\,.$$

We first let $\delta\to 0$, then let $R\to\infty$. The term $\int_{B_{2R}\setminus B_R} |u^{(i)}|^2\,|\nabla\tau|^2\,dx \to 0$ since $|\nabla\tau|^2 \in L^{\frac{3}{2}}(\mathbb{R}^3)$ and $|u^{(i)}|^2 \in L^3(\mathbb{R}^3)$ (implying that $\|u^{(i)}\|_{L^6(B_{2R}\setminus B_R)} \to 0$). Hence we obtain

$$\lim_{\sigma\to 0}\int A_0 P(\rho^\sigma)\cdot(\rho^\sigma-\rho)\,dx = 0 \text{ in } L^2\,,$$

and the strong convergence $\rho_i^\sigma \to \rho_i$ follows from the monotonicity condition (1.23) and (5.3) follows. □

Acknowledgment. J. Málek thanks K. R. Rajagopal and L. Tao for many useful discussions explaining the modeling of the mixtures in the framework of continuum mechanics.

REFERENCES

[1] S. N. Antontsev, A. V. Kazhikhov, and V. N. Monakhov, *Boundary Value Problems in Mechanics of Nonhomogeneous Fluids*, North–Holland, Amsterdam, 1990 (in English).
[2] H. Darcy, *Les Fontaines Publiques de la Ville de Dijon*, Victor Dalmont, Paris, 1856.
[3] B. Desjardins, *Regularity of weak solutions of the compressible isentropic Navier-Stokes equations*, Comm. Partial Differential Equations, 22 (1997), pp. 977–1008.
[4] B. Desjardins, *Regularity results for two-dimensional flows of multiphase viscous fluids*, Arch. Rational Mech. Anal., 137 (1997), pp. 135–158.
[5] B. Desjardins, *On the regularity of solutions of the compressible isentropic Navier-Stokes equations*, in Hyperbolic Problems: Theory, Numerics, Applications, Vol. I (Zürich, 1998), Internat. Ser. Numer. Math 129, Birkhäuser, Basel, 1999, pp. 215–224.
[6] R. J. DiPerna and P.-L. Lions, *Ordinary differential equations, transport theory and Sobolev spaces*, Invent. Math., 98 (1989), pp. 511–547.
[7] E. Feireisl, *On compactness of solutions to the compressible isentropic Navier-Stokes equations when the density is not square integrable*, Comment. Math. Univ. Carolin., 42 (2001), pp. 83–98.
[8] E. Feireisl, *Dynamics of Viscous Compressible Fluids*, Oxford University Press, Oxford, UK, 2004.
[9] E. Feireisl, A. Novotný, and H. Petzeltová, *On the existence of globally defined weak solutions to the Navier-Stokes equations*, J. Math. Fluid Mech., 3 (2001), pp. 358–392.
[10] A. Fick, *Über Diffusion*, Annalen der Physik und Chemie, 94 (1855), pp. 59–86.
[11] J. Frehse, S. Goj, and J. Málek, *A Stokes-like system for mixtures*, in Nonlinear Problems in Mathematical Physics and Related Topics, II, Int. Math. Ser. 2, M. S. Birman, S. Hildebrandt, V. Solonnikov, and N. Uraltseva, eds., Kluwer Plenum, New York, 2002.
[12] J. Frehse, S. Goj, and J. Málek, *Existence of solutions to a Stokes-like system for mixtures*, Preprint 103, SFB611, University of Bonn, Bonn, Germany, 2003.
[13] J. Frehse, S. Goj, and J. Málek, *A uniqueness result for a model for mixtures in the absence of external forces and interaction momentum*, Appl. Math., accepted.
[14] J. Frehse and W. Weigant, in preparation.
[15] J. Frehse and W. Weigant, *On quasi-stationary compressible miscible mixtures*, in preparation.
[16] Y. Giga and S. Takahashi, *On global weak solutions of the nonstationary two-phase Stokes flow*, SIAM J. Math. Anal., 25 (1994), pp. 876–893.
[17] A. V. Kazhikhov and A. N. Petrov, *Well-posedness of the initial-boundary value problem for a model system of equations of a multicomponent mixture*, Dinamika Sploshn. Sredy, 174 (1978), pp. 61–73.
[18] P.-L. Lions, *Mathematical Topics in Fluid Mechanics.* Vol. 1. *Incompressible Models*, Oxford Lecture Ser. Math. Appl. 3, Clarendon Press, Oxford University Press, New York, 1996.

[19] P.-L. Lions, *Mathematical Topics in Fluid Mechanics.* Vol. 2. *Compressible Models*, Oxford Lecture Ser. Math. Appl. 10, Clarendon Press, Oxford University Press, New York, 1998.
[20] A. Nouri and F. Poupaud, *An existence theorem for the multifluid Navier-Stokes problem*, J. Differential Equations, 122 (1995), pp. 71–88.
[21] A. Nouri, F. Poupaud, and Y. Demay, *An existence theorem for the multi-fluid Stokes problem*, Quart. Appl. Math., 55 (1997), pp. 421–435.
[22] A. Novotný, *Some remarks to the compactness of steady compressible isentropic Navier-Stokes equations via the decomposition method*, Comment. Math. Univ. Carolin., 37 (1996), pp. 305–342.
[23] A. Novotný, *Compactness of steady compressible isentropic Navier-Stokes equations via the decomposition method (the whole 3-D space)*, in Theory of the Navier-Stokes Equations, World Scientific, River Edge, NJ, 1998, pp. 106–120.
[24] K. Rajagopal and L. Tao, *Mechanics of Mixtures*, World Scientific, River Edge, NJ, 1995.
[25] K. R. Rajagopal, *On the flow of mixtures between parallel plates*, in Recent Advances in the Mechanics of Structured Continua, M. Massoudi and K. Rajagopal, eds., ASME, New York, 2000, pp. 125–137.
[26] C. Truesdell, *Rational Thermodynamics. With an Appendix by C. C. Wang,* 2nd ed., Springer-Verlag, New York, 1984.
[27] A. A. Zlotnik, *Uniform estimates and stabilization of the solutions of a system of equations of the one-dimensional motion of a multicomponent barotropic mixture*, Mat. Zametki, 58 (1995), pp. 307–312.

SIAM J. MATH. ANAL.
Vol. 36, No. 4, pp. 1282–1292

OPTIMAL STABILITY OF RECONSTRUCTION OF PLANE LIPSCHITZ CRACKS*

LUCA RONDI†

Abstract. We establish an optimal stability estimate for the determination of a finite number of Lipschitz perfectly insulating cracks inside a planar conductor by performing two suitably chosen electrostatic boundary measurements.

Key words. inverse problems, stability, Lipschitz cracks, corkscrew condition

AMS subject classifications. Primary, 35R30; Secondary, 78A30

DOI. 10.1137/S0036141003435837

1. Introduction.

We study the inverse problem of determining a finite number of unknown perfectly insulating cracks σ_j, $j = 1, \dots, N$, whose union is denoted with Σ, inside a known, possibly inhomogeneous and anisotropic, planar conductor Ω, whose known background conductivity is given by A, through voltage and current electrostatic measurements at the boundary.

We prescribe two current densities ψ_1, ψ_2, and we measure on Γ_0, a subarc of the boundary of Ω, $\partial\Omega$, the corresponding electrostatic potentials u_i. We recall that the electrostatic potential u_i satisfies the following Neumann-type boundary value problem:

$$\begin{cases} \operatorname{div}(A\nabla u_i) = 0 & \text{in } \Omega\backslash\Sigma, \\ A\nabla u_i \cdot \nu = 0 & \text{on either side of } \sigma_j,\ j = 1, \dots, N, \\ A\nabla u_i \cdot \nu = \psi_i & \text{on } \partial\Omega, \end{cases} \tag{1.1}$$

where ν denotes the unit normal, with the outward orientation when on $\partial\Omega$.

If ψ_1 and ψ_2 are suitably chosen—for example, they can model a two-electrode configuration where the positive electrode is kept fixed whereas the negative one is moved in a different position as we change the current density from ψ_1 to ψ_2—then the measurements $u_i|_{\Gamma_0}$, $i = 1, 2$, uniquely determine the unknown multiple crack Σ.

This inverse problem was introduced in [8], where the first uniqueness result in two dimensions was proved. Since then, many results concerning uniqueness and stability have been obtained; we refer to [5] and the references therein for a detailed account of these issues in two and three dimensions.

We are interested in estimating the error, in the Hausdorff distance, on the determination of Σ from an estimate of the error on the measurements $u_i|_{\Gamma_0}$. It has already been proven that

(a) if the components of Σ are a priori known to be Lipschitz regular, then the stability estimate is of log-log type (see [12, Theorem 4.1, part (I)]);
(b) if we a priori know either the coordinate system with respect to which the components of Σ are Lipschitz regular, or that the components of Σ are $C^{1,\alpha}$

*Received by the editors October 10, 2003; accepted for publication (in revised form) April 23, 2004; published electronically February 3, 2005. This research was supported by MIUR under grant 2002013279.

http://www.siam.org/journals/sima/36-4/43583.html

†Dipartimento di Scienze Matematiche, Università degli Studi di Trieste, via Valerio, 12/1 34127 Trieste, Italy (rondi@mathsun1.univ.trieste.it).

regular, with $0 < \alpha \leq 1$, then the stability estimate is of log type (see [12, Theorem 4.1, parts (II) and (III)]).

These results were proved first in the case of a single crack. More precisely, part (a) in [3] and part (b), at least for what concerns the $C^{1,\alpha}$ case, in [11]. Their extension to the case of multiple cracks is essentially based on arguments developed in [4] for the treatment of the multiple cavities case.

The aim of the present paper is to fill the gap between cases (a) and (b). In fact, we prove that, under the same assumptions as those of [12, Theorem 4.1, part (I)], Lipschitz regularity is enough to obtain a stability estimate of log type. We remark that single log estimates are usually obtained through a two-step procedure; see [1], where this argument was developed for the first time. For example, the proof of (b) relies on (a), as the first step, and, as the second step, on carefully studying the relation between two unknown multiple cracks Σ and Σ' (corresponding to two different sets of measurements) if they are close enough in the Hausdorff distance, in particular, on proving a uniform interior cone property for the open set $\Omega\backslash(\Sigma \cup \Sigma')$. However, if we consider case (a), it might happen that no kind of uniform interior cone property holds for all the points of $\Sigma \cup \Sigma'$, no matter how close the two multiple cracks are in the Hausdorff distance. On the other hand, if we consider the proof of [12, Theorem 4.1], it is clear that the arguments, at a given stage, in particular in the proof of Proposition 4.9 in [12] (or of Proposition 5.1 in [3] for the case of a single crack), are developed only locally in a suitable neighborhood of the point z where the Hausdorff distance between Σ and Σ' is reached. In this paper we establish that if Σ and Σ' are Lipschitz and close enough, then the points in $\Sigma \cup \Sigma'$ belonging to such a suitable neighborhood of the point z can be reached through a suitable sequence of discs contained in $\Omega\backslash(\Sigma \cup \Sigma')$; see Lemma 3.3. Such a condition, which allows us to carry over the second step of the procedure, is similar to the so-called *corkscrew condition* used in [9] to define nontangentially accessible domains.

We wish to emphasize that logarithmic stability estimates are optimal for this inverse problem. In fact, the abstract method developed in [6] from an idea of Mandache [10] provides the instability character of the problem; see [7], an expanded version of [6], for details.

Finally we wish to remark that, with a completely analogous procedure, we can extend this stability result to the inverse problem of multiple cavities. That is, if we perform two measurements of this kind, corresponding to prescribed current densities ψ_1 and ψ_2 as above, then we can obtain a stability estimate of log type for the determination of Lipschitz multiple cavities. We notice that we have uniqueness and stability results for the determination of multiple cavities with a single measurement, which can be of the most general type; see [3] and the references therein for the two-dimensional case and [2] for the higher-dimensional one. However, in the planar case, with a single measurement the stability estimate is of log-log type if the cavities are assumed to be Lipschitz, and it is of log type if the cavities satisfy the conditions described in case (b) above. Unfortunately, the technique used to prove the stability results with a single measurement is quite different, even if it has many common features with the one used for the inverse crack problem. In particular, in order to exploit the fact that the defects Σ and Σ' are the closures of open sets, we need to study the stability of a Cauchy-type problem up to the whole boundary of $\Omega\backslash(\Sigma \cup \Sigma')$; thus we cannot restrict our analysis to a neighborhood of the point where the Hausdorff distance is reached. This can be clearly observed once we notice that, locally, we are not able to distinguish between a portion of a crack and a portion of the boundary of

a cavity. Therefore the approach developed in this paper cannot be directly applied to improve the stability with a single measurement. It remains an interesting open problem to establish log-type estimates for the determination of Lipschitz multiple cavities by a single measurement.

The plan of the paper is as follows. In section 2 we precisely state the stability result Theorem 2.3 and make some preliminary considerations. In section 3 the proof of Theorem 2.3 is developed.

2. Statement of the stability result. We begin with the definition of a quantitative notion of smoothness for open and closed curves in $\mathbb{R}^2$. The following standard notation will be used. For every $z = x + \mathrm{i}y \in \mathbb{C}$, $x = \Re z$ and $y = \Im z$ being the real and imaginary parts of z, respectively, and for every $r > 0$, we denote with $B_r(z)$ the open disc with center z and radius r. As usual, we shall identify complex numbers $z = x + \mathrm{i}y \in \mathbb{C}$ with points $(x, y) \in \mathbb{R}^2$. We shall use the following notation for complex derivatives:

$$f_{\overline{z}} = (f_x + \mathrm{i}f_y)/2, \quad f_z = (f_x - \mathrm{i}f_y)/2.$$

We denote by $J = \left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right]$ the counterclockwise rotation of 90° and by $(\cdot)^T$ transpose.

DEFINITION 2.1. *Let $\gamma \subset \mathbb{R}^2$ be a bounded, simple curve, either open or closed. Then, with two fixed positive constants δ and M, we say that γ is* Lipschitz *with constants δ, M if for every $z \in \gamma$ there exists a coordinate system (x, y) with origin in z such that with respect to these coordinates $\gamma \cap B_\delta(z)$ is a* Lipschitz graph *with constant M, that is, $\gamma \cap B_\delta(z) = \{y = \phi(x) : a \le x \le b\} \cap B_\delta(z)$, where ϕ is a Lipschitz function on $[-\delta, \delta]$ such that $\|\phi\|_{C^{0,1}[-\delta,\delta]} \le M$, a and b satisfy $-\delta \le a \le 0 \le b \le \delta$, and at least one of them has modulus equal to δ.*

Let $\Omega \subset \mathbb{R}^2$ be a bounded, simply connected domain. We say that $\sigma \subset \Omega$ is a *crack* in Ω if it is a closed set in Ω, which can be represented as the image of a simple open curve. We say that $\Sigma \subset \Omega$ is a *multiple crack* in Ω if it is the finite union of pairwise disjoint cracks in Ω.

We suppose that the following assumptions on the data of the inverse problem and the following a priori information on the unknown multiple crack present in Ω hold. We wish to remark that these assumptions and a priori information are essentially minimal and coincide with those used in previous papers, and we repeat them here for the convenience of the reader.

Assumptions on the domain. Let Ω be a bounded, simply connected domain in $\mathbb{R}^2$. We assume that the diameter of Ω is bounded by a given positive constant L and that its boundary $\partial\Omega$ is a simple closed curve which is Lipschitz with given positive constants δ, M.

From these assumptions we may deduce the following properties of Ω. We may find a constant L_1 depending on δ, M, and L only such that

$$0 < \delta \le \mathrm{length}(\partial\Omega) \le L_1.$$

Furthermore, there exists a constant M_1, depending on δ, M, and L only, such that

$$\mathrm{length}_{\partial\Omega}(z_0, z_1) \le M_1 |z_0 - z_1| \quad \text{for any } z_0, z_1 \in \partial\Omega. \tag{2.1}$$

Here $\mathrm{length}_{\partial\Omega}(z_0, z_1)$ is the length of the smallest arc in $\partial\Omega$ connecting z_0 to z_1. Moreover the measure of Ω, $|\Omega|$, is bounded from below and above by positive constants depending on δ, M, and L only.

Assumptions on the background conductivity. Let $A = A(z)$, $z \in \Omega$, be a conductivity tensor with bounded measurable entries satisfying, for given positive constants λ and Λ,

$$\begin{aligned} &A(z)\xi \cdot \xi \geq \lambda |\xi|^2 \quad \text{for every } \xi \in \mathbb{R}^2 \text{ and for a.e. } z \in \Omega, \\ &|a_{ij}(z)| \leq \Lambda \quad \text{for every } i,j = 1,2 \text{ and for a.e. } z \in \Omega. \end{aligned} \tag{2.2}$$

Assumptions on the boundary data. Let γ_0, γ_1, γ_2 be three fixed simple arcs in $\partial\Omega$, pairwise internally disjoint.

Given $H > 0$, let us fix three functions η_0, η_1, $\eta_2 \in L^2(\partial\Omega)$ such that for every $i = 0, 1, 2$

$$\begin{aligned} &\eta_i \geq 0 \text{ on } \partial\Omega, \quad \operatorname{supp}(\eta_i) \subset \gamma_i, \\ &\int_{\partial\Omega} \eta_i = 1, \quad \|\eta_i\|_{L^2(\partial\Omega)} \leq H. \end{aligned} \tag{2.3}$$

Then we prescribe the current densities on the boundary ψ_1, ψ_2 to be given by

$$\psi_1 = \eta_0 - \eta_1, \quad \psi_2 = \eta_0 - \eta_2. \tag{2.4}$$

We have

$$\int_{\partial\Omega} \psi_i = 0, \quad \|\psi_i\|_{L^2(\partial\Omega)} \leq 2H \quad \text{for every } i = 1, 2. \tag{2.5}$$

We shall consider also the antiderivatives along $\partial\Omega$ of ψ_1, ψ_2,

$$\Psi_i(s) = \int \psi_i(s)\mathrm{d}s, \quad i = 1, 2, \tag{2.6}$$

where the indefinite integral is taken, as usual, with respect to arclength on $\partial\Omega$ in the counterclockwise direction. The functions Ψ_1, Ψ_2 are defined up to an additive constant.

We remark that from the assumptions on Ω, through (2.1), we have that, for every $i = 1, 2$, Ψ_i satisfies the following Hölder continuity property for any z_0, $z_1 \in \partial\Omega$:

$$|\Psi_i(z_0) - \Psi_i(z_1)| \leq 2H(\operatorname{length}_{\partial\Omega}(z_0, z_1))^{1/2} \leq H_1 |z_0 - z_1|^{1/2}, \tag{2.7}$$

where $H_1 = 2HM_1^{1/2}$, M_1 as in (2.1).

Assumptions on the measurements. Let $\Gamma_0 \subset \partial\Omega$ be a subarc whose length is greater than or equal to δ.

A priori information on the multiple interior crack. We assume that an admissible multiple crack $\Sigma \subset \Omega$ is the union of finitely many, pairwise disjoint cracks σ_j, $j = 1, \ldots, N$, $N \geq 1$.

We suppose that each crack σ_j, $j = 1, \ldots, N$, is Lipschitz with constants δ, M. Moreover we suppose that

$$\operatorname{dist}(\Sigma, \partial\Omega) \geq \delta \tag{2.8}$$

and that

$$\operatorname{dist}(\sigma_j, \sigma_l) \geq \delta \quad \text{for any } j \neq l. \tag{2.9}$$

Let us make some remarks about the properties of the admissible multiple cracks. First we notice that Σ is not empty and each component of Σ is a simple open curve whose length is bounded from below and above by positive constants depending on δ, M, and L only.

Let Σ and $\Sigma' = \bigcup_{l=1}^{N'} \sigma'_l$, $N' \geq 1$, be two multiple interior cracks satisfying the a priori information. Then the following lemma is easy to prove. We recall that we denote the Hausdorff distance with d_H and that throughout the paper we set

$$p = \mathrm{d}_H(\Sigma, \Sigma').$$

LEMMA 2.2. *There exists a constant $p_0 > 0$, depending on δ, M, and L only, such that if $p \leq p_0$, then these two properties hold.*

First, the number of connected components of Σ and Σ' is the same, for instance, equal to N, and, up to rearranging their order and swapping Σ with Σ', we can assume that

$$\mathrm{d}_H(\sigma_j, \sigma'_j) \leq \mathrm{d}_H(\Sigma, \Sigma') \quad \textit{for every } j = 1, \dots, N \tag{2.10}$$

and that there exists $z'_0 \in \sigma'_1$ so that

$$\mathrm{dist}(z'_0, \sigma_1) = \mathrm{d}_H(\sigma_1, \sigma'_1) = p. \tag{2.11}$$

Furthermore, $\Sigma \cup \Sigma' \subset \partial G$, where G is the connected component of $\Omega \backslash (\Sigma \cup \Sigma')$ whose boundary contains $\partial\Omega$.

For any $i = 1, 2$, let $u_i \in W^{1,2}(\Omega \backslash \Sigma)$ be the weak solution to (1.1). That is, we understand that u_i satisfies

$$\int_{\Omega \backslash \Sigma} A \nabla u_i \cdot \nabla \varphi = \int_{\partial\Omega} \psi_i \varphi \quad \text{for any } \varphi \in W^{1,2}(\Omega \backslash \Sigma).$$

We remark that u_i is unique up to additive constants. We denote by u'_i the solution to (1.1) when Σ is replaced with Σ'.

The set of constants δ, M, L, λ, Λ, and H will be referred to as the *a priori data*. We are now in position to state the main result.

THEOREM 2.3. *Under the previously stated assumptions, let $\varepsilon > 0$ be such that*

$$\max_{i=1,2} \|u_i - u'_i\|_{L^\infty(\Gamma_0)} \leq \varepsilon. \tag{2.12}$$

Then

$$\mathrm{d}_H(\Sigma, \Sigma') \leq \omega(\varepsilon), \tag{2.13}$$

where $\omega : (0, +\infty) \mapsto (0, +\infty)$ satisfies

$$\omega(\varepsilon) \leq K |\log \varepsilon|^{-\beta} \quad \textit{for every } \varepsilon,\ 0 < \varepsilon < 1/\mathrm{e}, \tag{2.14}$$

and K, $\beta > 0$ depend on the a priori data only.

We conclude this section by describing some properties of the solution to (1.1) and its stream function, assuming that the hypotheses of Theorem 2.3 are satisfied. For details and proofs we refer to [12, Chapter 4].

Let $i = 1, 2$ and let u_i solve (1.1). Then there exists a global single-valued function $v_i \in W^{1,2}(\Omega \backslash \Sigma)$ which satisfies

$$\nabla v_i = JA \nabla u_i \quad \text{almost everywhere in } \Omega \backslash \Sigma. \tag{2.15}$$

Such a function is referred to as the *stream function* associated to u_i. Moreover, letting $f_i = u_i + \mathrm{i}v_i$, we have

$$(f_i)_{\overline{z}} = \mu_1 (f_i)_z + \mu_2 \overline{(f_i)_z} \quad \text{almost everywhere in } \Omega\backslash\Sigma, \tag{2.16}$$

where μ_1 and μ_2 are bounded, measurable, complex-valued coefficients which depend on f_i and satisfy

$$|\mu_1| + |\mu_2| \leq k < 1 \quad \text{almost everywhere in } \Omega\backslash\Sigma, \tag{2.17}$$

where k is a constant depending on λ, Λ only.

Moreover, v_i satisfies in the weak sense the following Dirichlet-type boundary value problem:

$$\begin{cases} \operatorname{div}(B\nabla v_i) = 0 & \text{in } \Omega\backslash\Sigma, \\ v_i = d_j & \text{on } \sigma_j,\ j = 1, \dots, N, \\ v_i = \Psi_i & \text{on } \partial\Omega, \\ \int_\gamma B\nabla v_i \cdot \nu = 0 & \text{for any smooth Jordan curve } \gamma \subset \Omega\backslash\Sigma, \end{cases} \tag{2.18}$$

where $B = (\det A)^{-1} A^T$. We remark that the constants d_j are unknown and depend on $i = 1, 2$.

The weak formulation of (2.18) is the following. We want to find $v_i \in W^{1,2}(\Omega)$ such that v_i is constant in the trace sense on any crack σ_j, its trace on $\partial\Omega$ equals Ψ_i, and it satisfies

$$\int_{\Omega\backslash\Sigma} B\nabla v_i \cdot \nabla\varphi = 0 \quad \text{for any } \varphi \in W_0^{1,2}(\Omega) : \varphi = \text{const. on any crack.}$$

Let us finally remark that the stream function v_i is unique up to additive constants.

For any $i = 1, 2$, the following Hölder estimates hold (see [12, Proposition 4.6]):

$$|v_i(z_1) - v_i(z_2)| \leq C_1 |z_1 - z_2|^{\alpha_1} \quad \text{for every } z_1, z_2 \in \overline{\Omega}, \tag{2.19}$$

$$|u_i(z_1) - u_i(z_2)| \leq C_1 (\tilde{d}(z_1, z_2))^{\alpha_1} \quad \text{for every } z_1, z_2 \in \tilde{\Omega}. \tag{2.20}$$

Here C_1 and $\alpha_1 > 0$ depend on the a priori data only. We denote with $\tilde{\Omega}$ the compact manifold obtained by the appropriate gluing of $\overline{\Omega}\backslash\Sigma$ to the degenerate simple closed curve $\tilde{\sigma}_j$ obtained by overlapping two copies of σ_j, $j = 1, \dots, N$, and with $\tilde{d}$ the geodesic distance on $\tilde{\Omega}$.

It is useful to stress the difference between the estimates (2.19), (2.20). In fact, since v_i is constant on each σ_j, it is expected that v_i is continuous across each σ_j. Instead u_i may have different one-sided limits on σ_j. This is the main motivation for the introduction of the metric $\tilde{d}$.

For any $i = 1, 2$, let v_i' be the stream function associated to u_i' and $f_i' = u_i' + \mathrm{i}v_i'$. In what follows, we shall always normalize v_i and v_i' in such a way that $v_i = v_i'$ on $\partial\Omega$. Then we have that, for any $i = 1, 2$,

$$\|f_i - f_i'\|_{L^\infty(\Gamma_0)} \leq \varepsilon, \tag{2.21}$$

and, by (2.19), (2.20) and by assuming that $\varepsilon \leq 1/\mathrm{e}$,

$$\|f_i - f_i'\|_{L^\infty(\Omega)} \leq C_2, \tag{2.22}$$

where C_2 depends on the a priori data only.

Furthermore (see [12, Proposition 4.11]),

$$|v_i(z) - v_i'(z)| \le \eta(\varepsilon) \quad \text{for any } z \in \overline{\Omega}, \tag{2.23}$$

where η is a positive function defined on $(0, +\infty)$ such that

$$\eta(\varepsilon) \le C_3(\log|\log\varepsilon|)^{-\alpha_2} \quad \text{for every } \varepsilon,\ 0 < \varepsilon < 1/\mathrm{e}. \tag{2.24}$$

Here C_3 and α_2 are positive constants depending on the a priori data only.

3. Proof of Theorem 2.3. We begin with the following two results.

THEOREM 3.1. *Theorem* 2.3 *holds true if we replace* (2.14) *with*

$$\omega(\varepsilon) \le K_1(\log|\log(\varepsilon)|)^{-\beta_1} \quad \textit{for every } \varepsilon,\ 0 < \varepsilon < 1/\mathrm{e}, \tag{3.1}$$

K_1, $\beta_1 > 0$ depending on the a priori data only.

PROPOSITION 3.2. *Suppose that the assumptions of Theorem* 2.3*, with the exception of* (2.12)*, are satisfied. Let us further assume that $p \le p_0$, and hence* (2.10) *and* (2.11) *are satisfied.*

If there exist positive constants c_0 and η such that for every r, $0 \le r \le c_0 p$, there exists $z' \in \sigma_1' \cap \partial B_r(z_0')$ such that

$$|v_i(z') - v_i'(z')| \le \eta \quad \textit{for any } i = 1, 2, \tag{3.2}$$

then we have

$$p \le K_2 \eta^{\beta_2}, \tag{3.3}$$

where K_2 and β_2 are positive constants depending on c_0 and the a priori data only.

Theorem 3.1 is the first part of Theorem 4.1 in [12]. The proof of Proposition 3.2 follows exactly the same argument as that used to prove Proposition 4.9 in [12]. It appears clear that our aim is to improve the estimate (2.23)–(2.24) at least for points which are near to the point where the Hausdorff distance is reached.

The following geometric construction is crucial. Let us recall that G is the connected component of $\Omega\backslash(\Sigma \cup \Sigma')$ whose boundary contains $\partial\Omega$ and that whenever $p \le p_0$, we assume that the conclusions of Lemma 2.2 hold.

LEMMA 3.3. *Let Ω, Σ, and Σ' be as in Theorem* 2.3*. Then there exist positive constants p_1, $0 < p_1 \le p_0$, c_0, δ_0, C_4, and C_5, $0 < C_5 < 1$, depending on δ, M, and L only, such that if $p \le p_1$, then for every r, $0 \le r \le c_0 p$, there exists $z' \in \sigma_1' \cap \partial B_r(z_0')$ satisfying the following condition*

(a) *there exists a sequence of discs $D_n = B_{r_n}(z_n)$ such that, for any $n \in \mathbb{N}$,*

$$\begin{aligned} &\overline{D_n} \cap \overline{D_{n+1}} \ne \emptyset, \quad 2D_n = B_{2r_n}(z_n) \subset G, \\ &|z_n - z'| \le C_4 r_n, \quad r_{n+1} \le C_5 r_n, \end{aligned} \tag{3.4}$$

and, moreover,

$$\operatorname{dist}(D_1, \partial G) \ge \delta_0. \tag{3.5}$$

Let us remark that a point $z \in \Sigma \cup \Sigma'$ satisfies condition (a) provided there exists an open sector of a cone with vertex in z which is contained in G. In fact, in such a case, we can construct the discs satisfying condition (a) as follows. We take discs D_n

centered on the bisecting line of the sector so that $2D_n$ is contained in the sector and D_n is tangent to D_{n+1}, which is obviously chosen to be closer to z; see [1] for details. However, such a cone condition might not be satisfied if Σ and Σ' are only Lipschitz regular, even if they are very close in the Hausdorff distance; see Example 2.14 in [12]. Nevertheless, we show that, roughly speaking, in a neighborhood of the point where the Hausdorff distance is reached, condition (a) is satisfied, even if we might still lack the cone condition.

We begin with some preliminaries and by fixing some notation. Without loss of generality, by (2.8) and (2.9), we can restrict ourselves to the case of single cracks, σ_1 and σ_1', and take as G the unbounded connected component of $\mathbb{R}^2\backslash(\sigma_1 \cup \sigma_1')$.

Let us assume that $p \le \min\{p_0, \delta/4\}$ and that $z_0' \in \sigma_1'$ satisfies $\operatorname{dist}(z_0', \sigma_1) = p$, as in Lemma 2.2. Furthermore, let $z_0 \in \sigma_1$ be the point where this distance is reached, that is, such that $|z_0' - z_0| = p$.

Let us consider the coordinate system (x, y) with origin in z_0' such that with respect to these coordinates $\sigma_1 \cap B_\delta(z_0)$ is a *Lipschitz graph* with constant M and $z_0 = (x_0, y_0)$ with $x_0 \ge 0$ and $y_0 \le 0$. For any $i = 1, 2$, let $z_i = (x_i, y_i) \in \partial B_p(z_0')$ be such that $y_i \le 0$ and the tangent line to $\partial B_p(z_0')$ at the point z_i has slope $(-1)^i M$. Clearly we have $x_2 > 0$, $x_1 = -x_2$, and $y_1 = y_2$.

With the notation $S_r(\theta_1, \theta_2)$, where $r > 0$ and $\theta_1 < \theta_2$, we denote the open sector of a cone so defined:

$$S_r(\theta_1, \theta_2) = \{z = (x, y) : \ |z| < r \text{ and } \theta_1 < \arg z < \theta_2\}.$$

We say that r is the radius and $\theta_2 - \theta_1$ is the amplitude of such a sector. Let θ, $0 < \theta < \pi/2$, be the angle between a line of slope M and the y-axis and let

$$S_0 = S_{\delta/2}(\pi/2 - \theta, \pi/2 + \theta).$$

The proof of Lemma 3.3 is rather technical. Before entering into details, let us sketch its main features. Let us consider, for simplicity, the case in which $x_1 \le x_0 \le x_2$. Let us consider the function

$$f(x) = \begin{cases} -M(x - x_1) + y_1 & \text{if } x \le x_1, \\ -\sqrt{1 - x^2} & \text{if } x_1 \le x \le x_2, \\ M(x - x_2) + y_2 & \text{if } x \ge x_2. \end{cases}$$

We have that since $\sigma_1 \cap B_\delta(z_0)$ is a *Lipschitz graph* with constant M, then σ_1 must be, locally, below the graph of f. In particular, $S_0(z_i) = z_i + S_0$, for any $i = 0, 1, 2$, and $S_0(z_0') = z_0' + S_0$ do not contain points of σ_1. Moreover, by construction, any point z belonging to $S_0(z_0')$ has distance greater than p from the graph of f and, consequently, also from σ_1. Since p is the Hausdorff distance between σ_1 and σ_1', we infer that $S_0(z_0') \cap \sigma_1'$ is empty.

We have shown that there exists an open sector of a cone with vertex in z_0' which is contained in G, and thus condition (a) is satisfied for $r = 0$. For $r > 0$, we proceed as follows. Let $z \in \sigma_1' \cap \partial B_r(z_0')$, and let us consider the two opposite sectors with vertex in z which do not intersect σ_1', which exist by the Lipschitz character of σ_1'. Such sectors, at least for small r and near z, are contained in the epigraph of f, and thus do not intersect σ_1, too. Then the main idea is the following. We proceed from z along the bisecting line of one of these two sectors until we meet the bisecting line of one among the sectors $S_0(z_i)$, $i = 1, 2$, or $S_0(z_0')$. Then we turn and continue along this other bisecting line. This piecewise linear curve will be the direction along which we approach z with the sequence of balls contained in G that provides condition (a).

Proof of Lemma 3.3. Under the previous hypotheses and notation, we have two cases: Either

(i) the tangent line to $\partial B_p(z_0')$ at the point z_0 has slope m less than or equal to M; or

(ii) the tangent line to $\partial B_p(z_0')$ at the point z_0 has slope m greater than M (including the extreme case of a vertical tangent, when $z_0 = (p, 0)$).

If (i) holds, that is, when $x_1 \le x_0 \le x_2$, then $S_0(z_0') = z_0' + S_0$ satisfies $\operatorname{dist}(S_0(z_0'), \sigma_1) \ge p$. Therefore, since $p = \mathrm{d}_H(\sigma_1 \cup \sigma_1')$, we have $S_0(z_0') \cap (\sigma_1 \cup \sigma_1') = \emptyset$.

On the other hand, if (ii) holds, then we pick $\tilde{S}_0(z_0')$ as $\tilde{S}_0(z_0') = z_0' + \tilde{S}_0 = z_0' + S_{\delta/2}(\pi/2, 3\pi/2 - \theta)$, and we still have that $\tilde{S}_0(z_0') \cap (\sigma_1 \cup \sigma_1') = \emptyset$. We wish to remark that (ii) can hold only if z_0 is an endpoint of σ_1.

The construction of $S_0(z_0')$ or $\tilde{S}_0(z_0')$, respectively, already proves that the lemma is true for $r = 0$, that is, for z_0'.

We have that, with respect to a coordinate system which is rotated in the counterclockwise sense of an angle $\tilde{\theta}$, $0 \le \tilde{\theta} < \pi$, with respect to the system (x, y), $\sigma_1' \cap B_\delta(z_0')$ is a Lipschitz graph with constant M, which implies that for any $z' \in \sigma_1' \cap \overline{B}_p(z_0')$, the sectors $S_0'(z')^{\pm} = z' \pm S_{\delta/2}(\pi/2 - \theta + \tilde{\theta}, \pi/2 + \theta + \tilde{\theta})$ do not intersect σ_1'. Furthermore, $S_0(z') = z' + S_0$ does not contain points of σ_1.

If $\tilde{\theta} \in [0, 15\theta/8] \cup [\pi - 15\theta/8, \pi)$, then for any $z' \in \sigma_1' \cap \overline{B}_p(z_0')$, we have that the intersection of $S_0(z')$ either with $S_0'(z')^+$ or with $S_0'(z')^-$ contains a sector of a cone of radius $\delta/2$ and amplitude at least $\theta/8$. Such a sector has empty intersection with $\sigma_1 \cup \sigma_1'$; thus a uniform cone property holds and the lemma easily follows.

We now consider the case in which $\tilde{\theta} \in (15\theta/8, \pi - 15\theta/8)$. We restrict ourselves to the case in which (i) holds; the case in which (ii) holds can be treated in a completely analogous way.

Let us consider the ball $B_{\delta/2}(z_0')$. We have that $B_{\delta/2}(z_0') \setminus (S_0'(z_0')^+ \cup S_0'(z_0')^-)$ consists of two closed sectors $\tilde{S}^+$ and $\tilde{S}^-$, where the first one is the only one whose intersection with $S_0(z_0')$ is not empty. If we further subtract $S_0(z_0')$, then we obtain at most three closed sectors, $\tilde{S}^-$, $\hat{S}_1$, and $\hat{S}_2$, the last ones being contained in $\tilde{S}^+$. We order $\hat{S}_1$ and $\hat{S}_2$ in the counterclockwise direction; that is, we take $\hat{S}_1$ as the one contained in $\{x \ge 0\}$ and $\hat{S}_2$ as the one contained in $\{x \le 0\}$, keeping in mind that one or both of them can be empty or have an empty interior. We observe that at most one between $\hat{S}_1$ and $\hat{S}_2$ contains points of σ_1'.

Assume that $\sigma_1' \cap \hat{S}_1$ is not empty. Then there exist constants c_1, $0 < c_1 < 1$, and $c_2 > 0$, depending on δ and M only, such that for every $z' \in \sigma_1' \cap \hat{S}_1 \cap \overline{B}_{c_1 p}(z_0')$ we have that $S_0'(z')^+ \cap S_0(z_0')$ is not empty, $S_0'(z')^+ \cap S_0(z_0') \subset z_0' + (S_{\delta/4}(\pi/2 - \theta, \pi/2 + \theta) \setminus S_{c_2|z' - z_0'|}(\pi/2 - \theta, \pi/2 + \theta))$, and the angle between the bisecting lines of $S_0'(z')^+$ and $S_0(z_0')$ is greater than a positive constant depending on M only. Then we can prove that for every $z' \in \sigma_1' \cap \hat{S}_1 \cap \overline{B}_{c_1 p}(z_0')$, condition (a) holds. We take a sequence of discs in $S_0(z_0')$, each one so that its center is on the bisecting line of $S_0(z_0')$, it is tangential to the next one, and the disc with double radius and same center is still contained in $S_0(z_0')$, till we reach the intersection of the bisecting lines of $S_0'(z')^+$ and $S_0(z_0')$. From that point on, we continue the construction by taking discs, with analogous properties as before, along the sector $S_0'(z')^+$.

If $\sigma_1' \cap \hat{S}_2$ is not empty, then we can repeat the same reasoning using $S_0'(z')^-$ instead of $S_0'(z')^+$.

It might happen that $\sigma_1' \cap (\hat{S}_1 \cup \hat{S}_2)$ is strictly contained in $B_{c_1 p}(z_0')$, and therefore the proof is not yet concluded. In this case, we can find positive constants p_1, $0 < p_1 \le \min\{p_0, \delta/4\}$, c_3, c_4, and θ_1, $0 < \theta_1 \le \theta$, depending on δ and M only, such that

if $p \le p_1$, then for any r, $0 < r \le c_3 p$, there exists $z' \in \sigma_1' \cap \tilde{S}^-$ such that $|z' - z_0'| = r$ and the following holds. Let $\tilde{S}_0'(z')^{\pm}$ be the sector with vertex in z', radius $\delta/2$, the same bisecting line as $S_0'(z')^{\pm}$, and amplitude $2\theta_1$. Then either (i) $\tilde{S}_0'(z')^+ \cap S_0(z_1)$ is not empty, $\tilde{S}_0'(z')^+ \cap S_0(z_1) \subset z_1 + (S_{\delta/4}(\pi/2-\theta, \pi/2+\theta) \backslash S_{c_4 p}(\pi/2-\theta, \pi/2+\theta))$, the angle between the bisecting lines of $\tilde{S}_0'(z')^+$ and $S_0(z_1)$ is greater than a positive constant depending on M only, and $S_0(z_1) \backslash (z_1 + S_{c_4 p}(\pi/2-\theta, \pi/2+\theta))$ does not contain points of σ_1'; or (ii) the same properties are satisfied by $\tilde{S}_0'(z')^-$ and $S_0(z_2)$. Then we repeat the construction used before using either the two sectors $\tilde{S}_0'(z')^+$ and $S_0(z_1)$ or $\tilde{S}_0'(z')^-$ and $S_0(z_2)$. ∎

We can now conclude the proof of our stability result.

Proof of Theorem 2.3. By Theorem 3.1, we can assume without loss of generality that $p \le p_1$. Then, by Lemma 3.3, we can find c_0, δ_0, C_4, and C_5, $0 < C_5 < 1$, depending on δ, M, and L only, such that for every r, $0 \le r \le c_0 p$, there exists $z' \in \sigma_1' \cap \partial B_r(z_0')$ satisfying condition (a).

Then, for any $i = 1, 2$, and any of these z' satisfying condition (a), we have

$$|v_i(z') - v_i'(z')| \le C_6 |\log \varepsilon|^{-\alpha_3}, \tag{3.6}$$

where C_6 and $\alpha_3 > 0$ depend on the a priori data only.

In fact, let us fix $i \in \{1, 2\}$ and let us call $f = u + \mathrm{i}v = u_i - u_i' + \mathrm{i}(v_i - v_i')$. We have that f is quasiregular inside $\Omega \backslash (\Sigma \cup \Sigma')$; that is, it satisfies a Beltrami-type equation like (2.16)–(2.17).

Let G_{δ_0} be the set of points in G whose distance from $\Sigma \cup \Sigma'$ is greater than or equal to δ_0. We assume, without loss of generality, that $\delta_0 \le \delta/4$; thus a neighborhood of $\partial\Omega$ in G is contained in G_{δ_0}.

Let $\Omega_1 = G_{\delta_0} \cup (\bigcup_{n \in \mathbb{N}} 2D_n)$, with D_n as in condition (a) applied to z'. We have that Ω_1 is a domain contained in G such that $\partial\Omega \subset \partial\Omega_1$.

Let us observe that for any r, $0 < r \le |z_1 - z'|$, with z_1 the center of D_1, there exists $w_r \in \bigcup_{n \in \mathbb{N}} \overline{D_n}$ such that $|w_r - z'| = r$. Furthermore, we can take such a w_r in D_n, where n satisfies $n < C_7(1 + |\log r|)$, with C_7 depending on C_4, C_5, and L only.

Then (3.6) can be obtained as follows. By recalling (2.21) and (2.22), we can estimate, in terms of ε, $|f|$ inside Ω_1 by using the method of harmonic measure, which has been generalized to operators with nonconstant and anisotropic coefficients in [3].

We can estimate $|v(z')|$ using the interior estimate of $|f|$ at the point w_r, $0 < r \le |z_1 - z|$, and (2.19). A precise estimate of $|f(w_r)|$ is obtained through a repeated use of the Harnack inequality along the sequence of discs D_n. We refer to the proof of Proposition 4.12 in [12] for details.

Then the conclusion follows immediately from (3.6) and Proposition 3.2. ∎

REFERENCES

[1] G. Alessandrini, *Stability for the crack determination problem*, in Inverse Problems in Mathematical Physics, L. Päivärinta and E. Somersalo, eds., Springer-Verlag, Berlin, Heidelberg, 1993, pp. 1–8.

[2] G. Alessandrini, E. Beretta, E. Rosset, and S. Vessella, *Optimal stability for inverse elliptic boundary value problem with unknown boundaries*, Ann. Scuola Norm. Sup. Pisa Cl. Sci., 29 (2000), pp. 755–806.

[3] G. Alessandrini and L. Rondi, *Stable determination of a crack in a planar inhomogeneous conductor*, SIAM J. Math. Anal., 30 (1998), pp. 326–340.

[4] G. Alessandrini and L. Rondi, *Optimal stability for the inverse problem of multiple cavities*, J. Differential Equations, 176 (2001), pp. 356–386.

[5] K. Bryan and M. S. Vogelius, *A review of selected works on crack identification*, in Geometric Methods in Inverse Problems and PDE Control, C. B. Croke, I. Lasiecka, G. Uhlmann, and M. S. Vogelius, eds., Springer-Verlag, New York, 2004, pp. 25–46.

[6] M. Di Cristo and L. Rondi, *Examples of exponential instability for inverse inclusion and scattering problems*, Inverse Problems, 19 (2003), pp. 685–701.

[7] M. Di Cristo and L. Rondi, *Examples of exponential instability for elliptic inverse problems*, preprint arXiv:math.AP/0303126, 2003; available online from http://arxiv.org/archive/math/.

[8] A. Friedman and M. Vogelius, *Determining cracks by boundary measurements*, Indiana Univ. Math. J., 38 (1989), pp. 527–556.

[9] D. S. Jerison and C. E. Kenig, *Boundary behavior of harmonic functions in non-tangentially accessible domains*, Adv. in Math., 46 (1982), pp. 80–147.

[10] N. Mandache, *Exponential instability in an inverse problem for the Schrödinger equation*, Inverse Problems, 17 (2001), pp. 1435–1444.

[11] L. Rondi, *Optimal stability estimates for the determination of defects by electrostatic measurements*, Inverse Problems, 15 (1999), pp. 1193–1212.

[12] L. Rondi, *Uniqueness and Optimal Stability for the Determination of Multiple Defects by Electrostatic Measurements*, Ph.D. thesis, S.I.S.S.A.-I.S.A.S., Trieste, 1999; available online from http://www.sissa.it/library/.

SIAM J. MATH. ANAL.
Vol. 36, No. 4, pp. 1293–1309

CONSERVATION LAWS WITH TIME DEPENDENT DISCONTINUOUS COEFFICIENTS*

GIUSEPPE MARIA COCLITE† AND NILS HENRIK RISEBRO‡

Abstract. We consider scalar conservation laws where the flux function depends discontinuously on both the spatial and temporal locations. Our main results are the existence and well-posedness of an entropy solution to the Cauchy problem. The existence is established by showing that a sequence of front tracking approximations is compact in L^1, and that the limits are entropy solutions. Then, using the definition of an entropy solution taken from [K. H. Karlsen, N. H. Risebro, and J. D. Towers, *Skr. K. Nor. Vidensk. Selsk.*, 3 (2003), pp. 1–49], we show that the solution operator is L^1 contractive. These results generalize the corresponding results from [S. N. Kružkov, *Math. USSR-Sb.*, 10 (1970), pp. 217–243] and also partially those from Karlsen, Risebro, and Towers.

Key words. conservation laws, entropy solutions, discontinuous coefficients

AMS subject classifications. 35L65, 65M25

DOI. 10.1137/S0036141002420005

1. Introduction.

In this paper we are concerned with the Cauchy problem for scalar conservation laws where the flux function depends on both the x and t coordinates. We study the case where this dependence takes the form $f(u,x,t) = f(u,a(x),g(t))$ through some functions a and g. Hence, we shall study the initial value problem

$$\begin{cases} u_t + f(u,a(x),g(t))_x = 0, & x \in \mathbf{R}, \quad t > 0, \\ u(x,0) = 0, & x \in \mathbf{R}, \end{cases} \tag{1.1}$$

where $f = f(u,a,g)$ is a smooth function. We regard the function $a(x)$ and $g(t)$ as coefficients, and if these are smooth, the classical results of Kružkov [16] and Oleĭnik [19] state that the above initial value problem is well posed in the class of entropy solutions.

In our case, the coefficients are allowed to be discontinuous, and we cannot apply the techniques of Kružkov and Oleĭnik directly to reach their conclusion. The main obstacle is that of the discontinuity of the spatial coefficient a. The equation where g is constant has recently received considerable attention, beginning with the paper of Temple [23], in which he studied a system of nonstrictly hyperbolic conservation laws. By a Lagrangian transformation, this system is equivalent to a scalar equation with discontinuous coefficients; see Wagner [26]. If one writes the scalar conservation law as a system by introducing a as a new component of the solution, we have

$$\begin{cases} u_t + f(u,a,g(t))_x = 0, \\ a_t = 0. \end{cases}$$

*Received by the editors December 18, 2002; accepted for publication (in revised form) April 2, 2004; published electronically February 3, 2005. The research was funded in part by the BeMatA program of the Research Council of Norway, by the European network HYKE, and by the EC as contract HPRN-CT-2002-00282.

http://www.siam.org/journals/sima/36-4/42000.html

†Centre of Mathematics for Applications, P.O. Box 1053, N-0316 Oslo, Norway (g.m.coclite@cma.uio.no).

‡Department of Mathematics, University of Oslo, P.O. Box 1053, Blindern, N-0316 Oslo, Norway (nilshr@math.uio.no).

This system has eigenvalues f_u and 0, and if $f_u(u,a,g)=0$ for some (u,a,g), then the system is nonstrictly hyperbolic, and the standard theory for systems (see Glimm [7] and, more recently, [8, 1]) does not apply. In particular, one can show by a concrete example (see, e.g., [23]) that the total variation of the approximate solutions produced by the Glimm scheme (and also by front tracking) is not bounded in terms of the discretization parameters. Such systems are commonly called *resonant*. For resonant systems, one cannot show compactness by the usual method of establishing BV estimates on a sequence of approximate solutions.

To overcome this difficulty, in [23] Temple introduced a nonlinear mapping $\Psi = \Psi(u)$ and used this mapping to prove that the sequence of approximations produced by the Glimm scheme is compact. This approach has since been used in a number of papers for related systems, using other approximations; see Gimse and Risebro [6] and Klingenberg and Risebro [14, 15] for front tracking approximations; Lin, Temple, and Wang [18] for Godunov-type approximations; Towers [24, 25] for monotone difference schemes; and Hong [9] for Godunov schemes for resonant $n \times n$ and for 2×2 systems with inhomogeneous source terms.

As an alternative to the use of Ψ to prove compactness, in [12, 10] Karlsen et al. used the Murat–Tartar compensated compactness approach to prove convergence of numerical approximations.

The conservation law (1.1) is formally equivalent to the Hamilton–Jacobi equation

$$v_t + f(v_x, a(x), g(t)) = 0.$$

In several papers, Ostrov (see, e.g., [20] and [21]) considered this type of equation, and he showed the existence of a viscosity-type solution and discussed the question of uniqueness.

Regarding uniqueness of weak solutions to (1.1) in the case where a and g are not smooth, this was first studied (for the constant g case) in [15] and [13]. In these papers it was shown that the solution is unique if it is the limit of solutions to equations where the coefficients are smoothed. More recently, L^1-contractivity was shown for piecewise smooth solutions in the case of convex flux functions in [25], and in a more general case by Karlsen, Risebro, and Towers [12]. Also, Seguin and Vovelle [22] proved uniqueness for L^∞ solutions for a special case of (1.1) with $g = \text{const.}$ and $a(\cdot)$ taking two values separated by a jump discontinuity. The techniques used in the present paper are heavily inspired by those used in [11], in which Karlsen, Risebro, and Towers show uniqueness of solutions in the case where g is constant, and where $u \mapsto f(u,a)$ is not required to have a single local maximum. The authors of [25, 12, 22, 11] all use a Kružkov-type entropy condition.

The purpose of the present paper is to extend the well-posedness theory for conservation laws with discontinuous coefficients by including a t dependent coefficient.

Conservation laws with discontinuous coefficients, both in x and t, occur in many models. The simplest such model is the hydrodynamic traffic flow model; see Lighthill and Whitham [17]. In this case the x and t dependency model the road conditions, specifically the maximal speed of any vehicle. Both of these dependencies can vary discontinuously—for instance, when modeling a traffic light. Another model in which such conservation laws occur is a clarifier-thickener model of continuous sedimentation; see Bürger et al. [4, 2, 3]. In the papers [2, 3] the actual models were simplified so that $g(t)$ was assumed to be constant.

Now we briefly state our main result and detail our assumptions. In order for the Riemann problem to have a bounded solution, it is convenient to assume that there is

a finite interval $[\alpha,\beta]$ such that $f(\alpha,a,g)=f(\beta,a,g)$ for all a and g, and we can choose $\alpha=0$ and $\beta=1$. This is not necessary for the solution of the Riemann problem to be bounded, but it is certainly sufficient; see [5], however, for less restrictive assumptions that yields the same conclusions.

So therefore we assume that $f:[0,1]\times\mathbf{R}^2\mapsto\mathbf{R}$, $g:\mathbf{R}^+\mapsto\mathbf{R}$ and $a:\mathbf{R}\mapsto\mathbf{R}$ are given functions which satisfy the following:

(A.1) a is piecewise C^1 with finitely many jump discontinuities at $x=x_1,\dots,x_M$.

(A.2) $\|a\|_{L^\infty}<\infty$, $\sup_{x\notin\{x_i\}_1^M}|a'(x)|<\infty$, and $a\in BV(\mathbf{R})$.

(A.3) $f\in C^2([0,1]\times\mathbf{R}^2;\mathbf{R})$, $f_{uu}(u,a,g)\le -c_{uu}<0$ for some positive constant c_{uu} for all a and g.

(A.4) $f(0,\cdot,\cdot)\equiv f(1,\cdot,\cdot)\equiv 0$, and there is a unique value u^* such that $f_u(u^*,\cdot,\cdot)\equiv 0$.

(A.5) $\partial f/\partial g\ge 0$ and $\partial f/\partial a\ge 0$. Furthermore

$$\frac{\partial^2 f}{\partial g\partial a}$$

is bounded.

(A.6) $g\in BV(\mathbf{R}^+)$, and $g(t)>0$ for all $t>0$.

Next, let $\Psi(u,a,g)$ be defined by

$$\Psi(u,g,a)=\operatorname{sign}(u-u^*)\,\frac{f(u^*,a,g)-f(u,a,g)}{f(u^*,a,g)}. \tag{1.2}$$

We demand that the initial data are such that $u_0\in L^1(\mathbf{R};[0,1])$ and

$$|\Psi(u_0,a,g)|_{BV}<\infty. \tag{1.3}$$

We use the following definition of a weak entropy solution of (1.1).

DEFINITION 1.1. *Let $T>0$, and let $u:\Pi_T=\langle 0,T\rangle\times\mathbf{R}\mapsto[0,1]$ be a measurable function. We call u an entropy weak solution of* (1.1) *if the following conditions hold:*

(D.1) *$u\in L^1(\Pi_T)$, and the map $\langle 0,T\rangle\ni t\mapsto u(\cdot,t)\in L^1(\mathbf{R})$ is Lipschitz continuous.*

(D.2) *The following entropy inequality holds for all constants c and all nonnegative test functions φ,*

$$\begin{aligned}\iint_{\Pi_T}&|u-c|\varphi_t+F(u,x,t,c)\varphi_x\,dtdx\\ &-\sum_{m=0}^{M}\int_{x_m}^{x_{m+1}}\int_0^T \operatorname{sign}(u-c)\,f_a(c,a(x),g(t))a'(x)\varphi\,dtdx\\ &+\sum_{m=1}^{M}\int_0^T\left|f(c,a\left(x_m^+\right),g(t))-f(c,a\left(x_m^-\right),g(t))\right|\,dt\ge 0,\end{aligned} \tag{1.4}$$

where we have set $x_0=-\infty$, $x_{M+1}=\infty$ and F is given by

$$F(u,x,t,c)=\operatorname{sign}(u-c)\left[f(u,a(x),g(t))-f(c,a(x),g(t))\right],\quad t>0,\quad x\in\mathbf{R}.$$

(D.3) *$u(\cdot,t)\to u_0$ in $L^1(\mathbf{R})$ as $t\downarrow 0$.*

(D.4) *$|\Psi(u(\cdot,t),a,g(t))|_{BV}<\infty$ for all $t\in\langle 0,T\rangle$.*

The inequality (1.4) implies that any entropy solution is a weak solution, as setting $c = 1$ and $c = 0$ will show. The condition (D.4) implies that the limits

$$\lim_{x \to x_m^{\pm}} \Psi(u, a, g)$$

exist for almost all t. Since $u \mapsto \Psi(u, a, g)$ is invertible, and the inverse is continuous, the limits

$$\lim_{x \to x_m^{\pm}} u(x, t)$$

also exist for almost all t. This will be needed to show uniqueness. Our main result is the following.

MAIN THEOREM. *Assume that f, a, and g satisfy the above assumptions,* (A.1)–(A.6). *If u_0 and v_0 are two functions that satisfy* (1.3), *then there exist corresponding entropy solutions u and v taking initial values u_0 and v_0, respectively. These entropy solutions satisfy*

$$\|u(\cdot, t) - v(\cdot, t)\|_{L^1(\mathbf{R})} \leq \|u_0 - v_0\|_{L^1(\mathbf{R})}.$$

The rest of this paper is organized as follows. In the next section, section 2, we define a sequence of approximate solutions by the front tracking method. This is based on the front tracking method defined in [14]. In section 3 we proceed to establish interaction estimates, which allows us to deduce that the total variation of Ψ is bounded for the front tracking approximations. Then we can use Helly's theorem and show that any limit is an entropy solution in the above sense. Then, using an adaptation of arguments taken from [11], one can show that the entropy solution operator is L^1 contractive. In this way our main theorem is proved. Finally, we conclude with a section showing the front tracking scheme used on a concrete example.

2. The front tracking scheme. We start this section by defining a front tracking scheme for the case where $g(t) \equiv \text{const}$. This scheme is slightly different from the front tracking scheme defined for this case in, e.g., [14]. The reason for this difference is that our front tracking scheme also must work when g is not constant.

Therefore we first consider the initial value problem,

$$\begin{cases} u_t + f(u, a)_x = 0 & \text{for } x \in \mathbf{R},\ t > 0, \\ u(x, 0) = u_0(x) & \text{for } x \in \mathbf{R}, \end{cases} \tag{2.1}$$

where f and a are as described above. The Riemann problem for (2.1) is the initial value problem where

$$u_0(x) = \begin{cases} u_l, & x \leq 0, \\ u_r, & x > 0, \end{cases} \qquad a(x) = \begin{cases} a_l, & x \leq 0, \\ a_r, & x > 0, \end{cases}$$

and its solution is detailed in [14]. This solution consists of at most one u-wave separating the u-values u_l and u_l', followed by a so-called a-wave separating the states (u_l', a_l) and (u_r', a_r). This wave is a contact discontinuity having zero speed. The solution is then completed by a u-wave separating u_r' and u_r. The first u-wave has nonpositive speed, and the second nonnegative. The intermediate states u_l' and u_r' are unique, provided (1.4) holds. Furthermore $u_{l,r}'$ can equal $u_{l,r}$.

Let

$$z(u,a) = \operatorname{sign}(u^* - u)\,(f(u,a) - f(u^*,a)) \quad \text{and} \quad \alpha(a) = f(u^*,a).$$

Since $a \mapsto f(u^*,a)$ is nondecreasing, $a \mapsto \alpha(a)$ is invertible. In the (z,α) plane, a-waves are straight lines of slope ± 1. An a-wave connecting two points (z_1,α_1) and (z_2,α_2) has slope 1 if z_1 and z_2 are nonpositive, and slope -1 if these values are nonnegative. If z_1 and z_2 have different sign, there is no a-wave connecting these points. Since u-waves connect points with the same a-values, these are horizontal lines in the (z,α) plane. Now fix a (small) number $\delta > 0$, and set $\alpha_i = i\delta$ and $z_j = j\delta$ for integers i and j. We define u_0^δ and a^δ as piecewise constant functions, with a finite number of jump discontinuities, such that

$$\left.\begin{aligned} \left\|a - a^\delta\right\|_{L^1(\mathbf{R})} &\to 0 \\ \left\|u_0 - u_0^\delta\right\|_{L^1(\mathbf{R})} &\to 0 \end{aligned}\right\} \quad \text{as } \delta \to 0. \tag{2.2}$$

Label the (finite number of) values of u^δ and a^δ $u_1,\dots,u_M$, and $a_1,\dots,a_N$, respectively. Let α_j be the jth member of the ordered set

$$\{\alpha_k\}_{k=m'}^{M'} \cup \{\alpha(a_k)\}_{k=1}^{M},$$

where m' and M' are chosen such that

$$0 < m' \le \min_x \alpha(a^\delta(x)) < \max_x \alpha(a^\delta(x)) \le M'.$$

For ease of notation, set

$$a_j = \alpha^{-1}(\alpha_j).$$

Next, for each α_j, we define $z_{j,k}$ to be the kth member of the ordered set

$$\{z_i\}_{i=-N'(j)}^{N'(j)} \cup \{z(u_i,a_j)\}_{i=1}^{M},$$

where $N'(j)$ is such that

$$z^{-1}\left(z_{-N'(j)},a_j\right) = 0 \quad \text{and} \quad z^{-1}\left(z_{N'(j)},a_j\right) = 1.$$

We also set

$$u_{j,k} = z^{-1}(z_{j,k},a_j) \quad \text{and} \quad f_{j,k} = f(u_{j,k},a_j).$$

Then, for each j, let the approximate flux function $f^\delta(u,a)$ be the piecewise linear interpolant,

$$f^\delta(u,a_j) = f_{j,k} + (u - u_{j,k})\,\frac{f_{j,k+1} - f_{j,k}}{u_{j,k+1} - u_{j,k}} \quad \text{for } u \in [u_{j,k},u_{j,k+1}]. \tag{2.3}$$

We have chosen the grid so that the entropy solution to the initial value problem

$$\begin{aligned} u_t + f^\delta\left(u,a^\delta\right)_x &= 0, \quad t > 0, \quad x \in \mathbf{R}, \\ u(x,0) &= u_0^\delta(x), \quad x \in \mathbf{R}, \end{aligned} \tag{2.4}$$

can be constructed by front tracking for any time t. We call this front tracking solution u^δ. Furthermore u^δ will take values that are grid points; i.e., for any point (x,t) such that u^δ and a^δ are constant at (x,t),

$$z\left(u^\delta(x,t), a^\delta(x)\right) = z_{j,k} \quad \text{for some } j \text{ and } k.$$

In particular, this means that

$$f^\delta\left(u^\delta, a^\delta\right) = f\left(u^\delta, a^\delta\right) \quad \text{almost everywhere.}$$

For an elaboration and proof of these statements, see [14]. The construction used here differs from the construction in [14] in that we have added grid points corresponding to the discretization of the initial function u_0 and the coefficient a, instead of choosing discretization that takes values on the fixed grid in the (z,α) plane.

Now we can define the front tracking approximation in the case where g is not constant; cf. (1.1). Let g^δ be a piecewise constant approximation to g, such that

$$\begin{aligned} \left\|g^\delta - g\right\|_{L^1(\mathbf{R}^+)} &\to 0 \quad \text{as } \delta \to 0, \\ \left|g^\delta\right|_{BV(\langle 0,T])} &\le |g|_{BV(\langle 0,T])}. \end{aligned} \tag{2.5}$$

Define t^n such that g^δ is constant on each interval $I^n = \langle t^n, t^{n+1}]$. Assuming that we can define front tracking for $t < t^n$, we can then use $u^\delta(\cdot, t^n)$ as initial values for a front tracking approximation defined in $[t^n, t^{n+1}\rangle$. In order to do this we must use a "new" mapping z, since $z = z(u,a,g)$, and redefine the grid on which we operate. However, we keep the grid points corresponding to $u^\delta(\cdot, t^n)$. In this way, the grid used in the interval I^{n+1} will contain more points than the one used in I^n, but since there are only a finite number of intervals I^n such that $t^n \le T$, for a fixed δ, we use a finite number of grid points for $t \le T$. If, for $t \in I^n$, $f^\delta(\cdot,\cdot,g^\delta(t))$ denotes the approximate flux function constructed above using $f(\cdot,\cdot,g^\delta\,\big|_{I^n})$ and $u^\delta(\cdot,t^n)$, then we have that the front tracking construction u^δ will be an entropy solution of

$$\begin{aligned} u^\delta_t + f^\delta\left(u^\delta, a^\delta(x), g^\delta(t)\right)_x &= 0, \quad t > 0, \quad x \in \mathbf{R}, \\ u^\delta(x,0) &= u^\delta_0(x), \quad x \in \mathbf{R}. \end{aligned} \tag{2.6}$$

We call the discontinuities in u^δ *fronts*, and we have three types: u-fronts, a-fronts, and g-fronts (that have infinite speed!).

3. Compactness. In this section we show that the sequence $\left\{u^\delta\right\}_{\delta>0}$ is compact in L^1 by estimating the variation of $\Psi(u^\delta, a^\delta, g^\delta)$. For each time t, such that g^δ is constant at t, we can view u^δ as consisting of a sequence of fronts, u-fronts and a-fronts.

We defined the map Ψ by (1.2), and we define the associated Temple functional of a front w by

$$T(w) = \begin{cases} |\Delta\Psi| & \text{if } w \text{ is a } u\text{-front}, \\ 2\,|\Delta f(u^*,a,g)| & \text{if } w \text{ is an } a\text{-front, and } \Psi_r < \Psi_l, \\ 4\,|\Delta f(u^*,a,g)| & \text{if } w \text{ is an } a\text{-front, and } \Psi_r > \Psi_l. \end{cases} \tag{3.1}$$

If $\Psi_l < \Psi_r$, we call an a-front *counterclockwise*; otherwise we call it *clockwise*. For sequence of fronts, define T additively. Next, for the front tracking approximation u^δ, we define the interaction estimate Q by

$$Q(t) = T(t)\left|g^\delta(\cdot)\right|_{BV([t,T])}, \tag{3.2}$$

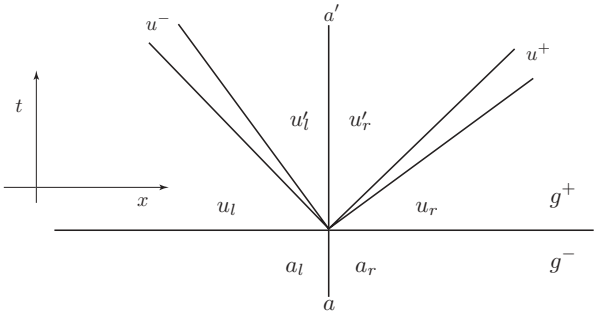

FIG. 1. *The states used in an interaction between an a-wave and a g-wave.*

where with a slight abuse of notation we write $T(t) = T(u^\delta(\cdot, t))$. With these definitions, we can state the following lemma.

LEMMA 3.1. *There exists a positive constant C, depending only on f, a, and g, such that for all $t > 0$, we have that the "Glimm functional"*

$$G(t) = T(t) + CQ(t) \tag{3.3}$$

is nonincreasing in time.

Proof. In each interval I^n, we know from [14] that T is nonincreasing, and the lemma holds. To prove the lemma we must study interactions between u-fronts and g-fronts, and between a-fronts and g-fronts.

We start by considering the interaction between a single a-front and a single g-front. The states involved are depicted in Figure 1. We label the "incoming" a-wave (front) a, the outgoing a-wave a', the left-moving outgoing u-wave u^-, and the right-moving outgoing u-wave u^+. See Figure 1.

In this case we claim that

$$T(u^-) + T(a') + T(u^+) - T(a) \leq C\,|\Delta g|\,|\Delta a| \tag{3.4}$$

for some constant C depending on f and its derivatives, but not on δ. In this case it will follow from the discussion of cases below that if a is (counter)clockwise, then a' is (counter)clockwise. Next, from this it follows that

$$\begin{aligned}
|T(a') - T(a)| &\leq 4\left(\left|f\left(u^*, a_l, g^+\right) - f\left(u^*, a_r, g^+\right)\right| - \left|f\left(u^*, a_l, g^-\right) - f\left(u^*, a_r, g^-\right)\right|\right) \\
&\leq 4\left[\int_{a_l}^{a_r} \left|\frac{\partial f}{\partial a}\right| (u^*, z, g^+)\, dz - \int_{a_l}^{a_r} \left|\frac{\partial f}{\partial a}\right| (u^*, z, g^-)\, dz\right] \\
&\leq 4\int_{a_l}^{a_r}\int_{g^-}^{g^+} \left|\frac{\partial^2 f}{\partial g \partial a}\right| (u^*, z, y)\, dy dz \\
&\leq C\,|\Delta a \Delta g|
\end{aligned}$$

for some C depending on the partial derivatives of f. Therefore it suffices to show that

$$T(u^-) + T(u^+) \leq C\,|\Delta g|\,|\Delta a|\,. \tag{3.5}$$

First observe that since an a-wave cannot cross the line $z = 0$, either both u_l and u_r are less than or equal to u^* or both are greater than or equal to u^*. If this is not so,

then the "a-wave" is in fact a stationary u-wave followed by an a-wave, or vice versa. If this is so, we can perturb u^δ an arbitrarily small amount by shifting the stationary u-wave a small distance and then treat the interaction of the g-wave and the u-wave separately.

We now let

$$G\left(a_l, a_r, g^-, g^+\right) = T\left(u^-\right) + T\left(u^+\right).$$

For simplicity, we regard a_l and g^- as fixed, and the emerging waves as functions of $a = a_r$ and $g = g^+$. Trivially we have that

$$G(a_l, a_l, g^-, g) = G(a_l, a, g^-, g^-) = 0,$$

and (3.5) follows if G is continuous and

$$\frac{\partial^2 G}{\partial a \partial g}$$

is bounded, since

$$G(a_l, a_r, g^-, g^+) = \int_{a_l}^{a_r} \int_{g^-}^{g^+} \frac{\partial^2 G}{\partial a \partial g}(a_l, a, g^-, g)\, dg da.$$

First we assume that both u_l and u_r are less than or equal to u^*. In this case, if

$$f\left(u_l, a_l, g\right) \le f\left(u^*, a, g\right), \tag{3.6}$$

then there are no left-moving waves u^-, while if

$$f\left(u_l, a_l, g\right) > f\left(u^*, a, g\right), \tag{3.7}$$

there will be emerging u-waves of both positive and negative speeds. These two case are depicted in Figure 2. Note that in both cases both a and a' are counterclockwise. So we find that

$$\begin{aligned} G\left(a_l, a, g^-, g\right) = {} & \frac{\operatorname{sign}\left(u_l - u_r\right)}{f\left(u^*, a, g\right)} \left[f\left(u_l, a_l, g\right) - f\left(u_r, a, g\right)\right] \chi_{\{f(u_l,a_l,g) \le f(u^*,a,g)\}} \\ & + \left\{ \frac{1}{f\left(u^*, a_l, g\right)} \left[f\left(u_l, a_l, g\right) - f\left(u^*, a, g\right)\right] \right. \\ & \left. + \frac{1}{f\left(u^*, a, g\right)} \left[f\left(u^*, a, g\right) - f\left(u_r, a, g\right)\right] \right\} \chi_{\{f(u_l,a_l,g) > f(u^*,a,g)\}}. \end{aligned}$$

From this expression it is straightforward to check that G is sufficiently regular, and (3.5) holds.

The case where $u_{l,r} \ge u^*$ is similar: If

$$f\left(u_r, a, g\right) \le f\left(u^*, a_l, g\right), \tag{3.8}$$

there is only one outgoing u-wave, with negative speed. If

$$f\left(u_r, a, g\right) > f\left(u^*, a_l, g\right), \tag{3.9}$$

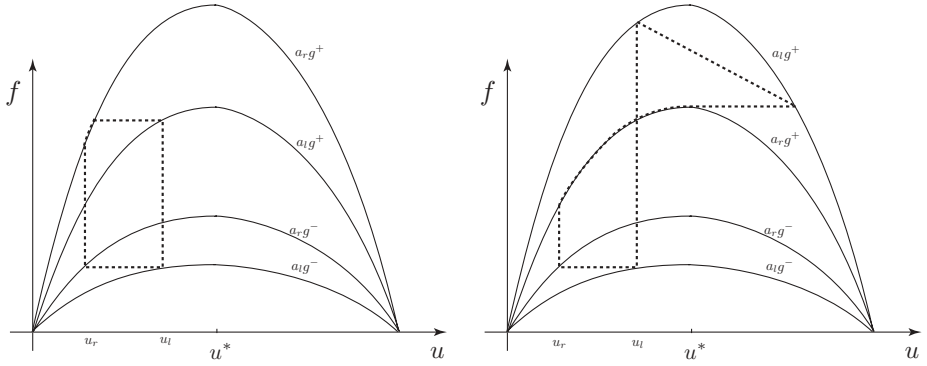

FIG. 2. *The possible results of an interaction if* $u_{l,r} \leq u^*$. *Left:* (3.6) *holds. Right:* (3.7) *holds.*

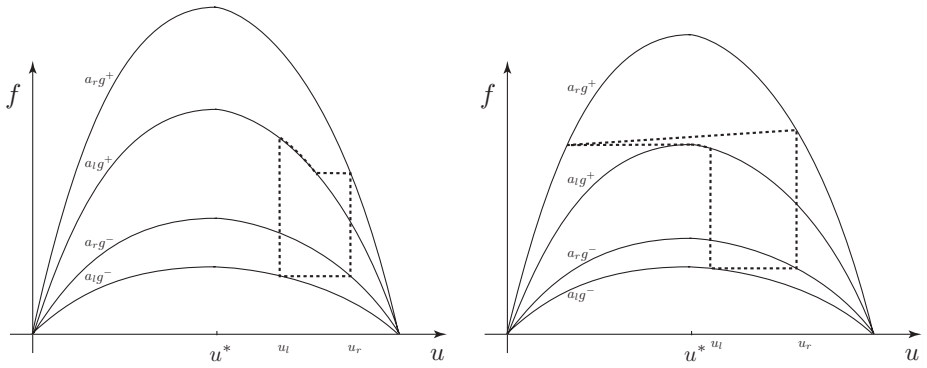

FIG. 3. *The possible results of an interaction if* $u_{l,r} \geq u^*$. *Left:* (3.8) *holds. Right:* (3.9) *holds.*

there are two outgoing u-waves. See Figure 3. Note that in both cases a and a' are clockwise. In this case we have

$$G\left(a_l, a, g^-, g\right) = \frac{\operatorname{sign}\left(u_l - u_r\right)}{f\left(u^*, a_l, g\right)} \left[f\left(u_l, a_l, g\right) - f\left(u_r, a, g\right)\right] \chi_{\{f(u_r,a,g) \leq f(u^*,a_l,g)\}}$$
$$+ \left\{ \frac{1}{f\left(u^*, a, g\right)} \left[f\left(u_r, a, g\right) - f\left(u^*, a_l, g\right)\right] \right.$$
$$\left. + \frac{1}{f\left(u^*, a_l, g\right)} \left[f\left(u^*, a_l, g\right) - f\left(u_l, a_l, g\right)\right] \right\} \chi_{\{f(u_r,a,g) > f(u^*,a_l,g)\}}.$$

Also in this case G is sufficiently regular for (3.5) to hold and thereby (3.4). This finishes the study of the interaction of a- and g-fronts

Now we consider the interaction of a single u-wave and a single g-wave. The situation is depicted in Figure 4. For this interaction we claim that

(3.10)
$$\left|\Psi\left(u_r, a, g^+\right) - \Psi\left(u_l, a, g^+\right)\right| - \left|\Psi\left(u_r, a, g^-\right) - \Psi\left(u_l, a, g^-\right)\right|$$
$$\leq C\left|g^+ - g^-\right| \left|\Psi\left(u_r, a, g^-\right) - \Psi\left(u_l, a, g^-\right)\right|.$$

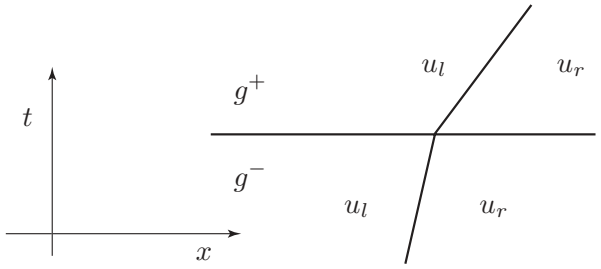

FIG. 4. *The states used in an interaction between a u-wave and a g-wave.*

Since $\Psi(u^*,\cdot,\cdot)=\Psi_u(u^*,\cdot,\cdot)=0$, we can write

$$\begin{aligned}\Psi\left(u_r,a,g^+\right)-\Psi\left(u_l,a,g^+\right)-\Psi\left(u_r,a,g^-\right)+\Psi\left(u_r,a,g^-\right)&\\ =\int_{u_l}^{u_r}\left(\Psi_u\left(\sigma,a,g^+\right)-\Psi_u\left(\sigma,a,g^-\right)\right)\,d\sigma&\\ =\int_{u_l}^{u_r}\int_{u^*}^{\sigma}\left(\Psi_{uu}\left(\eta,a,g^+\right)-\Psi_{uu}\left(\eta,a,g^-\right)\right)\,d\eta\,d\sigma&\\ =\int_{u_l}^{u_r}\int_{u^*}^{\sigma}\int_{g^-}^{g^+}\Psi_{uug}\left(\eta,a,g\right)\,dg\,d\eta\,d\sigma.&\end{aligned}\tag{3.11}$$

We also find that

$$\begin{aligned}\Psi\left(u_r,a,g^-\right)-\Psi\left(u_l,a,g^-\right)&=\int_{u_l}^{u_r}\Psi_u\left(\sigma,a,g^-\right)\,d\sigma\\ &=\int_{u_l}^{u_r}\left(\Psi_u\left(\sigma,a,g^-\right)-\Psi_u\left(u^*,a,g^-\right)\right)\,d\sigma\\ &=\int_{u_l}^{u_r}\int_{u^*}^{\sigma}\Psi_{uu}\left(\eta,a,g^-\right)\,d\eta\,d\sigma.\end{aligned}\tag{3.12}$$

We also have that

$$\Psi_{uu}(u,a,g)=\operatorname{sign}(u-u^*)\,\frac{-f_{uu}(u,a,g)}{f(u^*,a,g)}\geq\frac{c_{uu}}{C_{u^*}},$$
$$|\Psi_{uug}(u,a,g)|=\left|\frac{f_{uu}(u,a,g)f_g(u^*,a,g)-f_{uug}(u,a,g)f(u^*,a,g)}{f^2(u^*,a,g)}\right|\leq C_1$$

for some constant C_1. To fix ideas, assume that $u_l\leq u_r$, so that also

$$\Psi\left(u_l,a,g^\pm\right)\leq\Psi\left(u_r,a,g^\pm\right).$$

To show (3.10), we consider different cases.

Case 1. $u^*\leq u_l\leq u_r$. By (3.11) we have

$$\begin{aligned}\left|\Psi\left(u_r,a,g^+\right)-\Psi\left(u_l,a,g^+\right)\right|-\left|\Psi\left(u_r,a,g^-\right)-\Psi\left(u_l,a,g^-\right)\right|&\\ \leq C'\left|g^+-g^-\right|\int_{u_l}^{u_r}\int_{u^*}^{\sigma}d\eta\,d\sigma&\\ =C'\left|g^+-g^-\right|\left(\frac{\left|u_r^2-u_l^2\right|}{2}-u^*\left(u_r-u_l\right)\right)&\end{aligned}$$

for some constant C' depending on the partial derivatives of f. By (3.12) we also have that

$$\Psi\left(u_r, a, g^-\right) - \Psi\left(u_l, a, g^-\right) \geq \frac{c_{uu}}{C_{u^*}} \int_{u_l}^{u_r} \int_{u^*}^{\sigma} d\eta \, d\sigma,$$

so (3.10) follows with $C = C' C_{u^*}/c_{uu}$.

Case 2. $u_l \leq u_r \leq u^*$. In this case,

$$\begin{aligned} \left|\Psi\left(u_r, a, g^+\right) - \Psi\left(u_l, a, g^+\right)\right| - \left|\Psi\left(u_r, a, g^-\right) - \Psi\left(u_l, a, g^-\right)\right| & \\ \leq \hat{C}\left|g^+ - g^-\right| \int_{u_l}^{u_r} \int_{\sigma}^{u^*} d\eta \, d\sigma & \\ = \hat{C}\left|g^+ - g^-\right| \left[u^*\left(u_r - u_l\right) - \frac{\left|u_r^2 - u_l^2\right|}{2}\right] & \end{aligned}$$

for some constant $\hat{C}$, and also by (3.12)

$$\Psi\left(u_r, a, g^-\right) - \Psi\left(u_l, a, g^-\right) \geq \frac{c_{uu}}{C_{u^*}} \int_{u_l}^{u_r} \int_{\sigma}^{u^*} d\eta \, d\sigma.$$

Hence (3.10) follows as in the first case.

Case 3. $u_l \leq u^* \leq u_r$. Now we write

$$\begin{aligned} \left|\Psi\left(u_r, a, g^+\right) - \Psi\left(u_l, a, g^+\right)\right| - \left|\Psi\left(u_r, a, g^-\right) - \Psi\left(u_l, a, g^-\right)\right| & \\ \leq C_2\left|g^+ - g^-\right| \left[\int_{u_l}^{u^*} \int_{u^*}^{\sigma} d\eta \, d\sigma + \int_{u^*}^{u_r} \int_{\sigma}^{u^*} d\eta \, d\sigma\right] & \\ = C_2\left|g^+ - g^-\right| \left(\frac{u_r^2 - 2\left(u^*\right)^2 + u_l^2}{2} - u^*\left(u_r - 2u^* + u_l\right)\right). & \end{aligned}$$

Also, by (3.12) we can estimate

$$\Psi\left(u_r, a, g^-\right) - \Psi\left(u_l, a, g^-\right) \geq \frac{c_{uu}}{C_{u^*}} \left[\int_{u_l}^{u^*} \int_{u^*}^{\sigma} d\eta \, d\sigma + \int_{u^*}^{u_r} \int_{\sigma}^{u^*} d\eta \, d\sigma\right].$$

So again (3.10) follows.

If $u_l > u_r$, we can use the same arguments as in Cases 1 or 2 above to show (3.10). Since $T(t) \geq \left|a^\delta\right|_{BV}$, the lemma now follows. □

Let $T^n = T\,\big|_{I^n}$ and $g^n = g^\delta\,\big|_{I^n}$. Since T is nonincreasing in each interval I^n, from Lemma 3.1, we have that

$$T^{n+1} \leq T^n\left(1 + C\left|g^{n+1} - g^n\right|\right).$$

By the Grönwall inequality it follows that

$$\begin{aligned} T(t) &\leq T^1(0+) \exp\left(C \sum_n \left|g^n - g^{n-1}\right|\right) \\ &\leq \lim_{s \downarrow 0} T(s) \exp\left(C\left|g\right|_{BV}\right) \\ &\leq \left(\left|\Psi\left(u_0, a, g(0)\right)\right|_{BV} + 4\left|a\right|_{BV}\left|g(0)\right|\right) e^{C|g|_{BV}}, \end{aligned} \tag{3.13}$$

where the sum in the first line above is over those n such that $t_n < t$.

From this it immediately follows that the total variation of $\Psi(u^\delta, a^\delta, g^\delta(t))$ is bounded independently of δ and t. Furthermore

$$-1 \le \Psi\left(u^\delta(x,t), a^\delta(x), g^\delta(t)\right) \le 1.$$

By Helly's theorem, for each fixed $t \in [0,T]$,

$$\Psi\left(u^\delta(\cdot,t), a^\delta, g^\delta(t)\right) \to \psi \quad \text{almost everywhere as } \delta \downarrow 0,$$

and by the Lebesgue dominated convergence theorem also in $L^1(\mathbf{R})$. Furthermore, by a diagonal argument, we can achieve this convergence for a dense countable set $\{t^\gamma\} \subset [0,T]$. For t^γ in this set, define

$$u(\cdot,t^\gamma) = \Psi^{-1}\left(\psi, a, g\left(t^\gamma\right)\right).$$

Hence also $u^\delta(\cdot,t^\gamma) \to u(\cdot,t^\gamma)$. For any $t \in [0,T]$ we have that

$$\begin{aligned}\left\|u^{\delta_1}(\cdot,t) - u^{\delta_2}(\cdot,t)\right\|_{L^1(\mathbf{R})} &\le \left\|u^{\delta_1}(\cdot,t^\gamma) - u^{\delta_1}(\cdot,t)\right\|_{L^1(\mathbf{R})} \\ &\quad + \left\|u^{\delta_1}(\cdot,t^\gamma) - u^{\delta_2}(\cdot,t^\gamma)\right\|_{L^1(\mathbf{R})} + \left\|u^{\delta_2}(\cdot,t^\gamma) - u^{\delta_2}(\cdot,t)\right\|_{L^1(\mathbf{R})},\end{aligned}$$

where t^γ is such that $u^\delta(\cdot,t^\gamma) \to u(\cdot,t^\gamma)$. By Lemma 3.2, $t \mapsto u^\delta(\cdot,t)$ is L^1 Lipschitz continuous, so the first and third terms above can be made arbitrarily small by choosing δ_1 and δ_2 small, and the middle term can be made small by choosing t^γ close to t. Hence we have that u^δ converges to some function u in $L^1(\mathbf{R} \times [0,T])$. For the reader's convenience we show the following.

LEMMA 3.2. *There exists a positive constant C, independent of t, s, and δ, such that*

$$\left\|u^\delta(\cdot,t) - u^\delta(\cdot,s)\right\|_{L^1(\mathbf{R})} \le C\,|t-s|. \tag{3.14}$$

Proof. We start by noting that since

$$\begin{aligned}&\left|\Psi\left(u^\delta(x,t), a^\delta(x), g^\delta(t)\right) - \Psi\left(u^\delta(y,t), a^\delta(y), g^\delta(t)\right)\right| \\ &\qquad \ge \left|f\left(u^\delta(x,t), a^\delta(x), g^\delta(t)\right) - f\left(u^\delta(y,t), a^\delta(y), g^\delta(t)\right)\right|,\end{aligned}$$

it follows that the total variation of f is bounded by some constant C, and C is independent of t and δ. Next, assume that $0 \le s < t \le T$, and let α_h be a smooth approximation to the characteristic function of the interval $[s,t]$, so that

$$\alpha_h \to \chi_{[s,t]} \quad \text{and} \quad \alpha_h' \to \delta_s - \delta_t,$$

as $h \downarrow 0$, where δ_s denotes the Dirac delta function centered at s. Choose a test function $\varphi(x)$ such that $|\varphi| \le 1$, and set $\varphi_h(x,t) = \varphi(x)\alpha_h(t)$. Since u^δ is a weak solution, we have that

$$\iint_{\Pi_T} u^\delta \partial_t \varphi_h + f\left(u^\delta, a^\delta, g^\delta\right) \partial_x \varphi_h \, dtdx = 0,$$

and sending $h \downarrow 0$ we find that

$$\int_{\mathbf{R}} \varphi(x)\left(u^\delta(x,t) - u^\delta(x,s)\right) dx = \int_s^t \int_{\mathbf{R}} \varphi_x(x) f\left(u^\delta, a^\delta, g^\delta\right) dtdx.$$

Now

$$\begin{aligned}\left\|u^\delta(\cdot,t)-u^\delta(\cdot,s)\right\|_{L^1(\mathbf{R})} &= \sup_{|\varphi|\le 1}\int \varphi(x)\left(u^\delta(x,t)-u^\delta(x,s)\right)\,dx\\ &= \sup_{|\varphi|\le 1}\int_s^t\int_{\mathbf{R}} \varphi_x(x) f\left(u^\delta,a^\delta,g^\delta\right)\,dxd\sigma\\ &\le \int_s^t \left|f\left(u^\delta(\cdot,\sigma),a^\delta,g^\delta(\sigma)\right|_{BV}\,d\sigma\\ &\le (t-s)C. \qquad \square\end{aligned}$$

Next, we shall show that the limit u is an entropy solution. First we study how u^δ differs from an entropy solution in each interval $\langle x_m, x_{m+1}\rangle$. Assume that y_k are the discontinuity points of a^δ inside this interval, such that

$$x_m = y_0 < y_1 < \cdots < y_K = x_{m+1},$$

and we have that $a = a_k$ for $x \in \langle y_k, y_{k+1}\rangle$. Since u^δ is an entropy solution inside each interval $\langle y_k, y_{k+1}\rangle$,

(3.15)
$$\int_{y_k}^{y_{k+1}}\int_0^T \left|u^\delta - c\right|\varphi_t + F^\delta(u^\delta,x,t,c)\,dtdx - \int_0^T F^\delta(u^\delta,x,t,c)\Big|_{x=y_k^+}^{x=y_{k+1}^-}\,dt \ge 0,$$

where

$$F^\delta(u,x,t,c) = \operatorname{sign}(u-c)\left(f^\delta(u,a^\delta(x),g^\delta(t)) - f^\delta(c,a^\delta(x),g^\delta(t))\right).$$

If we set $y_k^{l,r} = y_k^{\mp}$, and observe that since u^δ is a weak solution,

$$f\left(u^\delta\left(t,y_k^l\right),a_k,g^\delta\right) = f\left(u^\delta\left(t,y_k^r\right),a_{k+1},g^\delta\right) =: f_k$$

for almost all t. Summing (3.15) for $k = 0,\dots,K-1$, we obtain

(3.16)
$$\int_{x_m}^{x_{m+1}}\int_0^T \left|u^\delta - c\right|\varphi_x + F^\delta\left(u^\delta,x,t,c\right)\varphi_x\,dtdx - \int_0^T \varphi F^\delta(u^\delta,x,t,c)\Big|_{x=x_m^+}^{x=x_{m+1}^-}\,dt$$

(3.17)
$$-\int_0^T\sum_{k=1}^{K-1}\varphi(x_k,t)\left[\operatorname{sign}(u_k^r-c)\left(f_k-f_k^r(c)\right)-\operatorname{sign}\left(u_k^l-c\right)\left(f_k-f_k^l(c)\right)\right]\,dt \ge 0,$$

where we have used the notation

$$u_k^{l,r} = u^\delta(y_k^{\mp},t), \quad f_k^{l,r}(c) = f^\delta(c,a^\delta(y_k^{\mp}),g^\delta).$$

Since at each discontinuity y_k, u^δ is the solution of a Riemann problem, either *both* u_k^l and u_k^r are less than or equal to u^* or both are greater than or equal to u^*. Using this we can label those discontinuities where both u-values are less than or equal to

u^* as L, and the remaining ones as G. Hence we can write the integrand in (3.17) as (for brevity we use a notation where $\varphi(x_k,t)$ is invisible; it will reappear later)

(3.18)
$$-\sum_L \operatorname{sign}\left(u_k^l-c\right)\left(f_k^l(c)-f_k^r(c)\right)+\left[\operatorname{sign}\left(u_k^r-c\right)-\operatorname{sign}\left(u_k^l-c\right)\right]\left(f_k-f_k^r(c)\right)$$

(3.19)
$$-\sum_G \operatorname{sign}\left(u_k^r-c\right)\left(f_k^l(c)-f_k^r(c)\right)+\left[\operatorname{sign}\left(u_k^l-c\right)-\operatorname{sign}\left(u_k^r-c\right)\right]\left(f_k^l(c)-f_k\right).$$

Since $f_u^\delta(u,\cdot,\cdot)>0$ for $u<u^*$, if $u_k^l\le c\le u_k^r$, the second term in (3.18) equals

$$2\left(f_k^r(u_k^r)-f_k^r(c)\right)\ge 0,$$

and if $u_r\le c\le u_l$, the second term equals

$$-2\left(f_k^r(u_k^r)-f_k^r(c)\right)\ge 0.$$

Similarly we find that the second term in (3.19) is always nonnegative. Hence

$$\begin{aligned}(3.17)&\le -\int_0^T\sum_{k=1}^{K-1}\operatorname{sign}\left(u_k^{l,r}-c\right)\left(f^\delta(c,a_{k+1},g^\delta)-f^\delta(c,a_k,g^\delta)\right)\,dt\\&=-\int_0^T\sum_{k=1}^{K-1}\operatorname{sign}\left(u_k^{l,r}-c\right)\frac{f^\delta(c,a_{k+1},g^\delta)-f^\delta(c,a_k,g^\delta)}{\Delta y_k}\,dt\Delta y_k,\end{aligned}$$

where $\Delta y_k=y_{k+1}-y_k$, and we use u_k^l for discontinuities in L, and u_k^r for discontinuities in G. Since a is continuously differentiable in $\langle x_m,x_{m+1}\rangle$ and $a^\delta\to a$, $g^\delta\to g$ and $u^\delta\to u$ as $\delta\downarrow 0$, we find that

$$\lim_{\delta\downarrow 0}(3.17)\ \le -\int_0^T\int_{x_m}^{x_{m+1}}\operatorname{sign}(u-c)\,f(c,a,g)_x\varphi\,dxdt. \tag{3.20}$$

For a rigorous proof of this inequality, see [11, Lemma 4.4]. By the same arguments, we find that for each discontinuity x_m

$$\begin{aligned}&F^\delta\left(u^\delta\left(x_m^+,t\right),x_m^+,t,c\right)-F^\delta\left(u^\delta\left(x_m^-,t\right),x_m^-,t,c\right)\\&\qquad=\operatorname{sign}\left(u_m^r-c\right)\left(f_m-f_m^r(c)\right)-\operatorname{sign}\left(u_m^l-c\right)\left(f_m-f_m^l(c)\right)\\&\qquad\ge\left|f_m^l(c)-f_m^r(c)\right|.\end{aligned}$$

Finally, adding (3.16)–(3.17) for m and using the above, we find that

$$\begin{aligned}&\iint_{\Pi_T}|u-c|\,\varphi_t+F(u,x,t,c)\varphi_x\,dtdx-\sum_m\int_{x_m}^{x_{m+1}}\int_0^T\operatorname{sign}(u-c)\,f_a(u,a,g)a'(x)\varphi\,dtdx\\&\qquad+\sum_m\int_0^T\left|f(c,a(x_m^+,g(t))-f(c,a(x_m^-),g(t))\right|\varphi(x_m,t)\,dt\\&\ge\lim_{\delta\downarrow 0}\sum_m\Big[(3.16)+(3.17)\Big]\\&\ge 0.\end{aligned}$$

Hence the limit u is an entropy solution of (1.1), since by the Lemma 3.2 also $u(\cdot,t) \to u_0$ as $t \downarrow 0$. Summing up, we have shown the following.

THEOREM 3.3. *Assume that the assumptions* (A.1), (A.2), (A.3), (A.4), (A.5), *and* (A.6) *hold. Then the sequence of front tracking solutions defined in section* 2 *converges in* $L^1(\Pi_T)$ *to an entropy weak solution of* (1.1).

Now it is straightforward to use methods from [11] to show that this problem has a unique solution. This is contained in the following theorem.

THEOREM 3.4. *Assume that the assumptions* (A.1), (A.2), (A.3), (A.4), (A.5), *and* (A.6) *hold. Let* $u = u(x,t)$ *and* $v = v(x,t)$ *be two entropy weak solutions of*

$$u_t + f(u,a,g)_x = 0 \quad \textit{in the strip } \Pi_T$$

for some $T > 0$, *satisfying the initial conditions*

$$u(x,0) = u_0(x), \qquad v(x,0) = v_0(x), \qquad x \in \mathbf{R}.$$

Then we have that

$$\|u(t,\cdot) - v(t,\cdot)\|_{L^1(\mathbf{R})} \le \|u(s,\cdot) - v(s,\cdot)\|_{L^1(\mathbf{R})} \tag{3.21}$$

for each $0 \le s \le t < T$.

4. An example. To demonstrate that the front tracking construction also has some potential as a practical numerical method, in this section we show an example of how the front tracking construction works on a concrete example. We use the simple flux function

$$f(u,a,g) = 4agu(1-u), \tag{4.1}$$

and

$$\begin{aligned} a(x) &= \left(1 + 8\chi_{\{x<0.25\}}\,|x|\right) + 2\cos^2(\pi x), \\ g(t) &= 0.6 + 0.55\cos\left(2\pi t\left(\chi_{\{0.15<t<0.65\}} + 1\right)\right), \\ u_0(x) &= 0.5(1 - 0.8\sin(\pi x)). \end{aligned} \tag{4.2}$$

In the example we present, we have used periodic initial data in the interval $x \in [0,1]$, $\delta = 1/20$, and $\Delta t = 0.025$. In Figure 5 we show the initial data u_0^δ and the approximate coefficients a^δ and g^δ. In Figure 6 we show the front tracking solution at $t = 0.5$ and the fronts for $0 \le t \le 0.5$. The a-fronts are depicted as broken horizontal lines, the g-fronts as vertical broken lines, and the u-fronts as solid lines.

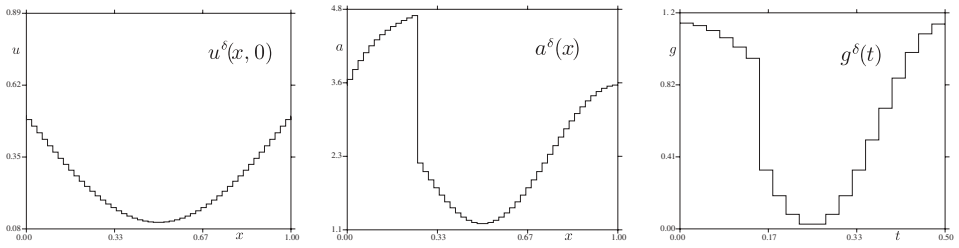

FIG. 5. *The initial function* u_0^δ *(left), the coefficients* a^δ *(middle) and* g^δ *(right).*

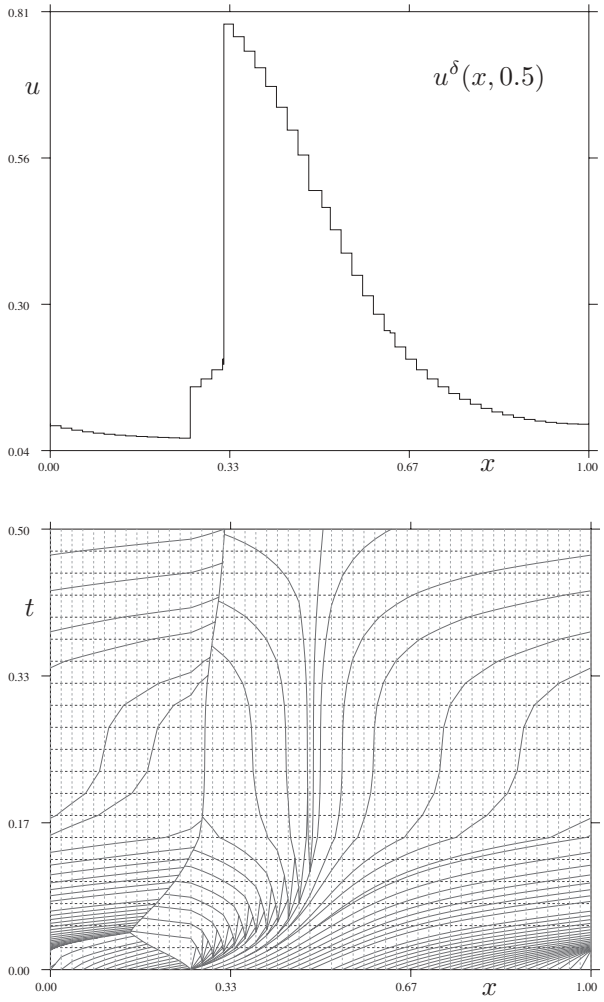

FIG. 6. *The front tracking solution* $u^\delta(x, 0.5)$ *(top) and the fronts (bottom).*

REFERENCES

[1] A. BRESSAN, T.-P. LIU, AND T. YANG, *L^1 stability estimates for $n \times n$ conservation laws*, Arch. Ration. Mech. Anal., 149 (1999), pp. 1–22.

[2] R. BÜRGER, K. H. KARLSEN, C. KLINGENBERG, AND N. H. RISEBRO, *A front tracking approach to a model of continuous sedimentation in ideal clarifier-thickener units*, Nonlinear Anal. Real World Appl., 4 (2003), pp. 457–481.

[3] R. BÜRGER, K. H. KARLSEN, N. H. RISEBRO, AND J. D. TOWERS, *Numerical methods for the simulation of continuous sedimentation in ideal clarifier-thickener units*, Int. J. Mineral Process., 73 (2004), pp. 209–228.

[4] M. BUSTOS, F. CONCHA, R. BÜRGER, AND E. TORY, *Sedimentation and Thickening: Phenomenological Foundation and Mathematical Theory*, Kluwer Academic, Dordrecht, The Netherlands, 1999.

[5] G. M. COCLITE AND N. H. RISEBRO, *Viscosity solutions of Hamilton-Jacobi equations with discontinuous coefficients*, in preparation.

[6] T. GIMSE AND N. H. RISEBRO, *Solution of the Cauchy problem for a conservation law with a discontinuous flux function*, SIAM J. Math. Anal., 23 (1992), pp. 635–648.

[7] J. GLIMM, *Solutions in the large for nonlinear hyperbolic systems of equations*, Comm. Pure

Appl. Math., 18 (1965), pp. 697–715.
[8] H. Holden and N. H. Risebro, *Front Tracking for Hyperbolic Conservation Laws*, Appl. Math. Sci. 152, Springer-Verlag, New York, 2002.
[9] J. M.-K. Hong, *Part* I: *An Extension of the Riemann Problems and Glimm's Method to General Systems of Conservation Laws with Source Terms. Part* II: *A Total Variation Bound on the Conserved Quantities for a Generic Resonant Nonlinear Balance Laws*, Ph.D. thesis, University of California, Davis, 2000.
[10] K. H. Karlsen, C. Klingenberg, and N. H. Risebro, *A relaxation scheme for conservation laws with a discontinuous coefficient*, Math. Comp., 73 (2004), pp. 1235–1259.
[11] K. H. Karlsen, N. H. Risebro, and J. D. Towers, *L^1 stability for entropy solutions of nonlinear degenerate parabolic convection-diffusion equations with discontinuous coefficients*, Skr. K. Nor. Vidensk. Selsk., 3 (2003), pp. 1–49.
[12] K. H. Karlsen, N. H. Risebro, and J. D. Towers, *On a nonlinear degenerate parabolic transport-diffusion equation with a discontinuous coefficient*, Electron. J. Differential Equations, 2002 (2002), pp. 1–23.
[13] R. A. Klausen and N. H. Risebro, *Stability of conservation laws with discontinuous coefficients*, J. Differential Equations, 157 (1999), pp. 41–60.
[14] C. Klingenberg and N. H. Risebro, *Convex conservation laws with discontinuous coefficients. Existence, uniqueness and asymptotic behavior*, Comm. Partial Differential Equations, 20 (1995), pp. 1959–1990.
[15] C. Klingenberg and N. H. Risebro, *Stability of a resonant system of conservation laws modeling polymer flow with gravitation*, J. Differential Equations, 170 (2001), pp. 344–380.
[16] S. N. Kružkov, *First order quasi-linear equations in several independent variables*, Math. USSR-Sb., 10 (1970), pp. 217–243.
[17] M. J. Lighthill and G. B. Whitham, *On kinematic waves.* II. *Theory of traffic flow on long crowded roads*, Proc. Roy. Soc. London. Ser. A, 229 (1955), pp. 317–345.
[18] L. Lin, B. Temple, and J. Wang, *Suppression of Oscillations in Godunov's Method for a Resonant Non-strictly Hyperbolic System*, tech. rep., University of California, Davis, 1992.
[19] O. A. Oleĭnik, *Construction of a generalized solution of the Cauchy problem for a quasi-linear equation of first order by the introduction of "vanishing viscosity,"* Amer. Math. Soc. Transl., 33 (1963), pp. 277–283.
[20] D. N. Ostrov, *Extending viscosity solutions to Eikonal equations with discontinuous spatial dependence*, Nonlinear Anal., 42 (2000), pp. 709–736.
[21] D. N. Ostrov, *Viscosity solutions and convergence of monotone schemes for synthetic aperture radar shape-from-shading equations with discontinuous intensities*, SIAM J. Appl. Math., 59 (1999), pp. 2060–2085.
[22] N. Seguin and J. Vovelle, *Analysis and approximation of a scalar conservation law with a flux function with discontinuous coefficients*, Math. Models Methods Appl. Sci., 13 (2003), pp. 221–257.
[23] B. Temple, *Global solution of the Cauchy problem for a class of 2×2 non-strictly hyperbolic conservation laws*, Adv. in Appl. Math., 3 (1982), pp. 335–375.
[24] J. D. Towers, *Convergence of a difference scheme for conservation laws with a discontinuous flux*, SIAM J. Numer. Anal., 38 (2000), pp. 681–698.
[25] J. D. Towers, *A difference scheme for conservation laws with a discontinuous flux: The nonconvex case*, SIAM J. Numer. Anal., 39 (2001), pp. 1197–1218.
[26] D. Wagner, *Equivalence of the Euler and Lagrangian equations of gas dynamics for weak solutions*, J. Differential Equations, 68 (1987), pp. 118–136.

SIAM J. MATH. ANAL.
Vol. 36, No. 4, pp. 1310–1345

BLOWUP SOLUTIONS FOR A LIOUVILLE EQUATION WITH SINGULAR DATA*

PIERPAOLO ESPOSITO†

Abstract. We study the existence of multiple blowup solutions for a semilinear elliptic equation with homogeneous Dirichlet boundary condition, exponential nonlinearity, and a singular source term given by Dirac masses. In particular, we extend the result of Baraket and Pacard [*Calc. Var. Partial Differential Equations*, 6 (1998), pp. 1–38] by allowing the presence, in the equation, of a weight function possibly vanishing in some points.

Key words. Liouville equation, singular data, exponential nonlinearity, blowup solutions

AMS subject classifications. 35J20, 35J25, 35J60

DOI. 10.1137/S0036141003430548

1. Main results and examples.

Let $\Omega \subset \mathbb{R}^2$ be a smooth bounded open set. We are concerned with the existence of solutions in the distributional sense for the problem

$$\begin{cases} -\triangle u = \rho^2 e^u - 4\pi \sum_{i=1}^N \alpha_i \delta_{p_i} & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega \end{cases} \tag{1}$$

with the property that $\rho^2 e^u$ "concentrates" when the parameter $\rho \to 0$. Here $\{\alpha_1, \ldots, \alpha_N\}$ are positive numbers, δ_p defines the Dirac mass at p, and $\Gamma := \{p_1, \ldots, p_N\} \subset \Omega$ is the set of singular sources in (1).

Problem (1) with $\Gamma = \emptyset$ has been largely studied in connection with many physical models such as thermionic emission [21], the theory of the isothermal gas sphere [14], gas combustion [25], and in the context of statistical mechanics in [11], [12], and [23]. The asymptotic analysis for blowup solutions to problem (1) as $\rho \to 0$ is contained in [36] (see also [27]) and alternatively it can be obtained as a by-product of the general blowup analysis of [8]: it leads in the limit to a quantization property of the energy $\rho^2 \int_\Omega e^u$ in terms of the number of blowup points and to a characterization of the location of the blowup points. For the converse question, namely, the construction of solutions to (1) which do blow up at the "admissible" points as $\rho \to 0$, the first result is due to Weston [38] who constructed a sequence of solutions on simply connected domains "concentrating" on a single blowup point according to [36] (see also [26] for more general nonlinearities). The general case of the existence of multiple blowup solutions has been treated only in the beautiful paper [4]. Subsequently Chen and Lin in [17] have given an alternative proof in the special case of an annulus. Thus, perturbative problems with exponential nonlinearities in dimension two seem to be very difficult to handle. So far only a few results have been derived that cover some special cases (beside [4] and its extensions [2], [3], and [5], see also [13] and [29]) in contrast to the vast literature available in higher dimensions; see, for example, [1], [15], [28], and [32].

*Received by the editors June 24, 2003; accepted for publication (in revised form) April 30, 2004; published electronically March 17, 2005. This research was supported by MIUR, national project *Variational Methods and Nonlinear Differential Equations*.

http://www.siam.org/journals/sima/36-4/43054.html

†Dipartimento di Matematica, Università di Roma "Tor Vergata," Via della Ricerca Scientifica, 00133 Roma, Italy (pesposit@axp.mat.uniroma2.it).

Motivated by some singular elliptic equations arising in the study of Chern–Simons vortex theory (we refer the reader to [39] and the references therein), we are interested in analyzing (1) with $\Gamma \neq \emptyset$. We mention that some of the progress made about condensates in Chern–Simons models is contained in [10], [30], [33], [34], [35], and [37] to quote a few.

Let $G(z, z')$ denote the Green's function of $-\triangle$ on Ω with Dirichlet boundary condition, namely,

$$\begin{cases} -\triangle_z G(z, z') = \delta_{z'} & \text{in } \Omega, \\ G(z, z') = 0 & \text{on } \partial\Omega, \end{cases}$$

and let $H(z, z') = \frac{1}{2\pi} \ln|z - z'| + G(z, z')$ be the regular part of $G(z, z')$. Problem (1) is equivalent to solving for $v = u + 4\pi \sum_{i=1}^{N} \alpha_i G(z, p_i)$, the regular part of u, the equation

$$(2) \qquad \begin{cases} -\triangle v = \rho^2 |z - p_1|^{2\alpha_1} \cdots |z - p_N|^{2\alpha_N} e^{-4\pi \sum_{i=1}^{N} \alpha_i H(z, p_i)} e^v & \text{in } \Omega, \\ v = 0 & \text{on } \partial\Omega. \end{cases}$$

Thus, we may consider the following general model problem:

$$(Q)_\rho \qquad \begin{cases} -\triangle v = \rho^2 |z - p_1|^{2\alpha_1} \cdots |z - p_N|^{2\alpha_N} f(z) e^v & \text{in } \Omega, \\ v = 0 & \text{on } \partial\Omega, \end{cases}$$

where $\Gamma = \{p_1, \ldots, p_N\} \subset \Omega$ and $\{\alpha_1, \ldots, \alpha_N\}$ are positive numbers, $f : \Omega \to \mathbb{R}$ is a smooth function such that $f(p_i) > 0$ for any $i = 1, \ldots, N$. An extension to the singular case of the blowup analysis in [8] is due to [7] (see also [6]). It permits us to perform an asymptotic analysis in the spirit of [36] (see [20] for a proof). To this purpose, set $\Gamma = \{p_1, \ldots, p_N\}$ and $\Omega' = \Omega \cap \{f > 0\}$, and for given $m \in \mathbb{N}$ and $s \in \{1, \ldots, N\}$ define

$$\tilde{\mathcal{F}}(z_1, \ldots, z_m) = \sum_{i=1}^{m} H(z_i, z_i) + \sum_{i \neq j} G(z_i, z_j) + \frac{1}{4\pi} \sum_{i=1}^{m} \ln\left(|z_i - p_1|^{2\alpha_1} \cdots |z_i - p_N|^{2\alpha_N} f(z_i)\right)$$

which is well defined in $(\Omega' \setminus \Gamma)^m$ for $z_i \neq z_j$ whenever $i \neq j$, and let

$$\mathcal{G}(z_1, \ldots, z_m, \omega_1, \ldots, \omega_s) = \frac{1}{4\pi} \left(\sum_{i=1}^{m} \sum_{j=1}^{s} 8\pi(1 + \alpha_j) G(z_i, \omega_j) \right)$$

be well defined for $z_i \neq \omega_j$, with $z_i \in \Omega$, $\omega_j \in \mathbb{C}$, $i = 1, \ldots, m$, $j = 1, \ldots, s$.

THEOREM 1.1. *Let $\Omega \subset \mathbb{R}^2$ be a smooth bounded open set and let f be a smooth positive function. Let v_ρ be a sequence of solutions of $(Q)_\rho$ such that $\sup_\rho T_\rho < +\infty$, $T_\rho = \rho^2 \int_\Omega |z - p_1|^{2\alpha_1} \cdots |z - p_N|^{2\alpha_N} f(z) e^{v_\rho}$. If $T_\rho \to 0$ as $\rho \to 0$, then $v_\rho \to 0$ in $C^{2,\beta}(\Omega)$ and, for ρ small, v_ρ coincides with the unique minimal solution of $(Q)_\rho$. If $T_\rho \to L \neq 0$, then (up to a subsequence) there exists a nonempty finite set $S = \{q_1, \ldots, q_K\} \subset \Omega$ (blowup set) such that $\rho^2 |z - p_1|^{2\alpha_1} \cdots |z - p_N|^{2\alpha_N} f(z) e^{v_\rho} \rightharpoonup \sum_{i=1}^{K} b_i \delta_{q_i}$ in the sense of measures and $v_\rho \to \sum_{i=1}^{K} b_i G(z, q_i)$ in $C^{2,\beta}_{loc}(\Omega \setminus S)$ for some $\beta \in (0, 1)$, with $b_i = 8\pi$ if $q_i \notin \Gamma$, or $b_i = 8\pi(1 + \alpha_j)$ if $q_i = p_j$ for some $j = 1, \ldots, N$.*

Moreover, if $S\cap\Gamma=\emptyset$, *then* $(q_1,\dots,q_K)$ *is a critical point for the function* $\tilde{\mathcal{F}}$; *if* $S\cap\Gamma=\{p_{j_1},\dots,p_{j_s}\}$ *and* $S\setminus\Gamma=\{q_{i_1},\dots,q_{i_m}\}$ *with* $m+s=K$, *then* $(q_{i_1},\dots,q_{i_m})$ *is a critical point for the function* $\tilde{\mathcal{F}}+\mathcal{G}(\cdot,p_{j_1},\dots,p_{j_s})$.

For the existence of a minimal solution of $(Q)_\rho$, we refer the reader to [19]. As a vice versa of Theorem 1.1, we establish the following result.

THEOREM 1.2. *Let* $\Omega\subset\mathbb{R}^2$ *be a smooth bounded open set,* f *be a smooth function, and* $\{\alpha_1,\dots,\alpha_N\}\subset(0,+\infty)\setminus\mathbb{N}$ *be real numbers. We have*

(a) *let* $S=\{p_{j_1},\dots,p_{j_s}\}\subset\Gamma$, *then there exist* $\rho_0>0$ *small and a family* $\{v_\rho\}_{0<\rho<\rho_0}$ *of solutions for equation* $(Q)_\rho$ *such that* $\rho^2|z-p_1|^{2\alpha_1}\cdots|z-p_N|^{2\alpha_N}f(z)e^{v_\rho}\rightharpoonup\sum_{i=1}^s 8\pi(1+\alpha_{j_i})\delta_{p_{j_i}}$ *in the sense of measures and* $v_\rho\to\sum_{i=1}^s 8\pi(1+\alpha_{j_i})G(z,p_{j_i})$ *in* $C^{2,\beta}_{loc}(\Omega\setminus S)$ *for some* $\beta\in(0,1)$;

(b) *let* $S=\{q_1,\dots,q_m\}\subset\Omega'\setminus\Gamma$ *and* $(q_1,\dots,q_m)$ *be a nondegenerate critical point of* $\tilde{\mathcal{F}}$ *such that* $\Delta\ln f(q_1)=\cdots=\Delta\ln f(q_m)=0$, *then there exist* $\rho_0>0$ *small and a family* $\{v_\rho\}_{0<\rho<\rho_0}$ *of solutions for* $(Q)_\rho$ *such that* $\rho^2|z-p_1|^{2\alpha_1}\cdots|z-p_N|^{2\alpha_N}f(z)e^{v_\rho}\rightharpoonup\sum_{i=1}^m 8\pi\delta_{q_i}$ *in the sense of measures and* $v_\rho\to\sum_{i=1}^m 8\pi G(z,q_i)$ *in* $C^{2,\beta}_{loc}(\Omega\setminus S)$ *for some* $\beta\in(0,1)$;

(c) *let* S *be such that* $S\cap\Gamma=\{p_{j_1},\dots,p_{j_s}\}$, $S\setminus\Gamma=\{q_1,\dots,q_m\}$, *and* $(q_1,\dots,q_m)$ *is a critical point of* $\tilde{\mathcal{F}}+\mathcal{G}(\cdot,p_{j_1},\dots,p_{j_s})$ *such that* $\Delta\ln f(q_1)=\cdots=\Delta\ln f(q_m)=0$, *then there exist* $\rho_0>0$ *small and a family* $\{v_\rho\}_{0<\rho<\rho_0}$ *of solutions for* $(Q)_\rho$ *such that* $\rho^2|z-p_1|^{2\alpha_1}\cdots|z-p_N|^{2\alpha_N}f(z)e^{v_\rho}\rightharpoonup\sum_{k=1}^s 8\pi(1+\alpha_{j_k})\delta_{p_{j_k}}+\sum_{j=1}^m 8\pi\delta_{q_j}$ *in the sense of measures and* $v_\rho\to\sum_{k=1}^s 8\pi(1+\alpha_{j_k})G(z,p_{j_k})+\sum_{j=1}^m 8\pi G(z,q_j)$ *in* $C^{2,\beta}_{loc}(\Omega\setminus S)$ *for some* $\beta\in(0,1)$.

Let us point out that the assumption $\Delta\ln f(q_i)=0$ for any $q_i\in S\setminus\Gamma$ is always fulfilled by the original problem (2), so in some sense it seems a "natural" assumption from a physical point of view. In case $\Gamma=\emptyset$, part (b) in Theorem 1.2 gives a direct extension of the result in [4], which has largely motivated our approach. More precisely, it states the following.

COROLLARY 1.3. *Let* $\Omega\subset\mathbb{R}^2$ *be a smooth bounded open set,* f *be a smooth function, and* $S=\{q_1,\dots,q_m\}\subset\Omega'$ *be a nonempty set. Assume that* $(q_1,\dots,q_m)$ *is a nondegenerate critical point of* $\tilde{\mathcal{F}}(z_1,\dots,z_m)=\sum_{i=1}^m H(z_i,z_i)+\sum_{i\neq j}G(z_i,z_j)+\frac{1}{4\pi}\sum_{i=1}^m\ln f(z_i)$ *in* $(\Omega')^m$ *such that* $\Delta\ln f(q_1)=\cdots=\Delta\ln f(q_m)=0$. *There exist* $\rho_0>0$ *small and a family* $\{v_\rho\}_{0<\rho<\rho_0}$ *of solutions for the equation*

$$\begin{cases}-\triangle v=\rho^2 f(z)e^v & \text{in } \Omega,\\ v=0 & \text{on } \partial\Omega,\end{cases}$$

such that $\rho^2 f(z)e^{v_\rho}\rightharpoonup\sum_{i=1}^m 8\pi\delta_{q_i}$ *in the sense of measures and* $v_\rho\to\sum_{i=1}^m 8\pi G(z,q_i)$ *in* $C^{2,\beta}_{loc}(\Omega\setminus S)$ *for some* $\beta\in(0,1)$.

Thus, from Corollary 1.3 the result in [4] is recovered by taking $f=\text{const}>0$.

To avoid technicalities, we derive the proof of Theorem 1.2 only in the following significant cases: (a) holds with $S=\{p\}$ and $p\in\Gamma$, (b) holds with $S=\{q\}$ and $q\notin\Gamma$, and (c) holds with $S=\{p,q\}$, $p\in\Gamma$, and $q\notin\Gamma$. Our approach generalizes to any number of "peaks," the technical details are worked out in [20]. So we restrict our attention to the problem

$$(P)_\rho\qquad\begin{cases}-\triangle v=\rho^2|z-p|^{2\alpha}f(z)e^v & \text{in } \Omega,\\ v=0 & \text{on } \partial\Omega,\end{cases}$$

where $\alpha\in(0,+\infty)\setminus\mathbb{N}$ is a real number and $f:\Omega\to\mathbb{R}$ is a smooth function not necessarily positive. We will prove the following result.

THEOREM 1.4. *Under the above assumptions we have*

(a) *if $p \in \Omega$ with $f(p) > 0$, there exist $\rho_0 > 0$ small and a family $\{v_\rho\}_{0<\rho<\rho_0}$ of solutions for $(P)_\rho$ such that $\rho^2|z-p|^{2\alpha} f(z) e^{v_\rho} \rightharpoonup 8\pi(1+\alpha)\delta_p$ in the sense of measures and $v_\rho \to 8\pi(1+\alpha)G(z,p)$ in $C^{2,\beta}_{loc}(\Omega \setminus \{p\})$ for some $\beta \in (0,1)$;*

(b) *if $q \in \Omega' \setminus \{p\}$ is a nondegenerate critical point of $\mathcal{F}(z) = H(z,z) + \frac{1}{4\pi}\ln[|z-p|^{2\alpha} f(z)]$ in $\Omega' \setminus \{p\}$ such that $\Delta \ln f(q) = 0$, then there exist $\rho_0 > 0$ small and a family $\{v_\rho\}_{0<\rho<\rho_0}$ of solutions for $(P)_\rho$ such that $\rho^2|z-p|^{2\alpha} f(z) e^{v_\rho} \rightharpoonup 8\pi\delta_q$ in the sense of measures and $v_\rho \to 8\pi G(z,q)$ in $C^{2,\beta}_{loc}(\Omega \setminus \{q\})$ for some $\beta \in (0,1)$;*

(c) *if $p \in \Omega$ with $f(p) > 0$ and $q \neq p$ is a nondegenerate critical point of $\mathcal{F}(z) = \tilde{\mathcal{F}}(z) + \mathcal{G}(z,p) = H(z,z) + \frac{1}{4\pi}\ln(|z-p|^{2\alpha} f(z)) + 2(1+\alpha)G(z,p)$ in $\Omega' \setminus \{p\}$ such that $\Delta \ln f(q) = 0$, then there exist $\rho_0 > 0$ small and a family $\{v_\rho\}_{0<\rho<\rho_0}$ of solutions for $(P)_\rho$ such that $\rho^2|z-p|^{2\alpha} f(z) e^{v_\rho} \rightharpoonup 8\pi(1+\alpha)\delta_p + 8\pi\delta_q$ in the sense of measures and $v_\rho \to 8\pi(1+\alpha)G(z,p) + 8\pi G(z,q)$ in $C^{2,\beta}_{loc}(\Omega \setminus \{p,q\})$ for some $\beta \in (0,1)$.*

We now discuss some applications of the results above. As it is well known, for $\alpha \geq 0$, the problem

$$\begin{cases} -\triangle v = \lambda \frac{|z-p|^{2\alpha} f(z) e^v}{\int_\Omega |z-p|^{2\alpha} f(z) e^v} & \text{in } B(0,1), \\ v = 0 & \text{on } \partial B(0,1) \end{cases} \tag{3}$$

with $p = 0$ and $f(z) = 1$ possesses a radial solution for $0 < \lambda < 8\pi(\alpha+1)$ and, as a consequence of a Pohozaev identity, has no solution for $\lambda \geq 8\pi(\alpha+1)$. By means of Theorem 1.4, we can show that such a threshold for existence of (3) is no longer valid if we perturb (3) either by replacing $f = 1$ with a suitable nonconstant function or by moving p close to $\partial B(0,1)$. In fact we will be able to produce solutions v_ρ for (3) with $\lambda_\rho = \rho^2 \int_\Omega |z-p|^{2\alpha} f(z) e^{v_\rho} \to 8\pi(\alpha+1) + 8\pi > 8\pi(\alpha+1)$ concentrating on two points. According to Theorem 1.4, for this purpose we need to exhibit a nondegenerate critical point q for $\mathcal{F}(z) = H(z,z) - \frac{2+\alpha}{2\pi}\ln|z-p| + 2(1+\alpha)H(z,p) + \frac{1}{4\pi}\ln f(z)$ such that $\Delta \ln f(q) = 0$. Let us recall that $H(z,p) = \frac{1}{4\pi}\ln\left(|p|^2|z|^2 - 2\langle p,z\rangle + 1\right)$ and $H(z,z) = \frac{1}{2\pi}\ln\left(1-|z|^2\right)$, where $\langle\cdot,\cdot\rangle$ denotes the inner product in $\mathbb{R}^2$. Hence we obtain for $\mathcal{F}(z)$ the expression

$$\mathcal{F}(z) = \frac{1}{2\pi}\ln\left(1-|z|^2\right) - \frac{2+\alpha}{2\pi}\ln|z-p| + \frac{1+\alpha}{2\pi}\ln\left(|p|^2|z|^2 - 2\langle p,z\rangle + 1\right) + \frac{1}{4\pi}\ln f(z).$$

Example 1.5. We study now the case $p = 0$. For fixed $q \in B(0,1) \setminus \{0\}$, we can define a function $f(z)$ such that in a small neighborhood of q it takes the form

$$f(z) = \exp\left((z_1-q_1)^2 - c_q(z_2-q_2)^2 - 2\ln\left(1-|z|^2\right) + 2(2+\alpha)\ln|z|\right),$$

where $c_q = 1 + \frac{4}{(1-|q|^2)^2} > 0$. For such a function $f(z)$, the function $\mathcal{F}(z)$ near q takes the form $\mathcal{F}(z) = \frac{1}{4\pi}\left[(z_1-q_1)^2 - c_q(z_2-q_2)^2\right]$ and hence q is a nondegenerate critical point of $\mathcal{F}(z)$ such that $\Delta \ln f(q) = 2 - 2c_q + \frac{8}{(1-|q|^2)^2} = 0$. Moreover, if we choose q such that $|q| = r_\alpha$, with $r_\alpha \in (0,1)$ satisfying $r_\alpha^{2+\alpha} + r_\alpha^2 - 1 = 0$, then such an f may be constructed as a small perturbation of the constant function 1. In fact, for ϵ small we can just take f_ϵ of the form

$$f_\epsilon(z) = \left(1 - \chi\left(\frac{z-q}{\epsilon}\right)\right) + \chi\left(\frac{z-q}{\epsilon}\right) \times \exp\left(\epsilon(z_1-q_1)^2 - c_\epsilon(z_2-q_2)^2 - 2\ln\left(1-|z|^2\right) + 2(2+\alpha)\ln|z|\right),$$

where $c_\epsilon = \epsilon + \frac{4}{(1-|q|^2)^2}$ and $0 \le \chi \le 1$ is a smooth cut-off function such that $\chi = 1$ in $B(0,1)$ and $\chi = 0$ in $\mathbb{R}^2 \setminus B(0,2)$.

Example 1.6. We study now the case $f(z) = 1$. The function $\mathcal{F}(z)$ becomes

$$\mathcal{F}(z) = \frac{1}{2\pi} \ln\big(1 - |z|^2\big) - \frac{2+\alpha}{2\pi} \ln|z-p| + \frac{1+\alpha}{2\pi} \ln\big(|p|^2|z|^2 - 2\langle p, z\rangle + 1\big).$$

Let us remark that, according to the nonexistence result stated above, for $p = 0$ the function $\mathcal{F}(z) = \frac{1}{2\pi} \ln\big(1 - |z|^2\big) - \frac{2+\alpha}{2\pi} \ln|z|$ has no critical points in $B(0,1) \setminus \{0\}$ and this remains true for p close to zero. On the other hand, we can take $p \in B(0,1)$ such that $p \to e \in \partial B(0,1)$ along a straight line. We consider a point $q = se$ for $s \in (-1,1)$. We have that

$$\nabla \mathcal{F}(q) = \left(\frac{(\alpha+2)s + \alpha}{2\pi(s^2-1)} + o(1) \right) e \quad \text{as } p \to e$$

for $|s-1|$ bounded away from zero. Let $s_0 = -\frac{\alpha}{\alpha+2}$, for p close to e we find a point s_p such that $\nabla \mathcal{F}(s_p e) = 0$ and $s_p \to s_0$ as $p \to e$. We evaluate now the determinant of $D^2\mathcal{F}(s_p e)$:

$$\det D^2\mathcal{F}(s_p e) = \frac{(\alpha+2)^6}{64\pi^2(\alpha+1)^3} + o(1) \quad \text{as } p \to e.$$

Hence $q_p = s_p e$ is a nondegenerate critical point of $\mathcal{F}(z)$ for p close to e such that $q_p \to -\frac{\alpha}{\alpha+2} e$ as $p \to e$.

As in [4], Theorem 1.4 is based on the construction of a suitable family of approximate solutions $v(\rho, \lambda, a)$ for problem $(P)_\rho$, with (λ, a) a suitable set of parameters, such that the linearized operator about $v(\rho, \lambda, a)$ is invertible. Thus, for ρ small a fixed point argument will provide a solution v_ρ close in some sense to $v(\rho, \lambda, a)$ with the required asymptotic properties.

2. Construction of approximating solutions. As far as part (a) in Theorem 1.4 is concerned, in view of the expected asymptotic behavior, the approximating function $v(\rho, 0, 0)$ will be constructed by gluing in a small neighborhood of p the limit function $8\pi(1+\alpha)G(z,p)$ with a suitable local solution of $-\Delta v = \rho^2|z-p|^{2\alpha} f(p) e^v$. Using the scale invariance $v(z) \to v_t(z) = v(tz) + 2(\alpha+1)\ln t$, $t > 0$, valid for the solutions of the equation

$$-\Delta v = \rho^2 |z|^{2\alpha} e^v, \tag{4}$$

we can construct local solutions which are very concentrated near p in such a way that the gluing with $8\pi(1+\alpha)G(z,p)$ is sufficiently accurate. This is possible in view of the fact that $8\pi(1+\alpha)G(z,p) \to +\infty$ as $z \to p$. For part (b) in Theorem 1.4, we glue in a small neighborhood of q the limit function $8\pi G(z,q)$ with a suitable local solution of $-\Delta v = \rho^2|q-p|^{2\alpha} f(q) e^v$. The scale invariance involved here is $v(z) \to v_t(z) = v(tz) + 2\ln t$, $t > 0$, valid for solutions of

$$-\Delta v = \rho^2 e^v. \tag{5}$$

Finally, for part (c) in Theorem 1.4 we combine the two previous constructions by gluing the limit function $8\pi(1+\alpha)G(z,p) + 8\pi G(z,q)$ with a local solution of $-\Delta v = \rho^2|z-p|^{2\alpha} f(p) e^v$ near p and with a local solution of $-\Delta v = \rho^2|q-p|^{2\alpha} f(q) e^v$ near q.

To this purpose, we recall some known facts. The solutions of $-\Delta u = e^u$ in $\mathbb{R}^2$ have been completely classified by Liouville in [24] and in complex notations they satisfy the so-called Liouville formula

$$\ln \frac{8|F'(z)|^2}{(1+|F(z)|^2)^2} \tag{6}$$

for some meromorphic function F with $F'(z) \neq 0$ whenever defined.

This representation formula generalizes to solutions in the punctured plane $\mathbb{C}\backslash\{0\}$, as proved in [18], by choosing some multivalued meromorphic function $F : \mathbb{C} \to \mathbb{C}$, locally univalent in $\mathbb{C} \setminus \{0\}$, satisfying

$$\text{either} \quad F(z) = G(z)z^\gamma,\ \gamma \in \mathbb{R}, \quad \text{or} \quad F(z) = \Phi(\sqrt{z}),$$

where G and Φ are single-valued holomorphic functions away from the origin and where $\Phi(z)\Phi(-z) = 1$.

A complete classification for solutions of

$$\begin{cases} -\Delta u = e^u & \text{in } \mathbb{R}^2, \\ \int_{\mathbb{R}^2} e^u < +\infty \end{cases} \tag{7}$$

can be performed either by the Liouville formula or via the moving plane method (see [16]) and it leads to the only possible choice of $F(z) = az + b$, with $a, b \in \mathbb{C}$. The complete classification for solutions of

$$\begin{cases} -\Delta u = e^u - 4\pi\alpha\delta_{p=0} & \text{in } \mathbb{R}^2, \\ \int_{\mathbb{R}^2} e^u < +\infty \end{cases} \tag{8}$$

is due to [31] and it corresponds to the choice $F(z) = az^{\alpha+1} + b$, with $a, b \in \mathbb{C}$ and $b = 0$ if $\alpha \notin \mathbb{N}$. By choosing $F(z) = \frac{1}{\tau\rho} z(1+\gamma z^2)$, $\tau > 0$, $\gamma \in \mathbb{C}$ such that $|\gamma| < \frac{1}{3}$, and $F(z) = \frac{1}{\tau\rho} z^{\alpha+1}$, we can provide, respectively, solutions for (7) and (8) in $B(0,1)$. By taking the regular part of this functions and adding a term $2\ln\frac{1}{\rho}$, we obtain a large class of solutions for (4) and (5) in $B(0,1)$, respectively, in the form

$$v_{\rho,\tau} = \ln \frac{8(\alpha+1)^2\tau^2}{(\tau^2\rho^2 + |z|^{2(\alpha+1)})^2}, \qquad v_{\rho,\tau,\gamma} = \ln \frac{8\tau^2|1+3\gamma z^2|^2}{(\tau^2\rho^2 + |z|^2|1+\gamma z^2|^2)^2}. \tag{9}$$

Let $h(z)$ be some smooth function such that $h(0) > 0$. The function $v_{\rho,\tau} - \ln h(0)$ satisfies the equation $-\Delta\left(v_{\rho,\tau} - \ln h(0)\right) = \rho^2 h(0)|z|^{2\alpha} e^{v_{\rho,\tau} - \ln h(0)}$ in $B(0,1)$. Similarly $v_{\rho,\tau,0} - \ln h(0)$ is a solution for $-\Delta\left(v_{\rho,\tau,0} - \ln h(0)\right) = \rho^2 h(0) e^{v_{\rho,\tau,0} - \ln h(0)}$ in $B(0,1)$. For $\rho > 0$ small, they can be viewed as approximating solutions when we replace $h(0)$ by $h(z)$: such an approximation, however, may not be accurate enough to carry out our fixed point argument. In fact, we will need to define the local approximating solution $U_{\rho,\tau}$ as the difference between, respectively, $v_{\rho,\tau}$, $v_{\rho,\tau,0}$ and a Taylor expansion of $\ln h(z)$ at $z = 0$, taking into account two basic facts:

(a) $U_{\rho,\tau}$ must be a "good" local approximating solution;

(b) translating $U_{\rho,\tau}$ at some point $q \in S$, the difference between this local function and the related limit function as $\rho \to 0$ must be small in a small annulus centered at q.

In case $\alpha > 0$, $v_{\rho,\tau} - \ln h(0)$ is satisfactory for (a). For (b), if $p \in S \cap \Gamma$, we choose some $\tau > 0$ such that the Taylor expansion corresponding to the difference function in

a small annulus centered in p contains powers of $z-p$ of degree 1. In case $\alpha=0$ the situation is more delicate as there is more degeneracy. Assuming $\Delta \ln h(0)=0$, for (a) we need to take the local function $U_{\rho,\tau}$ of the form $v_{\rho,\tau,\gamma}-\ln h(0)-2\bar{z}\cdot\partial_z \ln h(0)-\bar{z}^2\cdot\partial_{zz}\ln h(0)$. While for (b) we need the difference function to be an infinitesimal term of order 3 as $z\to q$. This condition will be attained by specifying $\tau>0$ and γ suitably and by the condition $\partial_z \mathcal{F}(q)=0$. The invertibility of $D^2\mathcal{F}(q)$ will guarantee the invertibility of the linearized operator around such an approximating solution at $\rho=0$.

Summarizing, an appropriate approximating solution for our problem near a blowup point should look like

$$U_{\rho,\tau}(z)=\begin{cases} v_{\rho,\tau}(z)-\ln h(0) & \text{if } \alpha>0,\\ v_{\rho,\tau,\gamma}(z)-\ln h(0)-2\bar{z}\cdot\partial_z\ln h(0)-\bar{z}^2\cdot\partial_{zz}\ln h(0) & \text{if } \alpha=0 \end{cases}$$

with τ and γ suitably chosen. Introduce the differential operators $\partial_z=\frac{1}{2}(\partial_1-i\partial_2)$, $\partial_{\bar{z}}=\frac{1}{2}(\partial_1+i\partial_2)$, and the notation $2z\cdot z'=z\overline{z'}+\bar{z}z'=2\mathrm{Re}\,(z\overline{z'})$. Thus $\Delta=4\partial_z\partial_{\bar{z}}$ and the Taylor expansion in 0 for any smooth function $h:\Omega\to\mathbb{R}$ takes the form

$$h(z)=h(0)+2\bar{z}\cdot\partial_z h(0)+\bar{z}^2\cdot\partial_{zz}h(0)+\frac{|z|^2}{4}\Delta h(0)+O(|z|^3).$$

Hence $U_{\rho,\tau}$ is a solution in $B(0,1)$ of

$$-\Delta U_{\rho,\tau}=\begin{cases} \rho^2|z|^{2\alpha}h(0)e^{U_{\rho,\tau}} & \text{if } \alpha>0,\\ \rho^2 e^{\ln h(0)+2\bar{z}\cdot\partial_z\ln h(0)+\bar{z}^2\cdot\partial_{zz}\ln h(0)}e^{U_{\rho,\tau}} & \text{if } \alpha=0, \end{cases} \tag{10}$$

and we see that the right-hand side (RHS) of (10) may be expressed as follows:

$$\mathrm{RHS}=\begin{cases} \rho^2|z|^{2\alpha}h(z)e^{U_{\rho,\tau}}+O\left(\rho^2|z|^{2\alpha+1}e^{U_{\rho,\tau}}\right) & \text{if } \alpha>0,\\ \rho^2 h(z)e^{U_{\rho,\tau}}+O\left(\rho^2|z|^3e^{U_{\rho,\tau}}\right) & \text{if } \alpha=0 \end{cases}$$

provided that when $\alpha=0$ we also satisfy $\Delta\ln h(0)=0$.

By the assumptions in Theorem 1.4, we may translate the function $U_{\rho,\tau}(z)$ around the points p and q by defining

$$\begin{cases} U^1_\rho(z)=v_{\rho,\tau_1}(z-p)-\ln f(p),\\ U^2_\rho(z)=v_{\rho,\tau_2,\gamma}(z-q)-\ln\left(|z-p|^{2\alpha}f\right)(q)-2\overline{z-q}\cdot\partial_z\ln\left(|z-p|^{2\alpha}f\right)(q)\\ \qquad\qquad -\overline{z-q}^2\cdot\partial_{zz}\ln\left(|z-p|^{2\alpha}f\right)(q) \end{cases}$$

with τ_1, τ_2, and γ to be specified below. Thus, we have

$$\Delta U^i_\rho(z)+\rho^2|z-p|^{2\alpha}f(z)e^{U^i_\rho(z)}=\begin{cases} O\left(\rho^2|z-p|^{2\alpha+1}e^{U^1_\rho(z)}\right) & \text{in } B(p,1),\\ O\left(\rho^2|z-q|^3e^{U^2_\rho(z)}\right) & \text{in } B(q,1). \end{cases} \tag{11}$$

Note that the following expansions hold as $\rho\to0$:

$$v_{\rho,\tau}(z-p)=\ln 8(1+\alpha)^2\tau^2-4(1+\alpha)\ln|z-p|+O\left(\frac{\tau^2\rho^2}{|z-p|^{2(\alpha+1)}}\right),$$

$$v_{\rho,\tau,\gamma}(z-q)=\ln 8\tau^2-4\ln|z-q|+2\overline{z-q}^2\cdot\gamma+O\left(|z-q|^4+\frac{\tau^2\rho^2}{|z-q|^2}\right).$$

Let us define the limit function $L(z)$ as

$$L(z)=\begin{cases}8\pi(1+\alpha)G(z,p) & \text{if } S=\{p\},\\ 8\pi G(z,q) & \text{if } S=\{q\},\\ 8\pi(1+\alpha)G(z,p)+8\pi G(z,q) & \text{if } S=\{p,q\}.\end{cases}$$

Hence in $|z-p|<1$ we get

$$U^1_\rho(z)-L(z)=\ln 8(\alpha+1)^2\tau_1^2-\mathcal{F}_1(p)+O\left(\frac{\tau_1^2\rho^2}{|z-p|^{2(1+\alpha)}}+|z-p|\right),$$

while for $|z-q|<1$,

$$U^2_\rho(z)-L(z)=\ln 8\tau_2^2-\mathcal{F}_2(q)-2\overline{z-q}\cdot\partial_z\mathcal{F}_2(q)$$
$$-\overline{z-q}^2\cdot(\partial_{zz}\mathcal{F}_2(q)-2\gamma)+O\left(\frac{\tau_2^2\rho^2}{|z-q|^2}+|z-q|^3\right),$$

where

$$\mathcal{F}_1(z)=\begin{cases}8\pi(1+\alpha)H(z,p)+\ln f(z) & \text{if } S=\{p\},\\ 8\pi(1+\alpha)H(z,p)+8\pi G(z,q)+\ln f(z) & \text{if } S=\{p,q\}\end{cases}$$

and

$$\mathcal{F}_2(z)=\begin{cases}8\pi H(z,q)+\ln\left(|z-p|^{2\alpha}f(z)\right) & \text{if } S=\{q\},\\ 8\pi H(z,q)+\ln\left(|z-p|^{2\alpha}f(z)\right)+8\pi(1+\alpha)G(z,p) & \text{if } S=\{p,q\}.\end{cases}$$

Let us remark that by assumption $\partial_z\mathcal{F}_2(q)=0$. Now we specify the values for τ_1, τ_2, and γ to be fixed as follows:

$$\tau_1=\frac{e^{\frac12\mathcal{F}_1(p)}}{\sqrt{8}(1+\alpha)},\qquad \tau_2=\frac{e^{\frac12\mathcal{F}_2(q)}}{\sqrt{8}},\qquad \gamma=\frac12\partial_{zz}\mathcal{F}_2(q).$$

In such a way we obtain

$$U^1_\rho(z)-L(z)=O\left(\frac{\tau_1^2\rho^2}{|z-p|^{2(1+\alpha)}}+|z-p|\right)\quad\text{in } |z-p|<1 \tag{12}$$

and

$$U^2_\rho(z)-L(z)=O\left(\frac{\tau_2^2\rho^2}{|z-q|^2}+|z-q|^3\right)\quad\text{in } |z-q|<1. \tag{13}$$

By scaling the variables, we can always assume that $\overline{B(p,2)}\cap\overline{B(q,2)}=\emptyset$, $\overline{B(p,2)}\subset\Omega$, $\overline{B(q,2)}\subset\Omega$, and $|\gamma|<\frac13$. For $i=1,2$ let $r_i=r_i(\rho)$ be a positive smooth function such that $\frac{\rho^2}{r_1^{4\alpha+5}}=O(1)$ as $\rho\to0$ and $\frac{\rho^2}{r_2^5}=O(1)$ as $\rho\to0$. Let χ be a radial smooth function such that $0\le\chi\le1$, $\chi=1$ in $B(0,1)$, and $\chi=0$ in $\mathbb{R}^2\setminus B(0,2)$.

To obtain part (a) in Theorem 1.4, for $\lambda_1\in\mathbb{R}$, $|\lambda_1|<\frac12\tau_1$, we consider the approximating function

$$v(\rho,\lambda_1)(z)=\left(1-\chi\left(\frac{z-p}{r_1}\right)\right)8\pi(1+\alpha)G(z,p)$$
$$+\chi\left(\frac{z-p}{r_1}\right)(v_{\rho,\tau_1+\lambda_1}(z-p)-\ln f(p)).$$

So $v(\rho,0)=U^1_\rho$ in $|z-p|<r_1$. For part (b) in Theorem 1.4, we need a three-parameter family of approximating functions and, for $(\lambda_2,a)\in\mathbb{R}\times\mathbb{C}$, $|\lambda_2|<\frac{1}{2}\tau_2$, $|a|<\frac{1}{2}$, and $g(z)=|z-p|^{2\alpha}f(z)$, we consider

$$\begin{aligned} v(\rho,\lambda_2,a)(z) \,=\, & \left(1-\chi\left(\frac{z-q-a}{r_2}\right)\right)8\pi G(z,q+a)\\ & +\chi\left(\frac{z-q-a}{r_2}\right)\left(v_{\rho,\tau_2+\lambda_2,\gamma}(z-q-a)-P_a(z)\right), \end{aligned}$$

where $P_a(z)=\ln g(q+a)+2\overline{z-q-a}\cdot\partial_z\ln g(q+a)+\overline{z-q-a}^2\cdot\partial_{zz}\ln g(q+a)$. So $v(\rho,0,0)=U^2_\rho$ in $|z-q|<r_2$. Finally, for part (c) in Theorem 1.4, we need a four-parameter family of approximating functions and, for $(\lambda,a)\in\mathbb{R}^2\times\mathbb{C}$, $\lambda=(\lambda_1,\lambda_2)$, $|\lambda|<\frac{1}{2}\min\{\tau_1,\tau_2\}$, $|a|<\frac{1}{2}$, and $g(z)=|z-p|^{2\alpha}f(z)$, we take

$$\begin{aligned} v(\rho,\lambda,a)(z) \,=\, & \left(1-\chi\left(\frac{z-p}{r_1}\right)\right)\left(8\pi(1+\alpha)G(z,p)+8\pi G(z,q+a)\right)\\ & +\chi\left(\frac{z-p}{r_1}\right)\left(v_{\rho,\tau_1+\lambda_1}(z-p)-\ln f(p)\right)\quad \text{in } B(p,1), \end{aligned}$$

$$\begin{aligned} v(\rho,\lambda,a)(z) \,=\, & \left(1-\chi\left(\frac{z-q-a}{r_2}\right)\right)\left(8\pi(1+\alpha)G(z,p)+8\pi G(z,q+a)\right)\\ & +\chi\left(\frac{z-q-a}{r_2}\right)\left(v_{\rho,\tau_2+\lambda_2,\gamma}(z-q-a)-P_a(z)\right)\quad \text{in } B(q,1), \end{aligned}$$

and

$$v(\rho,\lambda,a)(z)=8\pi(1+\alpha)G(z,p)+8\pi G(z,q+a)\qquad \text{in } \Omega\setminus(B(p,1)\cup B(q,1)).$$

To unify notation, from now on we will use the convention that

- $\lambda_2=0$, $a=0$ if $S=\{p\}$,
- $\lambda_1=0$ if $S=\{q\}$,
- every expression containing p (or q) does really exist only if $p\in S$ (or $q\in S$).

We remark that in such a way the last definition of $v(\rho,\lambda,a)$ contains the previous ones and (λ,a) always lie in $\mathbb{R}^2\times\mathbb{C}$.

3. A fixed point argument. In this section we obtain the desired existence result by means of a fixed point argument. To this end we have postponed the proof of the most technical aspects necessary to such an approach in the next two sections.

For $a\in\mathbb{C}$, $|a|<\frac{1}{2}$, it is possible to construct a diffeomorphism $\Psi(a,\cdot):\Omega\to\Omega$, smoothly depending on a, such that $\Psi(0,\cdot)=\mathrm{Id}$, $\Psi(a,z)=z-a$ for all $z\in B(q,\frac{3}{2})$, and $\Psi(a,\cdot)=\mathrm{Id}$ for all $z\in\Omega\setminus B(q,2)$. We can suppose that all derivatives of $\Psi(a,z)$ in a, $\bar a$, z, $\bar z$ up to order 3 are bounded in Ω.

We define now suitable function spaces of weighted Hölder type appropriate for our problem, which were introduced for the first time by Caffarelli, Hardt, and Simon in [9].

Definition 3.1. *For any $\nu\in\mathbb{R}$, $k\in\mathbb{N}$, $\beta\in[0,1]$, define the space*

$$C^{k,\beta}_\nu\left(B(0,1)\right):=\left\{w\in C^{k,\beta}\left(B(0,1)\setminus\{0\},\mathbb{R}\right)\,:\,\|w\|_{k,\beta,\nu}<+\infty\right\},$$

where

$$\|w\|_{k,\beta,\nu} := \sup_{r\le 1} r^{-\nu}\left\{\sup_{\{z:\ \frac{r}{2}<|z|<r\}}\left(\sum_{j=0}^{k} r^j|\nabla^j w(z)|\right)\right.$$
$$\left. + r^{k+\beta}\sup_{\{x,y:\ |x|,|y|\in(\frac{r}{2},r)\}}\left(\frac{|\nabla^k w(x)-\nabla^k w(y)|}{|x-y|^\beta}\right)\right\}.$$

Let $\nu_1 \in (0,1)$ and $\nu_2 \in (1,2)$ be two real numbers. Set $\tilde{\Omega} = \Omega \setminus B$, $B = B(p,1)\cup B(q,1)$, and define

$$X = \{w \in C^{2,\beta}(\Omega\setminus S,\mathbb{R}):\ w\equiv 0 \text{ on } \partial\Omega\,,\ \|w\|_X < +\infty\},$$

where $\|w\|_X = \|w\|_{2,\beta,\tilde{\Omega}} + \|w\|_{2,\beta,\nu_1,B(p,1)} + \|w\|_{2,\beta,\nu_2,B(q,1)}$, and

$$Y = \{w \in C^{0,\beta}(\Omega\setminus S,\mathbb{R}):\ \|w\|_Y < +\infty\},$$

where $\|w\|_Y = \|w\|_{0,\beta,\tilde{\Omega}} + \|w\|_{0,\beta,\nu_1-2,B(p,1)} + \|w\|_{0,\beta,\nu_2-2,B(q,1)}$.

We can replace the norm in X with an equivalent one (for ρ fixed) of the form

$$\|w\|_{X'} = \|w\|_{2,\beta,\tilde{\Omega}} + r_1^{\nu_1}\|w\|_{2,\beta,\nu_1,B(p,1)} + r_2^{\nu_2}\|w\|_{2,\beta,\nu_2,B(q,1)}$$

and we will refer to the space X, endowed with the norm $\|\cdot\|_{X'}$, as X'.

Finally, we define

$$\mathcal{E} = \{(w,\lambda,a) : w\in X,\ \lambda\in\mathbb{R}^2,\ a\in\mathbb{C}\}$$

with the norm $\|(w,\lambda,a)\|_{\mathcal{E}} = \|w\|_X + |\lambda| + |a|$, and $\mathcal{E}'$ as the space $\mathcal{E}$ endowed with the equivalent norm $\|(w,\lambda,a)\|_{\mathcal{E}'} = \|w\|_{X'} + |\lambda| + |a|$.

We can produce a solution $v(\rho,\lambda,a)+w\circ\Psi(a,\cdot)$, $(w,\lambda,a)\in\mathcal{E}'$, for problem $(P)_\rho$ if (w,λ,a) is a zero for the nonlinear map

$$N : \mathcal{E}' \to Y$$
$$(w,\lambda,a) \to N(w,\lambda,a) = \Delta\left[v(\rho,\lambda,a)+w\circ\Psi(a,\cdot)\right]\circ\Psi(a,\cdot)^{-1}$$
$$+\rho^2 g\circ\Psi(a,\cdot)^{-1}e^{v(\rho,\lambda,a)\circ\Psi(a,\cdot)^{-1}+w},$$

where $g(z) = |z-p|^{2\alpha}f(z)$. Define $L_{(0,\lambda,a)} : \mathcal{E}' \to Y$ as the linearized operator of N at $(0,\lambda,a)$. Hence,

$$L_{(0,\lambda,a)}(h,\sigma,b) = \Delta\left(h\circ\Psi(a,\cdot)\right)\circ\Psi(a,\cdot)^{-1} + \rho^2 g\circ\Psi(a,\cdot)^{-1}e^{v(\rho,\lambda,a)\circ\Psi(a,\cdot)^{-1}}h$$
$$+\sum_i \sigma_i\left[\Delta\partial_{\lambda_i}v(\rho,\lambda,a) + \rho^2 g(z)e^{v(\rho,\lambda,a)}\partial_{\lambda_i}v(\rho,\lambda,a)\right]\circ\Psi(a,\cdot)^{-1}$$
$$+2b\cdot\left[\Delta\partial_{\overline{a}}v(\rho,\lambda,a) + \rho^2 g(z)e^{v(\rho,\lambda,a)}\partial_{\overline{a}}v(\rho,\lambda,a)\right]\circ\Psi(a,\cdot)^{-1}$$
$$+2\partial_{\bar{z}}\left[\Delta v(\rho,\lambda,a) + \rho^2 g(z)e^{v(\rho,\lambda,a)}\right]|_{\Psi(a,\cdot)^{-1}}\cdot\left[\left(b\partial_a + \overline{b}\partial_{\overline{a}}\right)\Psi(a,\cdot)^{-1}\right].$$

In Theorem 4.13 below, we show that the map $L_{(0,0,0)} : \mathcal{E}' \to Y$ is uniformly invertible for ρ small. We can decompose

$$N(w,\lambda,a) - N(0,0,0) - L_{(0,0,0)}(w,\lambda,a)$$
$$= \left[N(w,\lambda,a) - N(0,\lambda,a) - L_{(0,\lambda,a)}(w,0,0)\right] + \left(L_{(0,\lambda,a)} - L_{(0,0,0)}\right)(w,0,0)$$
$$+\left[N(0,\lambda,a) - N(0,0,0) - L_{(0,0,0)}(0,\lambda,a)\right].$$

In the following three steps we estimate in Y each term above. For simplicity, we show only how to derive the estimates for the L^∞ part in $\|\cdot\|_Y$ since the estimates of the Hölder term can be established in a similar way: all along the paper we will use implicitly this fact to simplify all the computations.

Step 1. Let

$$\begin{aligned} f_1(w,\lambda,a) &= N(w,\lambda,a) - N(0,\lambda,a) - L_{(0,\lambda,a)}(w,0,0) \\ &= \rho^2 g \circ \Psi(a,\cdot)^{-1} e^{v(\rho,\lambda,a)\circ\Psi(a,\cdot)^{-1}} (e^w - 1 - w). \end{aligned}$$

Since

$$\Psi(a,\cdot)^{-1}: \begin{array}{ll} B(q,1) \to B(q+a,1) \\ z \qquad\quad \to z+a, \end{array}$$

we obtain the bounds

(14)
$$|\rho^2 g \circ \Psi(a,\cdot)^{-1} e^{v(\rho,\lambda,a)\circ\Psi(a,\cdot)^{-1}}| = \begin{cases} O\left(\frac{\rho^2|z-p|^{2\alpha}}{(\rho^2+|z-p|^{2\alpha+2})^2}\right) & \text{in } B(p,r_1), \\ O\left(\frac{\rho^2}{|z-p|^{2\alpha+4}}\right) & \text{in } B(p,1)\setminus B(p,r_1), \\ O\left(\frac{\rho^2}{(\rho^2+|z-q|^2)^2}\right) & \text{in } B(q,r_2), \\ O\left(\frac{\rho^2}{|z-q|^4}\right) & \text{in } B(q,1)\setminus B(q,r_2), \\ O(\rho^2) & \text{in } \tilde\Omega, \end{cases}$$

(15)
$$|\partial_{\overline a}\, \rho^2 g \circ \Psi(a,\cdot)^{-1} e^{v(\rho,\lambda,a)\circ\Psi(a,\cdot)^{-1}}| = \begin{cases} O\left(\frac{\rho^2}{(\rho^2+|z-q|^2)^2}\right) & \text{in } B(q,r_2), \\ O\left(\frac{\rho^2}{|z-q|^4}\right) & \text{in } B(q,1)\setminus B(q,r_2), \\ O(\rho^2) & \text{in } \Omega\setminus B(q,1), \end{cases}$$

(16)
$$|\partial_{\lambda_i}\, \rho^2 g \circ \Psi(a,\cdot)^{-1} e^{v(\rho,\lambda,a)\circ\Psi(a,\cdot)^{-1}}| = \begin{cases} O\left(\frac{\rho^2|z-p|^{2\alpha}}{(\rho^2+|z-p|^{2\alpha+2})^2}\right) & \text{in } B(p,2r_1) \text{ if } i=1, \\ O\left(\frac{\rho^2}{(\rho^2+|z-q|^2)^2}\right) & \text{in } B(q,2r_2) \text{ if } i=2, \\ 0 & \text{elsewhere.} \end{cases}$$

Hence, we can derive

$$\begin{aligned} &\|f_1(w_1,\lambda_1,a_1) - f_1(w_2,\lambda_2,a_2)\|_{0,\beta,\tilde\Omega} \\ &\quad = O\Big[\rho^2(\|w_1\|_{2,\beta,\tilde\Omega} + \|w_2\|_{2,\beta,\tilde\Omega})\|(w_1,\lambda_1,a_1) - (w_2,\lambda_2,a_2)\|_{\mathcal E}\Big], \\ &\|f_1(w_1,\lambda_1,a_1) - f_1(w_2,\lambda_2,a_2)\|_{0,\beta,\nu_1-2,B(p,1)} \\ &\quad = O\Big[\rho^{\frac{\nu_1}{\alpha+1}}(\|w_1 - w_2\|_{2,\beta,\nu_1,B(p,1)} \\ &\qquad + |\lambda_1-\lambda_2| + |a_1-a_2|)\left(\|w_1\|_{2,\beta,\nu_1,B(p,1)} + \|w_2\|_{2,\beta,\nu_1,B(p,1)}\right)\Big], \end{aligned}$$

and

$$\begin{aligned}&\|f_1(w_1,\lambda_1,a_1)-f_1(w_2,\lambda_2,a_2)\|_{0,\beta,\nu_2-2,B(q,1)}\\&\quad=O\left[\rho^{\nu_2}\left(\|w_1-w_2\|_{2,\beta,\nu_2,B(q,1)}\right.\right.\\&\qquad\left.\left.+|\lambda_1-\lambda_2|+|a_1-a_2|\right)\left(\|w_1\|_{2,\beta,\nu_2,B(q,1)}+\|w_2\|_{2,\beta,\nu_2,B(q,1)}\right)\right].\end{aligned}$$

Since $\frac{\rho^2}{r_1^{4\alpha+5}}+\frac{\rho^2}{r_2^5}=O(1)$, finally we get

$$\begin{aligned}\|f_1(w_1,\lambda_1,a_1)-f_1(w_2,\lambda_2,a_2)\|_Y=O\left[\|(w_1,\lambda_1,a_1)-(w_2,\lambda_2,a_2)\|_{\mathcal{E}'}\right.\\\left.\times\left(\|(w_1,\lambda_1,a_1)\|_{\mathcal{E}'}+\|(w_2,\lambda_2,a_2)\|_{\mathcal{E}'}\right)\right].\end{aligned}$$

Step 2. Define

$$f_2(w,\lambda,a)=\left(L_{(0,\lambda,a)}-L_{(0,0,0)}\right)(w,0,0)=f_2^1(w,a)+f_2^2(w,\lambda,a),$$

where

$$f_2^1(w,a)=\Delta\left[w\circ\Psi(a,\cdot)\right]\circ\Psi(a,\cdot)^{-1}-\Delta w$$

and

$$f_2^2(w,\lambda,a)=\rho^2 g\circ\Psi(a,\cdot)^{-1}e^{v(\rho,\lambda,a)\circ\Psi(a,\cdot)^{-1}}w-\rho^2 g e^{v(\rho,0,0)}w.$$

Using the identities

$$\begin{aligned}\partial_{\bar z}(w\circ\Psi)&=(\partial_z w\circ\Psi)\,\partial_{\bar z}\Psi+(\partial_{\bar z}w\circ\Psi)\,\partial_{\bar z}\bar\Psi,\\\Delta(w\circ\Psi)&=(\Delta w\circ\Psi)\left[|\partial_z\Psi|^2+|\partial_{\bar z}\Psi|^2\right]+8\mathrm{Re}\left[(\partial_{zz}w\circ\Psi)\,\partial_z\Psi\partial_{\bar z}\Psi\right]\\&\quad+8\mathrm{Re}\left[(\partial_z w\circ\Psi)\,\partial_{z\bar z}\Psi\right],\end{aligned}$$

we obtain

$$\begin{aligned}f_2^1(w_1,a_1)-f_2^1(w_2,a_2)&=\Delta(w_1-w_2)\left(|\partial_z\Psi(a_1,\cdot)|^2+|\partial_{\bar z}\Psi(a_1,\cdot)|^2-1\right)|_{\Psi(a_1,\cdot)^{-1}}\\&\quad+8\mathrm{Re}\left(\partial_{zz}(w_1-w_2)\partial_z\Psi(a_1,\cdot)|_{\Psi(a_1,\cdot)^{-1}}\,\partial_{\bar z}\Psi(a_1,\cdot)\,|_{\Psi(a_1,\cdot)^{-1}}\right)\\&\quad+8\mathrm{Re}\left(\partial_z(w_1-w_2)\partial_{z\bar z}\Psi(a_1,\cdot)\,|_{\Psi(a_1,\cdot)^{-1}}\right)\\&\quad+\Delta\left[w_2\circ\Psi(a_1,\cdot)\right]\circ\Psi(a_1,\cdot)^{-1}-\Delta\left[w_2\circ\Psi(a_2,\cdot)\right]\circ\Psi(a_2,\cdot)^{-1}.\end{aligned}$$

Therefore

$$\begin{aligned}f_2^1(w_1,a_1)-f_2^1(w_2,a_2)&=O\left(|D^2(w_1-w_2)||a_1|+|\nabla(w_1-w_2)||a_1|\right.\\&\quad\left.+|D^2w_2||a_1-a_2|+|\nabla w_2||a_1-a_2|\right)\end{aligned}$$

in $B(q,2)\setminus B(q,1)$ and $f_2^1(w_1,a_1)-f_2^1(w_2,a_2)=0$ outside this region. Hence

$$\begin{aligned}\|f_2^1(w_1,a_1)-f_2^1(w_2,a_2)\|_Y=O\left[\|(w_1,\lambda_1,a_1)-(w_2,\lambda_2,a_2)\|_{\mathcal{E}'}\right.\\\left.\times\left(\|(w_1,\lambda_1,a_1)\|_{\mathcal{E}'}+\|(w_2,\lambda_2,a_2)\|_{\mathcal{E}'}\right)\right].\end{aligned}$$

By (14), (15), and (16), for f_2^2 we get

$$\begin{aligned}&\|f_2^2(w_1,\lambda_1,a_1)-f_2^2(w_2,\lambda_2,a_2)\|_Y\\&\quad=O\left[\left(\|w_1\|_X+\|w_2\|_X\right)\|(w_1,\lambda_1,a_1)-(w_2,\lambda_2,a_2)\|_{\mathcal{E}'}\right.\\&\qquad\left.+\|w_1-w_2\|_X\left(\|(w_1,\lambda_1,a_1)\|_{\mathcal{E}'}+\|(w_2,\lambda_2,a_2)\|_{\mathcal{E}'}\right)\right].\end{aligned}$$

Step 3. Set $f_3(\lambda,a)=N(0,\lambda,a)-N(0,0,0)-L_{(0,0,0)}(0,\lambda,a)$. Let us write explicitly $N(0,\lambda,a)$ in $B(p,1)$,

$$\begin{aligned}N(0,\lambda,a)(z)=&\frac{1}{r_1^2}\Delta\chi\left(\frac{z-p}{r_1}\right)\Delta^1(z)+\frac{8}{r_1}\partial_{\bar z}\chi\left(\frac{z-p}{r_1}\right)\cdot\partial_{\bar z}\Delta^1(z)\\&-\chi\left(\frac{z-p}{r_1}\right)\rho^2|z-p|^{2\alpha}e^{v_{\rho,\tau_1+\lambda_1}(z-p)}\\&+\rho^2|z-p|^{2\alpha}f(z)e^{8\pi(1+\alpha)G(z,p)+8\pi G(z,q+a)+\chi(\frac{z-p}{r_1})\Delta^1(z)}\end{aligned}$$

and in $B(q,1)$

$$\begin{aligned}N(0,\lambda,a)(z)=&\frac{1}{r_2^2}\Delta\chi\left(\frac{z-q}{r_2}\right)\Delta^2(z)+\frac{8}{r_2}\partial_{\bar z}\chi\left(\frac{z-q}{r_2}\right)\cdot\partial_{\bar z}\Delta^2(z)\\&-\chi\left(\frac{z-q}{r_2}\right)\rho^2e^{v_{\rho,\tau_2+\lambda_2,\gamma}(z-q)}\\&+\rho^2g(z+a)e^{8\pi(1+\alpha)G(z+a,p)+8\pi G(z+a,q+a)+\chi(\frac{z-q}{r_2})\Delta^2(z)},\end{aligned}$$

where $g(z)=|z-p|^{2\alpha}f(z)$ and

$$\begin{aligned}&\Delta^1(z):=v_{\rho,\tau_1+\lambda_1}(z-p)-\ln f(p)-8\pi(1+\alpha)G(z,p)-8\pi G(z,q+a)\text{ in }B(p,1),\\&\Delta^2(z):=v_{\rho,\tau_2+\lambda_2,\gamma}(z-q)-P_a(z+a)-8\pi(1+\alpha)G(z+a,p)\\&\qquad\quad-8\pi G(z+a,q+a)\text{ in }B(q,1).\end{aligned}$$

In $B(p,1)$ we get

$$\begin{aligned}\partial_{\lambda_s\lambda_k}N(0,\lambda,a)(z)=&\frac{1}{r_1^2}\Delta\chi\left(\frac{z-p}{r_1}\right)v^1_{sk}(z)\delta_{s1}\delta_{k1}+\frac{8}{r_1}\partial_{\bar z}\chi\left(\frac{z-p}{r_1}\right)\cdot\partial_{\bar z}v^1_{sk}(z)\delta_{s1}\delta_{k1}\\&-\chi\left(\frac{z-p}{r_1}\right)\rho^2|z-p|^{2\alpha}e^{v_{\rho,\tau_1+\lambda_1}(z-p)}\left(v^1_s(z)v^1_k(z)+v^1_{sk}(z)\right)\delta_{s1}\delta_{k1}\\&+\rho^2|z-p|^{2\alpha}f(z)e^{8\pi(1+\alpha)G(z,p)+8\pi G(z,q+a)+\chi(\frac{z-p}{r_1})\Delta^1(z)}\\&\times\chi\left(\frac{z-p}{r_1}\right)\left(v^1_{sk}(z)+\chi\left(\frac{z-p}{r_1}\right)v^1_s(z)v^1_k(z)\right)\delta_{s1}\delta_{k1},\end{aligned}$$

$$\begin{aligned}\partial_{a\lambda_k}N(0,\lambda,a)(z)=&\rho^2|z-p|^{2\alpha}f(z)e^{v_{\rho,\tau_1+\lambda_1}(z-p)-\ln f(p)+(\chi(\frac{z-p}{r_1})-1)\Delta^1(z)}\\&\times\chi\left(\frac{z-p}{r_1}\right)\left(\chi\left(\frac{z-p}{r_1}\right)-1\right)v^1_k(z)\partial_a\Delta^1(z)\delta_{k1},\end{aligned}$$

and

$$\begin{aligned}\partial_{aa}N(0,\lambda,a)(z)=&\frac{1}{r_1^2}\Delta\chi\left(\frac{z-p}{r_1}\right)\partial_{aa}\Delta^1(z)+\frac{8}{r_1}\partial_{\bar z}\chi\left(\frac{z-p}{r_1}\right)\cdot\partial_{aa\bar z}\Delta^1(z)\\&+\rho^2|z-p|^{2\alpha}f(z)e^{v_{\rho,\tau_1+\lambda_1}(z-p)-\ln f(p)+(\chi(\frac{z-p}{r_1})-1)\Delta^1(z)}\\&\times\left\{\left(\chi\left(\frac{z-p}{r_1}\right)-1\right)\partial_{aa}\Delta^1(z)+\left(\chi\left(\frac{z-p}{r_1}\right)-1\right)^2\partial_a\Delta^1(z)\partial_a\Delta^1(z)\right\},\end{aligned}$$

where $v^1_j(z):=\partial_{\lambda_j}v_{\rho,\tau_1+\lambda_1}(z-p)$ and $v^1_{jm}(z):=\partial_{\lambda_j\lambda_m}v_{\rho,\tau_1+\lambda_1}(z-p)$ for any j,m.

Similarly, in $B(q,1)$ we get

$$\begin{aligned}\partial_{\lambda_s\lambda_k}N(0,\lambda,a)(z) &= \frac{1}{r_2^2}\Delta\chi\left(\frac{z-q}{r_2}\right)v_{sk}^2(z)\delta_{s2}\delta_{k2}+\frac{8}{r_2}\partial_{\bar z}\chi\left(\frac{z-q}{r_2}\right)\cdot\partial_{\bar z}v_{sk}^2(z)\delta_{s2}\delta_{k2}\\ &\quad-\chi\left(\frac{z-q}{r_2}\right)\rho^2e^{v_{\rho,\tau_2+\lambda_2,\gamma}(z-q)}\left(v_s^2(z)v_k^2(z)+v_{sk}^2(z)\right)\delta_{s2}\delta_{k2}\\ &\quad+\rho^2g(z+a)e^{8\pi(1+\alpha)G(z+a,p)+8\pi G(z+a,q+a)+\chi(\frac{z-q}{r_2})\Delta^2(z)}\\ &\quad\times\chi\left(\frac{z-q}{r_2}\right)\left(v_{sk}^2(z)+\chi\left(\frac{z-q}{r_2}\right)v_s^2(z)v_k^2(z)\right)\delta_{s2}\delta_{k2},\end{aligned}$$

$$\begin{aligned}\partial_{a\lambda_k}N(0,\lambda,a)(z) &= \rho^2e^{v_{\rho,\tau_2+\lambda_2,\gamma}(z-q)+(\chi(\frac{z-q}{r_2})-1)\Delta^2(z)}\chi\left(\frac{z-q}{r_2}\right)v_k^2(z)\delta_{k2}\\ &\quad\times\Bigg[\left(\chi\left(\frac{z-q}{r_2}\right)-1\right)\partial_a\Delta^2(z)g(z+a)e^{-P_a(z+a)}\\ &\quad+\partial_a\left(g(z+a)e^{-P_a(z+a)}\right)\Bigg],\end{aligned}$$

and

$$\begin{aligned}\partial_{aa}N(0,\lambda,a)(z) &= \frac{1}{r_2^2}\Delta\chi\left(\frac{z-q}{r_2}\right)\partial_{aa}\Delta^2(z)+\frac{8}{r_2}\partial_{\bar z}\chi\left(\frac{z-q}{r_2}\right)\cdot\partial_{aa\bar z}\Delta^2(z)\\ &\quad+\rho^2e^{v_{\rho,\tau_2+\lambda_2,\gamma}(z-q)+(\chi(\frac{z-q}{r_2})-1)\Delta^2(z)}\Bigg\{\partial_{aa}\left(g(z+a)e^{-P_a(z+a)}\right)\\ &\quad+\left(\chi\left(\frac{z-q}{r_2}\right)-1\right)\Big[2\partial_a\Delta^2(z)\partial_a\Big(g(z+a)e^{-P_a(z+a)}\Big)\\ &\qquad+\partial_{aa}\Delta^2(z)g(z+a)e^{-P_a(z+a)}\Big]\\ &\quad+\left(\chi\left(\frac{z-q}{r_2}\right)-1\right)^2(\partial_a\Delta^2(z))^2g(z+a)e^{-P_a(z+a)}\Bigg\},\end{aligned}$$

where $v_j^2(z):=\partial_{\lambda_j}v_{\rho,\tau_2+\lambda_2,\gamma}(z-q)$ and $v_{jm}^2(z):=\partial_{\lambda_j\lambda_m}v_{\rho,\tau_2+\lambda_2,\gamma}(z-q)$ for any j,m.

We have that

$$|\partial_\tau v_{\rho,\tau,\lambda}(z)|+|\partial_{\tau\tau}v_{\rho,\tau,\lambda}(z)|+|z||\nabla\partial_{\tau\tau}v_{\rho,\tau,\lambda}(z)|=O(1)$$

for $\lambda\in\{0,\gamma\}$, $f(z)e^{-\ln f(p)}=1+O(|z-p|)$, $g(z+a)e^{-P_a(z+a)}=1+O(|z-q|^2)$, and

$$|\partial_a\Delta^1(z)|+|\partial_{aa}\Delta^1(z)|+|z-p||\partial_{aa\bar z}\Delta^1(z)|=O(1)\quad\text{ in } B(p,1),$$

$$|\partial_a\Delta^2(z)|+|\partial_{aa}\Delta^2(z)|+|z-q||\partial_{aa\bar z}\Delta^2(z)|=O(1)\quad\text{ in } B(q,1),$$

$$\left|\partial_a\left(g(z+a)e^{-P_a(z+a)}\right)\right|+\left|\partial_{aa}\left(g(z+a)e^{-P_a(z+a)}\right)\right|=O(|z-q|^2).$$

So we can derive $\|\partial^2N(0,\lambda,a)\|_{0,\beta,\nu_1-2,B(p,\frac12)}=O(r_1^{-\nu_1})$, $\|\partial^2N(0,\lambda,a)\|_{0,\beta,\nu_2-2,B(q,\frac12)}=O(r_2^{-\nu_2})$, where ∂^2 denotes some second-order derivative of $N(0,\lambda,a)$ in the variables λ and a. Since $\|\partial^2N(0,\lambda,a)\|_{0,\beta}=O(\rho^2)$ in $\tilde\Omega$, we conclude that $\|\partial^2N(0,\lambda,a)\|_Y=O(\sum_{i=1}^2r_i^{-\nu_i})$. Finally, we obtain

$$\|f_3(\lambda_1,a_1)-f_3(\lambda_2,a_2)\|_Y=O\left(\left(\sum_{i=1}^2r_i^{-\nu_i}\right)\|(w_1,\lambda_1,a_1)-(w_2,\lambda_2,a_2)\|_{\mathcal{E}'}^2\right).$$

Step 4. We define

$$K : \mathcal{E}' \to \mathcal{E}'$$
$$(w, \lambda, a) \to -L^{-1}_{(0,0,0)} \left[N(0,0,0) + \big(N(w,\lambda,a) - N(0,0,0) - L_{(0,0,0)}(w,\lambda,a) \big) \right].$$

Let us remark that (w, λ, a) is a zero for $N \Leftrightarrow (w, \lambda, a)$ is a fixed point for K. Summarizing the previous steps and by means of the uniform estimates derived in Theorem 4.13 below, we have

$$\begin{aligned} &\|K(w_1,\lambda_1,a_1) - K(w_2,\lambda_2,a_2)\|_{\mathcal{E}'} \\ &\quad \le C_0 \left(\sum_{i=1}^{2} r_i^{-\nu_i} \right) (\|(w_1,\lambda_1,a_1)\|_{\mathcal{E}'} + \|(w_2,\lambda_2,a_2)\|_{\mathcal{E}'}) \\ &\quad \times \|(w_1,\lambda_1,a_1) - (w_2,\lambda_2,a_2)\|_{\mathcal{E}'} \end{aligned}$$

for some constant $C_0 > 0$, where we have taken into account that

$$\|w\|_X \le \left(\sum_{i=1}^{2} r_i^{-\nu_i} \right) \cdot \|(w,\lambda,a)\|_{\mathcal{E}'}.$$

We can choose $\nu_1 \in (0,1)$ and $\nu_2 \in (1,2)$ in such a way that $(\nu_1, 1-\nu_1) \cap (\nu_2 - 1, 2-\nu_2) \neq \emptyset$ and let us fix some $\delta > 0$ in this set. Define $\sigma = \frac{4\alpha+5}{2\nu_1} + 1$, $r_i = \rho^{\frac{1}{\sigma\nu_i}}$, and note that $N(0,0,0) = \eta$ where η is the error term defined and estimated in section 5. In fact, from the technical estimates contained in sections 4 and 5 we see that $\|L^{-1}_{(0,0,0)}\eta\|_{\mathcal{E}'} = O(r_1^{1-\delta} + r_2^{2-\delta})$ (see (37) below), and we get

$$\|K(w,\lambda,a)\|_{\mathcal{E}'} \le C_1 \left[\left(\sum_{i=1}^{2} r_i^{-\nu_i} \right) \|(w,\lambda,a)\|^2_{\mathcal{E}'} + r_1^{1-\delta} + r_2^{2-\delta} \right]$$

for some constant $C_1 > 0$, where we have used the fact that $\|K(w,\lambda,a) - K(0,0,0)\|_{\mathcal{E}'} \le C_0(\sum_{i=1}^{2} r_i^{-\nu_i})\|(w,\lambda,a)\|^2_{\mathcal{E}'}$. Thus, the suitable choice of r_1, r_2, as expressed by property (38) below, allows us to conclude that for ρ small the map K is a contraction of the space

$$\mathcal{E}' \cap \left\{ (w,\lambda,a) : \|(w,\lambda,a)\|_{\mathcal{E}'} \le 2C_1 \left(r_1^{1-\delta} + r_2^{2-\delta} \right) \right\}$$

into itself. So there exists a unique fixed point $(w^\rho, \lambda^\rho, a^\rho)$ of the map K for $0 < \rho < \rho_0$, $\rho_0 > 0$ small, such that

$$\|w^\rho\|_{2,\beta,\tilde{\Omega}} + r_1^{\nu_1}\|w^\rho\|_{2,\beta,\nu_1,B(p,1)} + r_2^{\nu_2}\|w^\rho\|_{2,\beta,\nu_2,B(q,1)} + |\lambda^\rho| + |a^\rho| \le 2C_1 \left(r_1^{1-\delta} + r_2^{2-\delta} \right).$$

Hence $v_\rho = v(\rho, \lambda^\rho, a^\rho) + w^\rho \circ \Psi(a^\rho, \cdot)$ is the solution we are looking for in Theorem 1.4. It admits the desired properties in view of the definition of $v(\rho,\lambda,a)$, the fact that $\frac{\rho^2}{r_1^{2\alpha+4}} + \frac{\rho^2}{r_2^4} \to 0$ as $\rho \to 0$ and $w^\rho \to 0$ uniformly in Ω and in $C^{2,\beta}_{loc}(\Omega \setminus S)$, as follows by (38).

4. Invertibility of the linearized operator $L_{(0,0,0)}$.

4.1. Some local operator. The radial case. We are interested in studying the linearized operator of the equation

$$-\Delta v = \rho^2 |z|^{2\alpha} e^v \quad \text{in } B(0,1)\,, \ \alpha \geq 0 \tag{17}$$

about the radial solutions $v_{\rho,\tau}$ defined in (9) in case either $\alpha = 0$ or $\alpha \notin \mathbb{N}$. We define the linearized operator about $v_{\rho,\tau}$ by setting

$$L_{\rho,\tau} w = \Delta w + \rho^2 |z|^{2\alpha} e^{v_{\rho,\tau}} w$$

and we investigate the invertibility of $L_{\rho,\tau}$ under Dirichlet boundary condition. Inspired by the work of Caffarelli, Hardt, and Simon in [9] also used in [4], we have the following result.

PROPOSITION 4.1. *Let $\alpha \notin \mathbb{N}$. For all $\nu \in (0,1)$ and $\tau > 0$, there exist $\rho_0 > 0$, a continuous linear form $H^0_{\rho,\tau} : C^{0,\beta}_{\nu-2}(B(0,1)) \to \mathbb{R}$, and a linear operator $G_{\rho,\tau} : C^{0,\beta}_{\nu-2}(B(0,1)) \to C^{2,\beta}_{\nu}(B(0,1))$, uniformly bounded for $0 < \rho < \rho_0$, such that for all $\rho \in (0,\rho_0)$ and for all $f \in C^{0,\beta}_{\nu-2}(B(0,1))$ there exists a unique bounded solution w of*

$$\begin{cases} L_{\rho,\tau} w = f & \text{in } B(0,1), \\ w = 0 & \text{on } \partial B(0,1) \end{cases} \tag{18}$$

which can be uniquely decomposed as follows:

$$w(z) = G_{\rho,\tau}(f)(z) + H^0_{\rho,\tau}(f) \frac{\tau^2\rho^2 - |z|^{2(\alpha+1)}}{\tau^2\rho^2 + |z|^{2(\alpha+1)}}.$$

Moreover, $H^0_{\rho,\tau}(f) = 0$ for any f such that $\int_0^{2\pi} f(re^{i\theta})d\theta = 0$ for all $r \in (0,1]$.

PROPOSITION 4.2. *Let $\alpha = 0$, $\nu \in (1,2)$, and $\tau > 0$. There exist $\rho_0 > 0$, two continuous linear forms $H^0_{\rho,\tau} : C^{0,\beta}_{\nu-2}(B(0,1)) \to \mathbb{R}$, $H^1_{\rho,\tau} : C^{0,\beta}_{\nu-2}(B(0,1)) \to \mathbb{C}$, and a linear operator $G_{\rho,\tau} : C^{0,\beta}_{\nu-2}(B(0,1)) \to C^{2,\beta}_{\nu}(B(0,1))$, uniformly bounded for $0 < \rho < \rho_0$, such that for all $\rho \in (0,\rho_0)$ and for all $f \in C^{0,\beta}_{\nu-2}(B(0,1))$ there exists a unique bounded solution w of*

$$\begin{cases} L_{\rho,\tau} w = f & \text{in } B(0,1), \\ w = 0 & \text{on } \partial B(0,1) \end{cases}$$

which can be uniquely decomposed as follows:

$$w(z) = G_{\rho,\tau}(f)(z) + H^0_{\rho,\tau}(f) \frac{\tau^2\rho^2 - |z|^2}{\tau^2\rho^2 + |z|^2} + 2H^1_{\rho,\tau}(f) \cdot \frac{z}{\tau^2\rho^2 + |z|^2}.$$

Moreover, $H^0_{\rho,\tau}(f) = 0$, $H^1_{\rho,\tau}(f) = 0$ for any f such that $\int_0^{2\pi} f(re^{i\theta})d\theta = 0$ and $\int_0^{2\pi} f(re^{i\theta})e^{-i\theta}d\theta = 0$ for all $r \in (0,1]$.

By the Liouville formula (6), we get that, for any $j \in \mathbb{Z}$ and $|a| < \frac{\alpha+1}{|j|+\alpha+1}$,

$$\ln \frac{8(\alpha+1)^2\tau^2|1 + \frac{j+\alpha+1}{\alpha+1} a z^j|^2}{(\tau^2\rho^2 + |z|^{2(\alpha+1)}|1 + az^j|^2)^2}$$

solves (17). Hence by taking its derivative with respect to a, evaluated at $a = 0$, we obtain a solution of $L_{\rho,\tau} w = 0$ in the form

$$\frac{1}{\alpha+1} \frac{(j+\alpha+1)\tau^2\rho^2 + (j-\alpha-1)|z|^{2(\alpha+1)}}{\tau^2\rho^2 + |z|^{2(\alpha+1)}} z^j, \quad j \in \mathbb{Z}.$$

Consequently,

$$a_j(r) := \frac{(j+\alpha+1)\tau^2\rho^2 + (j-\alpha-1)r^{2(\alpha+1)}}{\tau^2\rho^2 + r^{2(\alpha+1)}} r^j, \quad j \in \mathbb{Z}$$

is a solution for the ordinary differential equation

$$\ddot{a}_j + \frac{1}{r}\dot{a}_j - \frac{j^2}{r^2} a_j + \frac{8(\alpha+1)^2\tau^2\rho^2 r^{2\alpha}}{(\tau^2\rho^2 + r^{2(\alpha+1)})^2} a_j = 0 \quad \text{in } (0,1).$$

Let us remark that for $j > 0$, $\{a_j(r), a_{-j}(r)\}$ is a set of linearly independent solutions for the same homogeneous equation. Hence any other solution is obtained as a linear combination of $a_j(r)$ and $a_{-j}(r)$. For $j = 0$, another independent solution can be explicitly found and it behaves like $\ln r$ as $r \to 0$. Since $\ln r$ and $a_{-j}(r)$, $j > 0$, are not bounded in a neighborhood of $r = 0$ and $a_j(1) \neq 0$, $j \geq 0$, by means of Fourier decomposition, it is easy to derive the following lemma.

LEMMA 4.3. *Let w be a bounded solution of*

$$\begin{cases} L_{\rho,\tau} w = 0 & \text{in } B(0,1), \\ w = 0 & \text{on } \partial B(0,1). \end{cases}$$

Then $w = 0$.

We decompose w and f into Fourier series:

$$w(z) = w_0(r) + 2\sum_{j=1}^{+\infty} w_j(r) \cdot e^{-ij\theta}, \quad f(z) = f_0(r) + 2\sum_{j=1}^{+\infty} f_j(r) \cdot e^{-ij\theta}.$$

So problem (18) becomes equivalent to

$$\text{(Pj)} \qquad \begin{cases} \ddot{w}_j + \frac{1}{r}\dot{w}_j - \frac{j^2}{r^2} w_j + \frac{8(\alpha+1)^2\tau^2\rho^2 r^{2\alpha}}{(\tau^2\rho^2 + r^{2(\alpha+1)})^2} w_j = f_j, \quad \text{in } (0,1), \\ w_j(1) = 0 \end{cases}$$

for $j \in \mathbb{N}$. Set $j_\alpha = \min\{j \in \mathbb{N} : j > \alpha + 1\}$ and $m_\alpha = \max\{j \in \mathbb{N} : j < \alpha + 1\}$.

Step 1. By the variation of constants formula, for $j \geq j_\alpha$ and $\nu > -j$,

$$w_j(r) = \left(\int_1^r \frac{ds}{s a_j^2(s)} \int_0^s t a_j(t) f_j(t) dt \right) a_j(r), \quad r > 0$$

defines a solution of (Pj). Since $0 < j - \alpha - 1 \leq \frac{(j+\alpha+1)\tau^2\rho^2 + (j-\alpha-1)r^{2(\alpha+1)}}{\tau^2\rho^2 + r^{2(\alpha+1)}} \leq j + \alpha + 1$, we have that for $-j < \nu < j$ $r^{-\nu}|w_j(r)| \leq \frac{(j+\alpha+1)^2}{(j-\alpha-1)^2} \frac{1}{j^2-\nu^2} \|f_j\|_{0,\beta,\nu-2}$ and, by classical rescaled Schauder estimates (see [22]), we find

$$\|w_j\|_{2,\beta,\nu} \leq C \frac{(j+\alpha+1)^2}{(j-\alpha-1)^2} \frac{1}{j^2-\nu^2} \|f_j\|_{0,\beta,\nu-2}$$

for suitable $C > 0$. Finally, for $-j_\alpha < \nu < j_\alpha$, we can define $h(z) = 2\sum_{j=j_\alpha}^{+\infty} w_j(r) \cdot e^{-ij\theta}$ and, since $\|f_j\|_{0,\beta,\nu-2} \le \|f\|_{0,\beta,\nu-2}$, there holds the estimate

$$\sum_{j=j_\alpha}^{+\infty} \|w_j\|_{2,\beta,\nu} \le C \left(\sum_{j=j_\alpha}^{+\infty} \frac{(j+\alpha+1)^2}{(j-\alpha-1)^2} \frac{1}{j^2-\nu^2} \right) \|f\|_{0,\beta,\nu-2}.$$

So $h(z)$ is a well-defined function in $C_\nu^{2,\beta}(B(0,1))$ and satisfies

$$\begin{cases} L_{\rho,\tau} h = 2\sum_{j=j_\alpha}^{+\infty} f_j(r) \cdot e^{-ij\theta} & \text{in } B(0,1), \\ h = 0 & \text{on } \partial B(0,1) \end{cases}$$

together with the estimate $\|h\|_{2,\beta,\nu} \le C\|f\|_{0,\beta,\nu-2}$ for $-j_\alpha < \nu < j_\alpha$.

Step 2. For $0 < j \le m_\alpha$, $\nu > -j$, and $r > \bar r := (\frac{j+\alpha+1}{\alpha+1-j}\tau^2\rho^2)^{\frac{1}{2\alpha+2}}$, it is possible to define

$$\tilde w_j(r) = \left(\int_1^r \frac{ds}{s a_j^2(s)} \int_0^s t a_j(t) f_j(t) dt \right) a_j(r).$$

Note that $a_j(\bar r) = 0$ and hence $\tilde w_j(r)$ is not well defined up to $\bar r$. To be able to obtain an extension of $\tilde w_j$ for $r \le \bar r$, define $\psi_j(s,\rho) = (s-\bar r)^2 \frac{1}{s a_j^2(s)} \int_0^s t a_j(t) f_j(t) dt$ and set

$$w_j(r) = a_j(r) \left[\int_1^r \frac{\psi_j(s,\rho) - \psi_j(\bar r,\rho)}{(s-\bar r)^2} ds - \frac{1-r}{(1-\bar r)(r-\bar r)} \psi_j(\bar r,\rho) \right]. \tag{19}$$

The function w_j is well defined also for $r \le \bar r$ and gives an extension of $\tilde w_j$. We will refer to the first and second terms in the expression of $w_j(r)$ above as $w_j^1(r)$ and $w_j^2(r)$, respectively. Since for $0 < r < \bar r - \delta$, $\delta > 0$, we have

$$\begin{aligned} w_j(r) = a_j(r) \Bigg[& \int_1^{\bar r - \delta} \frac{\psi_j(s,\rho) - \psi_j(\bar r,\rho)}{(s-\bar r)^2} ds \\ & + \int_{\bar r - \delta}^r \frac{ds}{s a_j(s)^2} \int_0^s t a_j(t) f_j(t) dt + \left(\frac{1}{1-\bar r} + \frac{1}{\delta} \right) \psi_j(\bar r,\rho) \Bigg], \end{aligned}$$

the function $w_j(r)$ does solve (Pj) for $r > 0$. Note that for $\nu < j$, we have that $\sup_{r\in(\bar r,1)} r^{-\nu} |w_j(r)| \le C\|f_j\|_{0,\beta,\nu-2}$. In fact, for $r \ge 2\bar r$ we find $|w_j(r)| = |\tilde w_j(r)| \le C\|f_j\|_{0,\beta,\nu-2} r^\nu$ as $\frac{1}{a_j^2(s)} = O(\frac{1}{s^{2j}})$ for $s \ge r$. While for $\bar r \le r \le 2\bar r$ there holds

$$\begin{aligned} |w_j(r)| = |\tilde w_j(r)| &\le C\|f_j\|_{0,\beta,\nu-2} \left(\frac{r}{\bar r} - 1 \right) \bar r^j \int_r^1 \frac{(\tau^2\rho^2 + s^{2(\alpha+1)})^2 s^{\nu-j-1} ds}{[(j+\alpha+1)\tau^2\rho^2 + (j-\alpha-1)s^{2(\alpha+1)}]^2} \\ &\le C\|f_j\|_{0,\beta,\nu-2} \left(\frac{r}{\bar r} - 1 \right) \bar r^\nu \left[\bar r \int_r^{2\bar r} \frac{ds}{(s-\bar r)^2} + 1 \right] \le C\|f_j\|_{0,\beta,\nu-2} \bar r^\nu. \end{aligned}$$

Since $|\psi_j(s,\rho)| \le C\|f_j\|_{0,\beta,\nu-2} \bar r^2 s^{\nu-j-1}$ for $s \le \bar r$, then $|\psi_j(\bar r,\rho)| \le C\|f_j\|_{0,\beta,\nu-2} \bar r^{\nu-j+1}$ and in turn for $s \le \frac{\bar r}{2}$

$$\left| \frac{\psi_j(s,\rho) - \psi_j(\bar r,\rho)}{(s-\bar r)^2} \right| \le C\|f_j\|_{0,\beta,\nu-2} \left(s^{\nu-j-1} + \bar r^{\nu-j-1} \right). \tag{20}$$

For $s \in [\frac{\bar r}{2}, 2\bar r] \setminus \{\bar r\}$, we decompose

$$\frac{\psi_j(s,\rho)-\psi_j(\bar r,\rho)}{(s-\bar r)^2} = \left[\frac{1}{sa_j^2(s)} - \frac{\bar r^{-(2j-1)}(s-\bar r)^{-2}}{[(\alpha+1)^2-j^2]^2}\right]\int_0^{\bar r} ta_j(t)f_j(t)dt + \frac{1}{sa_j^2(s)}\int_{\bar r}^s ta_j(t)f_j(t)dt$$

and hence, using a homogeneity argument, we get

$$\begin{aligned}&\left|\frac{\psi_j(s,\rho)-\psi_j(\bar r,\rho)}{(s-\bar r)^2}\right| \\ \le\ & C\|f_j\|_{0,\beta,\nu-2}\bar r^{\nu-j-1}\left\{\frac{[1+z^{2(\alpha+1)}]^2}{(1-z^{2(\alpha+1)})^2 z^{2j+1}}\left|\int_1^z \frac{1-t^{2(\alpha+1)}}{1+t^{2(\alpha+1)}}t^{\nu+j-1}dt\right|\right. \\ &\left.+\left|\frac{[(\alpha+1-j)+(\alpha+1+j)z^{2(\alpha+1)}]^2}{(1-z^{2(\alpha+1)})^2 z^{2j+1}}-\frac{1}{(z-1)^2}\right|\right\}\Bigg|_{z=\frac{s}{\bar r}\in[\frac12,2]\setminus\{1\}}.\end{aligned}$$

Consequently, for $\frac{\bar r}{2} \le s \le 2\bar r$ we obtain

$$\left|\frac{\psi_j(s,\rho)-\psi_j(\bar r,\rho)}{(s-\bar r)^2}\right| \le C\|f_j\|_{0,\beta,\nu-2}s^{\nu-j-1}. \tag{21}$$

Finally, it is easy to see that for $r \le 2\bar r$ and for w_j^2 there holds

$$\sup_{r\in(0,2\bar r)} r^{-\nu}|w_j^2(r)| \le C\|f_j\|_{0,\beta,\nu-2}\sup_{0\le t\le 2} t^{-\nu+j}\frac{|1-t^{2(\alpha+1)}|}{|1-t|} \le C\|f_j\|_{0,\beta,\nu-2}$$

in view of the estimate available for $\psi_j(\bar r,\rho)$ and $\nu < j$. Since $|w_j^1(2\bar r)| \le |w_j(2\bar r)| + |w_j^2(2\bar r)| \le C\|f_j\|_{0,\beta,\nu-2}\bar r^{\nu}$, then $|\int_1^{2\bar r}\frac{\psi_j(s,\rho)-\psi_j(\bar r,\rho)}{(s-\bar r)^2}ds| \le C\|f_j\|_{0,\beta,\nu-2}\bar r^{\nu-j}$. So we derive, by splitting the integral in (19) as $\int_1^r = \int_1^{2\bar r} + \int_{2\bar r}^r$ and using (20) and (21), the estimate $\sup_{r\in(0,\bar r)} r^{-\nu}|w_j^1(r)| \le C\|f_j\|_{0,\beta,\nu-2}$ for $\nu < j$. Finally, the estimate $\sup_{r\in(0,1)} r^{-\nu}|w_j(r)| \le C\|f_j\|_{0,\beta,\nu-2}$ does hold and, using classical rescaled Schauder estimates, we get the existence of some constant $C > 0$ such that $\|w_j\|_{2,\beta,\nu} \le C\|f\|_{0,\beta,\nu-2}$ for $0 < j \le m_\alpha$ and $-j < \nu < j$.

Step 3. Analogously, for $j = 0$ and $\nu > 0$, it is possible to consider, for $0 < r < \bar r := (\tau\rho)^{\frac{1}{\alpha+1}}$,

$$\tilde w_0(r) = \left(\int_0^r \frac{ds}{sa_0^2(s)}\int_0^s ta_0(t)f_0(t)dt\right)a_0(r).$$

By defining $\psi_0(s,\rho) = (s-\bar r)^2\frac{1}{sa_0^2(s)}\int_0^s ta_0(t)f_0(t)dt$, we can extend $\tilde w_0$ for $r \ge \bar r$ by considering

$$\hat w_0(r) = a_0(r)\left[\int_0^r \frac{\psi_0(s,\rho)-\psi_0(\bar r,\rho)}{(s-\bar r)^2}ds + \frac{r}{\bar r(\bar r-r)}\psi_0(\bar r,\rho)\right]$$

which defines a solution for $(P0)$, with $\hat w_0(1) \ne 0$ in general. We have the estimate $\sup_{r\in(0,\bar r)} r^{-\nu}|\hat w_0(r)| \le C\|f_0\|_{0,\beta,\nu-2}$. In fact, for $r \le \frac{\bar r}{2}$ we see that $|\hat w_0(r)| = |\tilde w_0(r)| \le C\|f_0\|_{0,\beta,\nu-2}r^\nu$ since $\frac{1}{a_0^2(s)} = O(1)$ for $s \le r$. While for $\frac{\bar r}{2} \le r \le \bar r$ there holds

$$|\hat w_0(r)| = |\tilde w_0(r)| \le C\|f_0\|_{0,\beta,\nu-2}\left(1-\frac{r}{\bar r}\right)\bar r^{\nu+1}\int_0^r \frac{ds}{(s-\bar r)^2} \le C\|f_0\|_{0,\beta,\nu-2}\bar r^\nu.$$

Furthermore, since $|\psi_0(s,\rho)| \le C\|f_0\|_{0,\beta,\nu-2}\bar r^2 s^{\nu-1}$ for $s \le \bar r$, as above for $s \le \frac{\bar r}{2}$ we obtain

$$\left|\frac{\psi_0(s,\rho)-\psi_0(\bar r,\rho)}{(s-\bar r)^2}\right| \le C\|f_0\|_{0,\beta,\nu-2}\left(s^{\nu-1}+\bar r^{\nu-1}\right). \tag{22}$$

On the other hand, for $s \ge \frac{\bar r}{2}$ we have

$$\left|\frac{\psi_0(s,\rho)-\psi_0(\bar r,\rho)}{(s-\bar r)^2}\right| \le C\|f_0\|_{0,\beta,\nu-2}s^{\nu-1}. \tag{23}$$

In fact, (23) follows as in (21) when $s \in [\frac{\bar r}{2}, 2\bar r] \setminus \{\bar r\}$. While for $s \ge 2\bar r$, we have

$$\begin{aligned}\left|\frac{\psi_0(s,\rho)-\psi_0(\bar r,\rho)}{(s-\bar r)^2}\right| &= \left[\frac{1}{sa_0^2(s)}-\frac{\bar r}{(\alpha+1)^4(s-\bar r)^2}\right]\int_0^{\bar r} ta_0(t)f_0(t)dt\\ &\quad+\frac{1}{sa_0^2(s)}\int_{\bar r}^{s} ta_0(t)f_0(t)dt\\ &\le C\|f_0\|_{0,\beta,\nu-2}\left(\frac{\bar r^\nu}{s}+\frac{s^{\nu-1}}{(\frac{s}{\bar r})^{\nu-1}(\frac{s}{\bar r}-1)^2}+s^{\nu-1}\right).\end{aligned}$$

Finally, since $\nu > 0$ it is easy to see that $\sup_{r\in(\bar r,1)} r^{-\nu}|a_0(r)\psi_0(\bar r,\rho)\frac{r}{\bar r(r-\bar r)}| \le C\|f_0\|_{0,\beta,\nu-2}$. While by (22) and (23) for $r \ge \bar r$ we get $|a_0(r)\int_0^r \frac{\psi_0(s,\rho)-\psi_0(\bar r,\rho)}{(s-\bar r)^2}ds| \le C\|f_0\|_{0,\beta,\nu-2}r^\nu$. Hence $\sup_{r\in(0,1)} r^{-\nu}|\hat w_0(r)| \le C\|f_0\|_{0,\beta,\nu-2}$. Consequently, by classical rescaled Schauder estimates, we find a suitable constant $C > 0$ such that $\|\hat w_0\|_{2,\beta,\nu} \le C\|f\|_{0,\beta,\nu-2}$. We set now

$$w_0(r) = \hat w_0(r) + H^0_{\rho,\tau}(f)\frac{\tau^2\rho^2 - r^{2(\alpha+1)}}{\tau^2\rho^2 + r^{2(\alpha+1)}}, \tag{24}$$

where $H^0_{\rho,\tau}(f) \in \mathbb{R}$ is such that $w_0(1) = 0$. Hence $w_0(r)$ is a solution for $(P0)$ and for $\nu > 0$,

$$|H^0_{\rho,\tau}(f)| \le C|\hat w_0(1)| \le C\|f\|_{0,\beta,\nu-2}.$$

Notice that for $\alpha > 0$, $\alpha \notin \mathbb{N}$, Steps 1–3 lead to the proof of Proposition 4.1 by choosing $\nu \in (0,1)$ and $G_{\rho,\tau}(f) = \hat w_0(r) + 2\sum_{j=1}^{+\infty} w_j(r)\cdot e^{-ij\theta}$.

Step 4. To obtain Proposition 4.2, it remains for problem $(P1)$ to be considered with $\alpha = 0$ while for the validity of Steps 1–3 we must specify $\nu \in (0,2)$. To account also for $(P1)$ we further specify $1 < \nu < 2$. Then it is possible to define

$$\begin{aligned}\hat w_1(r) &= \left(\int_0^r \frac{ds}{sa_1^2(s)}\int_0^s ta_1(t)f_1(t)dt\right)a_1(r)\\ &= \frac{r}{\tau^2\rho^2+r^2}\int_0^r \frac{(\tau^2\rho^2+s^2)^2}{s^3}ds\int_0^s \frac{t^2}{\tau^2\rho^2+t^2}f_1(t)dt.\end{aligned}$$

To estimate $\|\hat w_1(r)\|_{2,\beta,\nu}$, introduce $z = \frac{r}{\tau\rho}$ and observe that

$$\begin{aligned}\sup_{r\in(0,1)} r^{-\nu}|\hat w_1(r)| &\le \|f_1\|_{0,\beta,\nu-2}\sup_{z\in(0,(\tau\rho)^{-1})}\frac{z^{1-\nu}}{1+z^2}\int_0^z\frac{(1+s^2)^2}{s^3}ds\int_0^s\frac{t^\nu}{1+t^2}dt\\ &\le C\|f\|_{0,\beta,\nu-2}.\end{aligned}$$

Set

$$w_1(r) = \hat{w}_1(r) + \overline{H^1_{\rho,\tau}(f)} \frac{r}{\tau^2\rho^2 + r^2}, \tag{25}$$

where $H^1_{\rho,\tau}(f) \in \mathbb{C}$ is such that $w_1(1) = 0$. Hence $w_1(r)$ is a solution for $(P1)$ and

$$|H^1_{\rho,\tau}(f)| \le C|\hat{w}_1(1)| \le C\|f\|_{0,\beta,\nu-2}.$$

Therefore, for $\alpha = 0$ Proposition 4.2 also follows with

$$G_{\rho,\tau}(f) = \hat{w}_0(r) + 2\hat{w}_1(r) \cdot e^{-i\theta} + 2\sum_{j=2}^{+\infty} w_j(r) \cdot e^{-ij\theta}$$

whenever $\nu \in (1,2)$. Finally, using Lemma 4.3 we can deduce the uniqueness of $w_j(r)$. The uniqueness of the decomposition (24) follows by evaluating $w_0(r)$ at $r = 0$. Similarly, if $\alpha = 0$ the uniqueness of the decomposition (25) follows by evaluating $\frac{w_1(r)}{r}$ at $r = 0$. Hence Propositions 4.1 and 4.2 are completely established.

Remark 4.4. The function $G_{\rho,\tau}(f)$ is the unique solution in $C^{2,\beta}_{\nu}(B(0,1))$ for $L_{\rho,\tau}w = f$ in $B(0,1)$ such that

$$G_{\rho,\tau}(f)|_{\partial B(0,1)} = \begin{cases} \hat{w}_0(1) & \text{if } \alpha \notin \mathbb{N}, \\ \hat{w}_0(1) + 2\hat{w}_1(1) \cdot e^{-i\theta} & \text{if } \alpha = 0. \end{cases}$$

4.2. Some local operator. The nonradial case. In case $\alpha = 0$, we discuss now the invertibility of the operator

$$L_{\rho,\tau,\gamma}w = \Delta w + \rho^2 e^{v_{\rho,\tau,\gamma}} w$$

under Dirichlet boundary condition. The following result holds.

PROPOSITION 4.5. *Let $\alpha = 0$. For all $\nu \in (1,2)$ and $\gamma \in \mathbb{C}$, $|\gamma| < \frac{1}{3}$, $\tau > 0$, there exist $\rho_0 > 0$, two continuous linear forms $H^0_{\rho,\tau,\gamma} : C^{0,\beta}_{\nu-2}(B(0,1)) \to \mathbb{R}$, $H^1_{\rho,\tau,\gamma} : C^{0,\beta}_{\nu-2}(B(0,1)) \to \mathbb{C}$, and a linear operator $G_{\rho,\tau,\gamma} : C^{0,\beta}_{\nu-2}(B(0,1)) \to C^{2,\beta}_{\nu}(B(0,1))$, uniformly bounded for $0 < \rho < \rho_0$, such that for all $\rho \in (0,\rho_0)$ and for all $f \in C^{0,\beta}_{\nu-2}(B(0,1))$ there exists a unique bounded solution w of*

$$\begin{cases} L_{\rho,\tau,\gamma}w = f & \text{in } B(0,1), \\ w = 0 & \text{on } \partial B(0,1) \end{cases}$$

which can be uniquely decomposed as

$$w(z) = G_{\rho,\tau,\gamma}(f)(z) + H^0_{\rho,\tau,\gamma}(f)\partial_\tau v_{\rho,\tau,\gamma} + 2H^1_{\rho,\tau,\gamma}(f) \cdot \partial_{\bar{z}} v_{\rho,\tau,\gamma}.$$

Moreover, the following estimates hold:

$$\|G_{\rho,\tau,\gamma}(f)\|_{2,\beta,\nu} \le C\left(\|G_{\rho,\tau}(f)\|_{2,\beta,\nu} + \rho^2|H^0_{\rho,\tau}(f)| + |H^1_{\rho,\tau}(f)|\right), \tag{26}$$

$$|H^0_{\rho,\tau,\gamma}(f)| \le C\left(\rho^2\|G_{\rho,\tau}(f)\|_{2,\beta,\nu} + |H^0_{\rho,\tau}(f)| + \rho^2|H^1_{\rho,\tau}(f)|\right), \tag{27}$$

$$|H^1_{\rho,\tau,\gamma}(f)| \le C\left(\rho^2\|G_{\rho,\tau}(f)\|_{2,\beta,\nu} + \rho^2|H^0_{\rho,\tau}(f)| + |H^1_{\rho,\tau}(f)|\right), \tag{28}$$

$$\begin{aligned}\|\partial_r G_{\rho,\tau,\gamma}(f)|_{\partial B(0,1)}\|_{1,\beta} \le C\big(&\|\partial_r G_{\rho,\tau}(f)|_{\partial B(0,1)}\|_{1,\beta} + \rho^2\|G_{\rho,\tau}(f)\|_{2,\beta,\nu}\\ &+ \rho^2|H^0_{\rho,\tau}(f)| + |H^1_{\rho,\tau}(f)|\big)\end{aligned} \tag{29}$$

for some constant $C > 0$.

Proof. In case $\alpha = 0$, we compute

$$\begin{aligned}\lim_{\rho\to 0}\partial_\tau v_{\rho,\tau,\gamma}(z) &= \lim_{\rho\to 0}\left(\frac{2}{\tau}\frac{|z|^2|1+\gamma z^2|^2-\tau^2\rho^2}{|z|^2|1+\gamma z^2|^2+\tau^2\rho^2}\right) = \frac{2}{\tau},\\ \lim_{\rho\to 0}\partial_{\bar z} v_{\rho,\tau,\gamma}(z) &= \lim_{\rho\to 0}\left(\frac{6\bar\gamma\bar z}{1+3\bar\gamma\bar z^2} - 2\frac{z(1+\gamma z^2)(1+3\bar\gamma\bar z^2)}{|z|^2|1+\gamma z^2|^2+\tau^2\rho^2}\right)\\ &= 6\bar\gamma\bar z\left(\sum_{k=0}^{+\infty}(-1)^k 3^k\bar\gamma^k\bar z^{2k}\right) - 2(1+3\bar\gamma\bar z^2)\left(\sum_{k=0}^{+\infty}(-1)^k\bar\gamma^k\bar z^{2k-1}\right)\\ &= -\frac{2}{\bar z} + 2\bar\gamma\bar z + \eta^\perp(z)\end{aligned}$$

uniformly on compact sets in $B(0,1)\setminus\{0\}$, where $\eta^\perp(z) = 2\sum_{k=0}^{+\infty}(-1)^{k+1}(3^{k+2}-2)\bar\gamma^{k+2}\bar z^{2k+3}$ is orthogonal to $\{1, e^{\pm i\theta}\}$ (in the sense $\eta^\perp(re^{i\theta})$ is orthogonal to 1 and $e^{\pm i\theta}$ in $L^2([0,2\pi])$ for any $r\in(0,1]$) and it is a harmonic function. Set $\mathrm{Span}\{1, e^{-i\theta}\} = \{a_0 + 2a_1\cdot e^{-i\theta} :\ a_0\in\mathbb{R},\ a_1\in\mathbb{C}\}$ and define π as the orthogonal projection over $\mathrm{Span}\{1, e^{-i\theta}\}$. Define the mapping $\psi_\rho : (h_0,h_1)\in\mathbb{R}\times\mathbb{C}\to(\psi^1_\rho(h_0,h_1),\psi^2_\rho(h_0,h_1))\in\mathbb{R}\times\mathbb{C}$ by setting

$$\psi^1_\rho(h_0,h_1) + 2\psi^2_\rho(h_0,h_1)\cdot e^{-i\theta} = \pi\left(h_0\partial_\tau v_{\rho,\tau,\gamma}(e^{i\theta}) + 2h_1\cdot\partial_{\bar z}v_{\rho,\tau,\gamma}(e^{i\theta})\right).$$

Note that $\psi_\rho\to\psi_0$ as $\rho\to 0$ in the operatorial norm, where $\psi_0(h_0,h_1) = (\frac{2}{\tau}h_0, -2\bar h_1 + 2\gamma h_1)$ is an invertible operator for $|\gamma|<1$ with inverse $\psi_0^{-1}(\tilde h_0,\tilde h_1) = (\frac{\tau}{2}\tilde h_0, -\frac{1}{2(1-|\gamma|^2)}\cdot(\bar\gamma\tilde h_1 + \bar{\tilde h}_1))$. Hence, for ρ small ψ_ρ is invertible with uniformly bounded inverse. Let $f\in C^{0,\beta}_{\nu-2}(B(0,1))$ be a given function and $\nu\in(1,2)$. By Remark 4.4, there exists a unique $w_0\in C^{2,\beta}_\nu(B(0,1))$ such that

$$\begin{cases} L_{\rho,\tau}w_0 = f \quad \text{in } B(0,1),\\ w_0\,|_{\partial B(0,1)} = \tilde h_0 + 2\tilde h_1\cdot e^{-i\theta}\in\mathrm{Span}\{1, e^{-i\theta}\}.\end{cases}$$

Let $(h_0,h_1) = \psi_\rho^{-1}(\tilde h_0,\tilde h_1)$. We define on $\partial B(0,1)$

$$\phi(\theta) := h_0\partial_\tau v_{\rho,\tau,\gamma}(e^{i\theta}) + 2h_1\cdot\partial_{\bar z}v_{\rho,\tau,\gamma}(e^{i\theta}) - \tilde h_0 - 2\tilde h_1\cdot e^{-i\theta}$$

in such a way that $\pi\phi = 0$. We extend ϕ in $B(0,1)$ as $\tilde\phi(z) = \sigma(r)\phi(\theta)$, where $0\le\sigma\le 1$ is a smooth function with $\sigma\equiv 1$ in $[\frac12,1]$ and $\sigma\equiv 0$ in $[0,\frac14]$. Since $L_{\rho,\tau}\tilde\phi\in\mathrm{Span}\{1, e^{-i\theta}\}^\perp$, by Proposition 4.2 we get $H^0_{\rho,\tau}(-L_{\rho,\tau}\tilde\phi) = 0$, $H^1_{\rho,\tau}(-L_{\rho,\tau}\tilde\phi) = 0$ and hence $w_1 := G_{\rho,\tau}(-L_{\rho,\tau}\tilde\phi)$ vanishes on $\partial B(0,1)$. The function $w_2 := w_0 + \tilde\phi + w_1\in C^{2,\beta}_\nu(B(0,1))$ solves

$$\begin{cases} L_{\rho,\tau}w_2 = f \quad \text{in } B(0,1),\\ w_2\,|_{\partial B(0,1)} = h_0\partial_\tau v_{\rho,\tau,\gamma}(e^{i\theta}) + 2h_1\cdot\partial_{\bar z}v_{\rho,\tau,\gamma}(e^{i\theta})\end{cases}$$

with $\|w_2\|_{2,\beta,\nu}\le C\|f\|_{0,\beta,\nu-2}$. Moreover, w_2 is the unique solution in $C^{2,\beta}_\nu(B(0,1))$ for the problem. If w_2' is a solution in $C^{2,\beta}_\nu(B(0,1))$ with

$$w_2'\,|_{\partial B(0,1)} = h_0'\partial_\tau v_{\rho,\tau,\gamma}(e^{i\theta}) + 2h_1'\cdot\partial_{\bar z}v_{\rho,\tau,\gamma}(e^{i\theta}) = \tilde h_0' + 2\tilde h_1'\cdot e^{-i\theta} + \phi',$$

then by the uniqueness part in Proposition 4.2 we derive $w_0 = w_2' - \tilde\phi' - w_1'$, $\tilde h_0' = \tilde h_0$, $\tilde h_1' = \tilde h_1$, where $w_1' = G_{\rho,\tau}(-L_{\rho,\tau}\tilde\phi')$ and $\tilde\phi'$ extends ϕ' as before. Since ψ_ρ is injective, then $h_0' = h_0$, $h_1' = h_1$, $\phi' = \phi$ and hence $\tilde\phi' + w_1' = \tilde\phi + w_1$ and $w_2' = w_2$.

Then $L_{\rho,\tau}$, as an operator between

$$\{w \in C^{2,\beta}_\nu(B(0,1)) : w\,|_{\partial B(0,1)} = h_0\partial_\tau v_{\rho,\tau,\gamma}(e^{i\theta}) + 2h_1\cdot\partial_{\bar z} v_{\rho,\tau,\gamma}(e^{i\theta}),\ (h_0,h_1)\in\mathbb{R}\times\mathbb{C}\}$$

and $C^{0,\beta}_{\nu-2}(B(0,1))$, is an isomorphism with inverse uniformly bounded with respect to $\|\cdot\|_{0,\beta,\nu-2}$ and $\|\cdot\|_{2,\beta,\nu}$. We will denote this inverse operator as $L^{-1}_{\rho,\tau}$. Moreover, we have the estimate $|h_0(f)| + |h_1(f)| \le C\|f\|_{0,\beta,\nu-2}$. We use now a perturbation argument to prove Proposition 4.5. Since for $z, x, y \in B(0,1)$ we have

$$|v_{\rho,\tau,\gamma} - v_{\rho,\tau}|(z) = \left|\ln\frac{(\tau^2\rho^2+|z|^2)^2|1+3\gamma z^2|^2}{(\tau^2\rho^2+|z|^2|1+\gamma z^2|^2)^2}\right| \le C|z|^2,$$

$$\begin{aligned}\left|\frac{(v_{\rho,\tau,\gamma}-v_{\rho,\tau})(x) - (v_{\rho,\tau,\gamma}-v_{\rho,\tau})(y)}{|x-y|^\beta}\right| &\le 2|\partial_z(v_{\rho,\tau,\gamma}-v_{\rho,\tau})(\xi)||x-y|^{1-\beta}\\ &\le C\left(\max\{|x|,|y|\}\right)^{2-\beta}\end{aligned}$$

for some point ξ on the segment joining x and y, we get that $\|v_{\rho,\tau,\gamma} - v_{\rho,\tau}\|_{0,\beta,2} \le C$. Hence, for $w \in C^{2,\beta}_\nu(B(0,1))$ we have the estimate

$$\|(L_{\rho,\tau,\gamma} - L_{\rho,\tau})w\|_{0,\beta,\nu-2} = \rho^2\|(e^{v_{\rho,\tau,\gamma}} - e^{v_{\rho,\tau}})w\|_{0,\beta,\nu-2} \le C\rho^2\|w\|_{2,\beta,\nu}. \tag{30}$$

A solution for the problem

$$\begin{cases} L_{\rho,\tau,\gamma}w = f \quad \text{in } B(0,1),\\ w\,|_{\partial B(0,1)} = -H^0_{\rho,\tau,\gamma}(f)\partial_\tau v_{\rho,\tau,\gamma}(e^{i\theta}) - 2H^1_{\rho,\tau,\gamma}(f)\cdot\partial_{\bar z}v_{\rho,\tau,\gamma}(e^{i\theta})\end{cases}$$

corresponds to a fixed point for the map $w \to L^{-1}_{\rho,\tau}f + L^{-1}_{\rho,\tau}(L_{\rho,\tau} - L_{\rho,\tau,\gamma})w$. By (30) we deduce that this map is a contraction. So it has a unique fixed point $w = G_{\rho,\tau,\gamma}(f)$ which satisfies $\|G_{\rho,\tau,\gamma}(f)\|_{2,\beta,\nu} \le C\|L^{-1}_{\rho,\tau}(f)\|_{2,\beta,\nu}$. At this point, we deduce (26)–(29): since

$$|\partial_\tau v_{\rho,\tau,\gamma}(e^{i\theta}) - \frac{2}{\tau}| + |\partial_{\bar z}v_{\rho,\tau,\gamma}(e^{i\theta}) - \left(-2e^{i\theta} + 2\bar\gamma e^{-i\theta} + \eta^\perp(e^{i\theta})\right)| \le C\rho^2,$$

there holds the estimate $\|\psi_\rho - \psi_0\| + \|\psi_\rho^{-1} - \psi_0^{-1}\| \le C\rho^2$. Therefore

$$\begin{aligned}(h_0,h_1) &= \psi_0^{-1}(\tilde h_0,\tilde h_1) + O\left(\rho^2|H^0_{\rho,\tau}(f)| + \rho^2|H^1_{\rho,\tau}(f)|\right)\\ &= \left(O(|H^0_{\rho,\tau}(f)| + \rho^2|H^1_{\rho,\tau}(f)|), O(\rho^2|H^0_{\rho,\tau}(f)| + |H^1_{\rho,\tau}(f)|)\right)\end{aligned} \tag{31}$$

as $\tilde h_0 = -H^0_{\rho,\tau}(f)\frac{\tau^2\rho^2-1}{\tau^2\rho^2+1}$ and $\tilde h_1 = -\overline{H^1_{\rho,\tau}(f)}\frac{1}{\tau^2\rho^2+1}$. On $\partial B(0,1)$ there holds

$$\begin{aligned}\phi(\theta) &= \frac{2}{\tau}h_0 + 2(-2\bar h_1 + 2\gamma h_1)\cdot e^{-i\theta} - \tilde h_0 - 2\tilde h_1\cdot e^{-i\theta} + 2h_1\cdot\eta^\perp(e^{i\theta})\\ &\quad + O\left(\rho^2|h_0| + \rho^2|h_1|\right) = 2h_1\cdot\eta^\perp(e^{i\theta}) + O\left(\rho^2|H^0_{\rho,\tau}(f)| + \rho^2|H^1_{\rho,\tau}(f)|\right)\\ &= O\left(\rho^2|H^0_{\rho,\tau}(f)| + |H^1_{\rho,\tau}(f)|\right)\end{aligned}$$

and we deduce

$$\|\tilde\phi\|_{2,\beta,\nu} + \|w_1\|_{2,\beta,\nu} = O\left(\rho^2|H^0_{\rho,\tau}(f)| + |H^1_{\rho,\tau}(f)|\right). \tag{32}$$

Since $G_{\rho,\tau,\gamma}(f) = L^{-1}_{\rho,\tau} f + L^{-1}_{\rho,\tau}(L_{\rho,\tau} - L_{\rho,\tau,\gamma})G_{\rho,\tau,\gamma}(f)$ with $L^{-1}_{\rho,\tau} f = G_{\rho,\tau}(f) + \tilde{\phi} + w_1$, we get (26) as follows:

$$\begin{aligned}\|G_{\rho,\tau,\gamma}(f)\|_{2,\beta,\nu} &\le C\left(\|G_{\rho,\tau}(f)\|_{2,\beta,\nu} + \|\tilde{\phi}\|_{2,\beta,\nu} + \|w_1\|_{2,\beta,\nu}\right)\\ &\le C\left(\|G_{\rho,\tau}(f)\|_{2,\beta,\nu} + \rho^2|H^0_{\rho,\tau}(f)| + |H^1_{\rho,\tau}(f)|\right)\end{aligned}$$

and in turn by (32) and (26) we obtain (29).

Letting $S = f + (L_{\rho,\tau} - L_{\rho,\tau,\gamma})G_{\rho,\tau,\gamma}(f)$, by (31) and (26) we find

$$\begin{aligned}|H^0_{\rho,\tau,\gamma}(f)| &= |h_0(S)| = O\left(\rho^2\|G_{\rho,\tau}(f)\|_{2,\beta,\nu} + |H^0_{\rho,\tau}(f)| + \rho^2|H^1_{\rho,\tau}(f)|\right),\\ |H^1_{\rho,\tau,\gamma}(f)| &= |h_1(S)| = O\left(\rho^2\|G_{\rho,\tau}(f)\|_{2,\beta,\nu} + \rho^2|H^0_{\rho,\tau}(f)| + |H^1_{\rho,\tau}(f)|\right),\end{aligned}$$

and the proof of Proposition 4.5 is completed. □

4.3. Some global operator. Let $\alpha \in (0,+\infty)\setminus\mathbb{N}$ be a fixed number. Let χ be a radial smooth function such that $0 \le \chi \le 1$, $\chi = 1$ in $B(0,1)$, $\chi = 0$ in $\mathbb{R}^2 \setminus B(0,2)$. In Theorem 1.4 we are interested in dealing with three possible cases:

(a) the concentration set S is a single point which is a singular source, that is, $S = \{p\}$, and in this case we consider the associated potential as given by $V_\rho(z) = \rho^2\chi(z-p)|z-p|^{2\alpha}e^{v_{\rho,\tau_1}(z-p)}$ where $\tau_1 > 0$ is defined in section 2;

(b) the concentration set S is a single point which is not a singular source, that is, $S = \{q\}$ with $q \ne p$, and the associated potential considered in this case is $V_\rho(z) = \rho^2\chi(z-q)e^{v_{\rho,\tau_2,\gamma}(z-q)}$ where $\tau_2 > 0$ and γ are defined in section 2;

(c) the concentration set $S = \{p,q\}$ and the associated potential is $V_\rho(z) = \rho^2\chi(z-p)|z-p|^{2\alpha}e^{v_{\rho,\tau_1}(z-p)} + \rho^2\chi(z-q)e^{v_{\rho,\tau_2,\gamma}(z-q)}$ where $\tau_1,\tau_2 > 0$ and γ are defined in section 2.

We are assuming that $\overline{B(p,2)} \cap \overline{B(q,2)} = \emptyset$, $\overline{B(p,2)} \subset \Omega$, and $\overline{B(q,2)} \subset \Omega$. Set $B = B(p,1) \cup B(q,1)$ and $\tilde{\Omega} = \Omega \setminus B$.

We introduce the operator $\mathcal{L}_\rho = \Delta + V_\rho$ where the potential V_ρ is defined above according to the cases (a), (b), and (c) we wish to deal with. We investigate the invertibility of $\mathcal{L}_\rho$ between X and Y (see section 3 for the definition of X and Y). We will prove the following result.

THEOREM 4.6. *There exist $\rho_0 > 0$ small, continuous linear forms $\mathcal{H}^0_{\rho,1}, \mathcal{H}^0_{\rho,2} : Y \to \mathbb{R}$ and $\mathcal{H}^1_{\rho,2} : Y \to \mathbb{C}$, a linear operator $\mathcal{G}_\rho : Y \to, X$, uniformly bounded for $\rho \in (0,\rho_0)$, such that for all $f \in Y$ and $\rho \in (0,\rho_0)$ there exists a unique solution $w(z)$ of*

$$\begin{cases}\mathcal{L}_\rho w = f & \text{in } \Omega,\\ w = 0 & \text{on } \partial\Omega\end{cases}$$

which can be decomposed in a unique way in the form

$$w(z) = \mathcal{G}_\rho(f)(z) + \chi(z-p)\mathcal{H}^0_{\rho,1}\partial_\tau v_{\rho,\tau_1}(z-p)$$

in case (a), *in the form*

$$w(z) = \mathcal{G}_\rho(f)(z) + \chi(z-q)\mathcal{H}^0_{\rho,2}\partial_\tau v_{\rho,\tau_2,\gamma}(z-q) + 2\chi(z-q)\mathcal{H}^1_{\rho,2}\cdot\partial_{\bar{z}} v_{\rho,\tau_2,\gamma}(z-q)$$

in case (b), *and in the form*

$$\begin{aligned}w(z) = \mathcal{G}_\rho(f)(z) &+ \chi(z-p)\mathcal{H}^0_{\rho,1}\partial_\tau v_{\rho,\tau_1}(z-p)\\ &+ \chi(z-q)\mathcal{H}^0_{\rho,2}\partial_\tau v_{\rho,\tau_2,\gamma}(z-q) + 2\chi(z-q)\mathcal{H}^1_{\rho,2}\cdot\partial_{\bar{z}} v_{\rho,\tau_2,\gamma}(z-q)\end{aligned}$$

in case (c).

We collect some preliminary results which will be crucial to the proof of Theorem 4.6. Since $|V_\rho| \le C\rho^2$ in $\tilde\Omega$, from classical elliptic theory we have the following lemma.

LEMMA 4.7. *There exists $\rho_0 > 0$ small such that for all $f \in C^{0,\beta}(\tilde\Omega)$ there exists a unique solution $w \in C^{2,\beta}(\tilde\Omega)$ for the problem*

$$\begin{cases} \mathcal{L}_\rho w = f & \text{in } \tilde\Omega, \\ w = 0 & \text{on } \partial\tilde\Omega. \end{cases}$$

Moreover, $\|w\|_{2,\beta,\tilde\Omega} \le C\|f\|_{0,\beta,\tilde\Omega}$.

We introduce now the exterior Dirichlet to Neumann map. Let $\Phi \in C^{2,\beta}(\partial B)$; we can extend Φ inside $\tilde\Omega$ in such a way that $\tilde\Phi \in C^{2,\beta}(\tilde\Omega)$, $\tilde\Phi = 0$ on $\partial\Omega$, and $\|\tilde\Phi\|_{2,\beta,\tilde\Omega} \le C\|\Phi\|_{2,\beta,\partial B}$.

By Lemma 4.7 we can find a solution $\bar w$ for

$$\begin{cases} \mathcal{L}_\rho \bar w = -\mathcal{L}_\rho \tilde\Phi & \text{in } \tilde\Omega, \\ \bar w = 0 & \text{on } \partial\tilde\Omega \end{cases}$$

and hence $w_\Phi = \bar w + \tilde\Phi$ solves

$$\begin{cases} \mathcal{L}_\rho w_\Phi = 0 & \text{in } \tilde\Omega, \\ w_\Phi = 0 & \text{on } \partial\Omega, \\ w_\Phi = \Phi & \text{on } \partial B \end{cases}$$

with $\|w_\Phi\|_{2,\beta,\tilde\Omega} \le C\|\Phi\|_{C^{2,\beta}(\partial B)}$.

Define

$$\begin{aligned} S_\rho : C^{2,\beta}(\partial B) &\to C^{1,\beta}(\partial B) \\ \Phi &\to S_\rho(\Phi) = \frac{\partial w_\Phi}{\partial n} \mid_{\partial B}, \end{aligned}$$

where n is the unit inward normal on ∂B to $\tilde\Omega$. If $\tilde w$ denotes the solution of

$$\begin{cases} \Delta \tilde w = -\Delta \tilde\Phi & \text{in } \tilde\Omega, \\ \tilde w = 0 & \text{on } \partial\tilde\Omega, \end{cases}$$

then

$$\begin{cases} \Delta(\bar w - \tilde w) = -V_\rho w_\Phi & \text{in } \tilde\Omega, \\ \bar w - \tilde w = 0 & \text{on } \partial\tilde\Omega \end{cases}$$

and so, by classical Schauder estimates, $\|\bar w - \tilde w\|_{2,\beta,\tilde\Omega} \le C\rho^2\|\Phi\|_{C^{2,\beta}(\partial B)}$. Hence, if S_0 denotes the Dirichlet to Neumann map corresponding to Δ on $\tilde\Omega$, we have that $S_\rho = S_0 + O(\rho^2)$. Summarizing, we have the following lemma.

LEMMA 4.8. *There exists $\rho_0 > 0$ small such that for $\rho \in (0, \rho_0)$ the map S_ρ is well defined and $S_\rho \to S_0$ as $\rho \to 0$ in the operatorial norm.*

We introduce now the interior Dirichlet to Neumann map. Let $\Phi \in C^{2,\beta}(\partial B)$, which we extend as $\tilde\Phi$ in B in such a way that $\|\tilde\Phi\|_{2,\beta,\nu_1,B(p,1)} + \|\tilde\Phi\|_{2,\beta,\nu_2,B(q,1)} \le$

$C\|\Phi\|_{2,\beta,\partial B}$. By Propositions 4.1 and 4.5, we see that there exists a unique solution $\bar{v}$ of

$$\begin{cases} \mathcal{L}_\rho \bar{v} = -\mathcal{L}_\rho \tilde{\Phi} & \text{in } B, \\ \bar{v} = 0 & \text{on } \partial B \end{cases}$$

and hence $v_\Phi = \bar{v} + \tilde{\Phi}$ uniquely solves

$$\begin{cases} \mathcal{L}_\rho v_\Phi = 0 & \text{in } B, \\ v_\Phi = \Phi & \text{on } \partial B \end{cases}$$

with $\|v_\Phi \mid_{B(p,1)}\|_{\mathcal{E}_1} + \|v_\Phi \mid_{B(q,1)}\|_{\mathcal{E}_2} \le C\|\Phi\|_{2,\beta,\partial B}$.

The space

$$\mathcal{E}_1 = \{w = h + \lambda \partial_\tau v_{\rho,\tau_1}(z-p) : \ h \in C^{2,\beta}_{\nu_1}(B(p,1)), \ \lambda \in \mathbb{R}\}$$

is endowed with the norm $\|w\|_{\mathcal{E}_1} = \|h\|_{2,\beta,\nu_1,B(p,1)} + |\lambda|$, and the space

$$\mathcal{E}_2 = \{w = h + \lambda \partial_\tau v_{\rho,\tau_2,\gamma}(z-q) + 2a\cdot\partial_{\bar{z}} v_{\rho,\tau_2,\gamma}(z-q) : h \in C^{2,\beta}_{\nu_2}(B(q,1)), \lambda \in \mathbb{R}, a \in \mathbb{C}\}$$

with the norm $\|w\|_{\mathcal{E}_2} = \|h\|_{2,\beta,\nu_2,B(q,1)} + |\lambda| + |a|$.

Define

$$\begin{aligned} T^1_\rho : C^{2,\beta}(\partial B(p,1)) &\to C^{1,\beta}(\partial B(p,1)) \\ \phi_1 &\to T^1_\rho(\phi_1) = \partial_{r_1} v_\Phi \mid_{\partial B(p,1)}, \\ T^2_\rho : C^{2,\beta}(\partial B(q,1)) &\to C^{1,\beta}(\partial B(q,1)) \\ \phi_2 &\to T^2_\rho(\phi_2) = \partial_{r_2} v_\Phi \mid_{\partial B(q,1)}, \end{aligned}$$

where $\Phi = (\phi_1, \phi_2)$, $r_1 = |z-p|$, and $r_2 = |z-q|$. T^i_ρ is a uniformly bounded operator such that the following lemma holds.

LEMMA 4.9. *$T^i_\rho \to T^i_0$ as $\rho \to 0$ in the operatorial norm, where*

$$T^1_0 \phi_1 = 2\sum_{n=1}^{+\infty} n a_n \cdot e^{-in\theta}$$

with $\phi_1 = a_0 + 2\sum_{n=1}^{+\infty} a_n \cdot e^{-in\theta}$, while

$$T^2_0 \phi_2 = -2a_1 \cdot (e^{i\theta} + \bar{\gamma} e^{-i\theta}) + 2\sum_{n=2}^{+\infty} n a_n \cdot e^{-in\theta},$$

where $\phi_2 = a_0 + 2a_1 \cdot (e^{i\theta} - \bar{\gamma} e^{-i\theta}) + 2\sum_{n=2}^{+\infty} a_n \cdot e^{-in\theta}$. The variable θ denotes the angular variable of $\frac{z-p}{|z-p|}$ and $\frac{z-q}{|z-q|}$, respectively.

Remark 4.10. (1) The map $a_1 \in \mathbb{C} \to \bar{a}_1 - \gamma a_1 \in \mathbb{C}$ is invertible; see the discussion for the invertibility of ψ_0. Since $a_1 \cdot (e^{i\theta} - \bar{\gamma} e^{-i\theta}) = (\bar{a}_1 - \gamma a_1) \cdot e^{-i\theta}$, the statement of Lemma 4.9 makes good sense.

(2) The operator T^2_0 is the interior Dirichlet to Neumann map associated with Δ on $B(q,1)\setminus\{q\}$ with a first-order singularity in q, since $w = a_0 + 2a_1 \cdot (\frac{z-q}{|z-q|^2} - \bar{\gamma}\overline{z-q}) + 2\sum_{n=2}^{+\infty} a_n \cdot \overline{z-q}^n$ is a harmonic extension of ϕ_2 in $B(q,1)\setminus\{q\}$ with $\partial_{r_2} w \mid_{\partial B(q,1)} = T^2_0 \phi_2$.

Proof. Assume for simplicity that $p = 0$. Write

$$\phi_1(\theta) = a_0 + 2\sum_{n=1}^{+\infty} a_n \cdot e^{-in\theta} = \frac{\tau_1(1+\tau_1^2\rho^2)}{2(1-\tau_1^2\rho^2)} a_0 \partial_\tau v_{\rho,\tau_1}(e^{i\theta}) + 2\sum_{n=1}^{+\infty} a_n \cdot e^{-in\theta}.$$

Then $w = \frac{\tau_1(1+\tau_1^2\rho^2)}{2(1-\tau_1^2\rho^2)} a_0 \partial_\tau v_{\rho,\tau_1}(z) + 2\sum_{n=1}^{+\infty} r^n a_n \cdot e^{-in\theta} + w_1$ solves

$$\begin{cases} L_{\rho,\tau_1} w = 0 & \text{in } B(p,1), \\ w = \phi_1 & \text{on } \partial B(p,1) \end{cases}$$

if and only if w_1 solves

$$\begin{cases} L_{\rho,\tau_1} w_1 = f_1 := -\rho^2 |z-p|^{2\alpha} e^{v_{\rho,\tau_1}} (2\sum_{n=1}^{+\infty} r^n a_n \cdot e^{-in\theta}) & \text{in } B(p,1), \\ w_1 = 0 & \text{on } \partial B(p,1). \end{cases}$$

The well-known estimate $\| \sum_{n=1}^{+\infty} r^n a_n \cdot e^{-in\theta} \|_{2,\beta,1} \le C \|\phi_1\|_{2,\beta}$ implies $\|f_1\|_{0,\beta,\nu_1-2} \le C\rho^{\frac{1-\nu_1}{\alpha+1}} \|\phi_1\|_{2,\beta}$. Since $\int_0^{2\pi} f_1(re^{i\theta}) d\theta = 0$ for all $r \in (0,1]$, by Proposition 4.1 $w_1(z) = G_{\rho,\tau_1}(f_1)(z)$ with $\|G_{\rho,\tau_1}(f_1)\|_{2,\beta,\nu_1} \le C\rho^{\frac{1-\nu_1}{\alpha+1}} \|\phi_1\|_{2,\beta}$. Therefore $\|\partial_r w_1 \mid_{\partial B(p,1)}\|_{1,\beta} \le C\rho^{\frac{1-\nu_1}{\alpha+1}} \|\phi_1\|_{2,\beta}$ and hence

$$\begin{aligned} T_\rho^1 \phi_1 &= \frac{\tau_1(1+\tau_1^2\rho^2)}{2(1-\tau_1^2\rho^2)} a_0 \partial_r \partial_\tau v_{\rho,\tau_1}(z) \mid_{\partial B(p,1)} + 2\sum_{n=1}^{+\infty} n a_n \cdot e^{-in\theta} + O\left(\rho^{\frac{1-\nu_1}{\alpha+1}} \|\phi_1\|_{2,\beta}\right) \\ &= \frac{4(\alpha+1)\tau_1^2\rho^2}{(1+\tau_1^2\rho^2)(1-\tau_1^2\rho^2)} a_0 + 2\sum_{n=1}^{+\infty} n a_n \cdot e^{-in\theta} + O\left(\rho^{\frac{1-\nu_1}{\alpha+1}} \|\phi_1\|_{2,\beta}\right) \\ &= T_0^1 \phi_1 + O\left(\rho^{\frac{1-\nu_1}{\alpha+1}} \|\phi_1\|_{2,\beta}\right). \end{aligned}$$

Assuming for simplicity that $q = 0$, for ϕ_2 as above we can write

$$\begin{aligned} \phi_2(\theta) &= \frac{\tau_2}{2} a_0 \partial_\tau v_{\rho,\tau_2,\gamma}(e^{i\theta}) + \frac{2\tau_2^2\rho^2}{\tau_2^2\rho^2 + |1+\gamma e^{2i\theta}|^2} a_0 - a_1 \cdot \partial_{\bar z} v_{\rho,\tau_2,\gamma}(e^{i\theta}) \\ &\quad + \frac{2\tau_2^2\rho^2}{\tau_2^2\rho^2 + |1+\gamma e^{2i\theta}|^2} a_1 \cdot \frac{1+3\bar\gamma e^{-2i\theta}}{1+\bar\gamma e^{-2i\theta}} e^{i\theta} + a_1 \cdot \eta^\perp(e^{i\theta}) + 2\sum_{n=2}^{+\infty} a_n \cdot e^{-in\theta}. \end{aligned}$$

Let $h(z) := \frac{2\tau_2^2\rho^2|z|^2}{\tau_2^2\rho^2+|1+\gamma z^2|^2} a_0 + \frac{2\tau_2^2\rho^2|z|^2}{\tau_2^2\rho^2+|1+\gamma z^2|^2} a_1 \cdot \frac{1+3\bar\gamma\bar z^2}{1+\bar\gamma\bar z^2} z$, then

$$w(z) = \frac{\tau_2}{2} a_0 \partial_\tau v_{\rho,\tau_2,\gamma}(z) - a_1 \cdot \partial_{\bar z} v_{\rho,\tau_2,\gamma}(z) + h(z) + a_1 \cdot \eta^\perp(z) + 2\sum_{n=2}^{+\infty} a_n \cdot \bar z^n + w_1$$

solves

$$\begin{cases} L_{\rho,\tau_2,\gamma} w = 0 & \text{in } B(q,1), \\ w = \phi_2 & \text{on } \partial B(q,1) \end{cases}$$

if and only if w_1 solves

$$\begin{cases} L_{\rho,\tau_2,\gamma} w_1 = f_2 := -\rho^2 e^{v_{\rho,\tau_2,\gamma}} (2\sum_{n=1}^{+\infty} a_n \cdot \bar z^n + a_1 \cdot \eta^\perp(z)) - L_{\rho,\tau_2,\gamma} h & \text{in } B(q,1), \\ w_1 = 0 & \text{on } \partial B(q,1). \end{cases}$$

Since $\|2\sum_{n=2}^{+\infty} a_n \cdot \bar{z}^n + a_1 \cdot \eta^{\perp}(z)\|_{2,\beta,2} \le C\|\phi_2\|_{2,\beta}$, we get $\|f_2\|_{0,\beta,\nu_2-2} \le C\rho^{2-\nu_2}\|\phi_2\|_{2,\beta}$. By Proposition 4.5,

$$w_1(z) = G_{\rho,\tau_2,\gamma}(f_2)(z) + H^0_{\rho,\tau_2,\gamma}(f_2)\partial_\tau v_{\rho,\tau_2,\gamma}(z) + 2H^1_{\rho,\tau_2,\gamma}(f_2)\cdot\partial_{\bar{z}} v_{\rho,\tau_2,\gamma}(z)$$

with $\|G_{\rho,\tau_2,\gamma}(f_2)\|_{2,\beta,\nu_2} + |H^0_{\rho,\tau_2,\gamma}(f_2)| + |H^1_{\rho,\tau_2,\gamma}(f_2)| \le C\rho^{2-\nu_2}\|\phi_2\|_{2,\beta}$.

Therefore $\|\partial_r w_1 \mid_{\partial B(q,1)}\|_{1,\beta} \le C\rho^{2-\nu_2}\|\phi_2\|_{2,\beta}$, and so

$$\begin{aligned} T^2_\rho\phi_2 = {} & \frac{\tau_2}{2}a_0\partial_r\partial_\tau v_{\rho,\tau_2,\gamma} \mid_{\partial B(q,1)} - a_1\cdot\partial_r\partial_{\bar{z}} v_{\rho,\tau_2,\gamma} \mid_{\partial B(q,1)} + a_1\cdot\partial_r\eta^{\perp} \mid_{\partial B(q,1)} \\ & + 2\sum_{n=2}^{+\infty} n a_n\cdot e^{-in\theta} + O\left(\rho^{2-\nu_2}\|\phi_2\|_{2,\beta}\right). \end{aligned}$$

By direct computation, we find $\partial_r\partial_\tau v_{\rho,\tau_2,\gamma} \mid_{\partial B(q,1)} = O(\rho^2)$ and

$$\begin{aligned} \partial_r\partial_{\bar{z}} v_{\rho,\tau_2,\gamma} \mid_{\partial B(q,1)} &= \partial_r\left(-\frac{2}{r}e^{i\theta} + 2\bar{\gamma}re^{-i\theta} + \eta^{\perp}(re^{i\theta})\right) \mid_{r=1} + O(\rho^2) \\ &= 2(e^{i\theta} + \bar{\gamma}e^{-i\theta}) + \partial_r\eta^{\perp}(re^{i\theta}) \mid_{r=1} + O(\rho^2). \end{aligned}$$

Consequently, $T^2_\rho\phi_2 = T^2_0\phi_2 + O(\rho^{2-\nu_2}\|\phi_2\|_{2,\beta})$ and the proof of the lemma is completed. □

Define

$$\begin{aligned} T_\rho : C^{2,\beta}(\partial B) &\to C^{1,\beta}(\partial B) \\ \Phi = (\phi_1,\phi_2) &\to T_\rho\Phi = (T^1_\rho\phi_1, T^2_\rho\phi_2) \end{aligned}$$

and similarly the operator T_0. We want to prove the following lemma.

LEMMA 4.11. *There exists $\rho_0 > 0$ small such that the operator $S_\rho - T_\rho$ is invertible with uniformly bounded inverse for $\rho \in (0,\rho_0)$.*

Proof. Since $S_\rho - T_\rho \to S_0 - T_0$ as $\rho \to 0$ in the operatorial norm, we want to prove that $S_0 - T_0$ is invertible. By an idea of R. Mazzeo used in [4], we claim that it is enough to prove that $S_0 - T_0$ is injective. Regarding $S_0 - T_0$ as an operator from $H^1(\partial B)$ into $L^2(\partial B)$, it is a self-adjoint first-order pseudodifferential operator. Since S_0 and T_0 are elliptic with principal symbols $-|\xi|$ and $|\xi|$, respectively, the difference $S_0 - T_0$ is also elliptic and semibounded. Hence, $S_0 - T_0$ has a discrete spectrum and the invertibility reduces to prove injectivity. The invertibility in Hölder spaces then will follow by classical regularity theory. Let $\Phi \in H^1(\partial B)$ such that $(S_0 - T_0)\Phi = 0 \in L^2(\partial B)$. In view of (2) in Remark 4.10, by Lemmas 4.8 and 4.9 there exists a solution w_0 for the problem

$$\begin{cases} \Delta w_0 = 0 & \text{in } \Omega\setminus S, \\ w_0 = 0 & \text{on } \partial\Omega, \\ w_0 = \Phi & \text{on } \partial B, \end{cases}$$

such that

$$w(z) = \begin{cases} s_1 + 2q_1\cdot\overline{z-p} + O(|z-p|^2) & \text{as } z\to p, \\ s_2 + 2q_2\cdot\left(\frac{z-q}{|z-q|^2} - \bar{\gamma}\overline{z-q}\right) + O(|z-q|^2) & \text{as } z\to q \end{cases}$$

for some $s_i \in \mathbb{R}$ and $q_i \in \mathbb{C}$. The assumption $(S_0 - T_0)\Phi = 0$ ensures that we are gluing harmonic functions in $\tilde{\Omega}$ and B which coincide with their normal derivative on ∂B. In this way the resulting function is harmonic in $\Omega\setminus S$.

In case $S = \{p\}$, w_0 is bounded near p and it extends to a harmonic function in Ω with homogeneous Dirichlet boundary condition, hence $w_0 = 0$, $\Phi = 0$, and the injectivity of $S_0 - T_0$ is proved. In the remaining cases S contains the point $q \neq p$, the solution w_0 must be equal to $8\pi q_2 \cdot \partial_{\bar{z}'} G(z,q)$ because their difference is a harmonic function in $\Omega \setminus S$ with removable singularities. Moreover, there holds $2\partial_z \left(q_2 \cdot \partial_{\bar{z}'} H(z,q)\right) |_{z=q} = -\frac{q_2\gamma}{4\pi}$ which can be rewritten as follows:

$$q_2 \partial_{zz'} H(q,q) + \bar{q}_2 \partial_{z\bar{z}'} H(q,q) = -\frac{q_2\gamma}{4\pi}. \tag{33}$$

Let us recall that, if $S = \{q\}$, $\mathcal{F}(z) = H(z,z) + \frac{1}{4\pi} \ln \left(|z-p|^{2\alpha} f(z)\right)$ and $\gamma = 4\pi \partial_{zz} H(q,q) + \frac{1}{2}\left[\partial_{zz}\left(|z-p|^{2\alpha} f(z)\right)\right](q)$; while if $S = \{p,q\}$, $\mathcal{F}(z) = H(z,z) + \frac{1}{4\pi} \ln(|z-p|^{2\alpha} f(z)) + 2(1+\alpha) G(z,p)$ and $\gamma = 4\pi \partial_{zz} H(q,q) + \frac{1}{2}\left[\partial_{zz}\left(|z-p|^{2\alpha} f(z)\right)\right](q) + 4\pi(1+\alpha)\partial_{zz} G(q,p)$. Hence (33) is equivalent to $D^2 \mathcal{F}(q)(\frac{q_2}{q_2}) = 0$ when we assume further that $\Delta \ln f(q) = 0$. The assumption that q is a nondegenerate critical point for $\mathcal{F}(z)$ provides $q_2 = 0$. Then w_0 is not singular in the points of S and as before $w_0 = 0$, $\Phi = 0$, and the injectivity of $S_0 - T_0$ follows. □

We are now in position to give the proof of Theorem 4.6.

Proof of Theorem 4.6. By Lemma 4.7 and Propositions 4.1 and 4.5, for any $f \in Y$ we can find $w_{\rm ext} \in C^{2,\beta}(\tilde{\Omega})$ and $w_{{\rm int},i} \in \mathcal{E}_i$, $i = 1,2$, which solve

$$\begin{cases} \mathcal{L}_\rho w_{\rm ext} = f & \text{in } \tilde{\Omega}, \\ w_{\rm ext} = 0 & \text{on } \partial\tilde{\Omega}, \end{cases} \quad \begin{cases} \mathcal{L}_\rho w_{\rm int},1 = f & \text{in } B(p,1), \\ w_{\rm int},1 = 0 & \text{on } \partial B(p,1), \end{cases} \quad \begin{cases} \mathcal{L}_\rho w_{\rm int},2 = f & \text{in } B(q,1), \\ w_{\rm int},2 = 0 & \text{on } \partial B(q,1). \end{cases}$$

Moreover, $\|w_{\rm ext}\|_{2,\beta,\tilde{\Omega}} + \sum_i \|w_{\rm int},i\|_{\mathcal{E}_i} \leq C\|f\|_Y$. By Lemma 4.11, we find $\Phi \in C^{2,\beta} \cdot (\partial B)$ such that

$$(S_\rho - T_\rho)\Phi = \left(-\partial_{r_1}(w_{\rm ext} - w_{\rm int},1) \,|_{\partial B(p,1)}, -\partial_{r_2}(w_{\rm ext} - w_{\rm int},2) \,|_{\partial B(q,1)}\right)$$

with $\|\Phi\|_{C^{2,\beta}(\partial B)} \leq C\|f\|_Y$. At this point, we define $w_{\rm ker} \in C(\Omega \setminus S)$ by solving

$$\begin{cases} \mathcal{L}_\rho w_{\rm ker} = 0 & \text{in } \Omega \setminus \partial B, \\ w_{\rm ker} = 0 & \text{on } \partial\Omega, \\ w_{\rm ker} = \Phi & \text{on } \partial B. \end{cases}$$

Define

$$w(z) = \begin{cases} w_{\rm ext}(z) + w_{\rm ker}(z) & \text{in } \tilde{\Omega}, \\ w_{\rm int},1(z) + w_{\rm ker}(z) & \text{in } B(p,1), \\ w_{\rm int},2(z) + w_{\rm ker}(z) & \text{in } B(q,1). \end{cases}$$

Since the external and internal normal derivative of $w(z)$ on ∂B coincide, we conclude that $w(z)$ is a solution for the problem

$$\begin{cases} \mathcal{L}_\rho w = f & \text{in } \Omega, \\ w = 0 & \text{on } \partial\Omega, \\ w \in C^{2,\beta}(\Omega \setminus S). \end{cases} \tag{34}$$

It remains to discuss the uniqueness of w: let w' be another solution of (34). Set $\Phi' = (w' \,|_{\partial B(p,1)}, w' \,|_{\partial B(q,1)})$. Then $(S_\rho - T_\rho)\Phi' = (S_\rho - T_\rho)\Phi = 0$ and, by injectivity of $S_\rho - T_\rho$, we deduce $\Phi' = \Phi$ and so $w' = w$. □

4.4. The linearized operator. Now we want to pass the information on the invertibility of $\mathcal{L}_\rho$ to $\Lambda_\rho = \Delta + W_\rho$, where $W_\rho(z) = \rho^2|z-p|^{2\alpha} f(z) e^{v(\rho,0,0)(z)}$, and in turn to $L_{(0,0,0)}$. To an element $(h,\lambda,a) \in \mathcal{E}$ we associate in a canonical way the function

$$\begin{aligned} w(z) = h(z) &+ \chi(z-p)\lambda_1 \partial_\tau v_{\rho,\tau_1}(z-p) + \chi(z-q)\lambda_2 \partial_\tau v_{\rho,\tau_2,\gamma}(z-q) \\ &+ \chi(z-q) 2a \cdot \partial_{\bar z} v_{\rho,\tau_2,\gamma}(z-q) \end{aligned}$$

(with the understanding that $\lambda_1 = 0$ if $p \notin S$ and $\lambda_2 = 0$, $a = 0$ if $q \notin S$) and we want to evaluate the difference $\Lambda_\rho - \mathcal{L}_\rho$ on $w(z)$. We have

$$\begin{aligned} \|(\Lambda_\rho - \mathcal{L}_\rho) w\|_Y \le C\rho^2 (\|h\|_{2,\beta,\tilde\Omega} + |\lambda| + |a|) \\ + \|\rho^2 |z-p|^{2\alpha} (f(z) e^{v(\rho,0,0)(z)} - e^{v_{\rho,\tau_1}(z-p)}) w\|_{0,\beta,\nu_1-2,B(p,1)} \\ + \|\rho^2 (|z-p|^{2\alpha} f(z) e^{v(\rho,0,0)(z)} - e^{v_{\rho,\tau_2,\gamma}(z-q)}) w\|_{0,\beta,\nu_2-2,B(q,1)}. \end{aligned}$$

Therefore,

$$(\Lambda_\rho - \mathcal{L}_\rho)(z) = \begin{cases} O\left(\rho^2 |z-p|^{2\alpha+1} e^{v_{\rho,\tau_1}(z-p)}\right) & \text{in } B(p,1), \\ O\left(\rho^2 |z-q|^3 e^{v_{\rho,\tau_2,\gamma}(z-q)}\right) & \text{in } B(q,1) \end{cases}$$

in view of (11), (12), and (13). Since $|\partial_\tau v_{\rho,\tau,\lambda}(z)| + |z||\partial_{\bar z} v_{\rho,\tau,\lambda}(z)| = O(1)$ in $B(0,1)$ when $\lambda \in \{0,\gamma\}$, we deduce $\|(\Lambda_\rho - \mathcal{L}_\rho) w\|_{0,\beta,\nu_1-2,B(p,1)} + \|(\Lambda_\rho - \mathcal{L}_\rho) w\|_{0,\beta,\nu_2-2,B(q,1)} = O\left(r^s \|(h,\lambda,a)\|_{\mathcal{E}'}\right)$, where $r := \max\{r_1, r_2\}$ and $s = \min\{1-\nu_1, 2-\nu_2\} > 0$. This implies $\|(\Lambda_\rho - \mathcal{L}_\rho) w\|_Y \le C r^s \|(h,\lambda,a)\|_{\mathcal{E}'}$. Note that Theorem 4.6 can be restated as follows: for any $f \in Y$ there exists $(h_0,\lambda_0,a_0) = \mathcal{L}_\rho^{-1} f \in \mathcal{E}$ such that

$$\begin{aligned} w_0(z) = h_0(z) &+ \chi(z-p)(\lambda_0)_1 \partial_\tau v_{\rho,\tau_1}(z-p) + \chi(z-q)(\lambda_0)_2 \partial_\tau v_{\rho,\tau_2,\gamma}(z-q) \\ &+ 2\chi(z-q) a_0 \cdot \partial_{\bar z} v_{\rho,\tau_2,\gamma}(z-q) \end{aligned}$$

is a solution for $\mathcal{L}_\rho w_0 = f$ in Ω and $\|(h_0,\lambda_0,a_0)\|_{\mathcal{E}} \le C\|f\|_Y$, provided $\rho > 0$ is small enough.

On the other hand, for given $(h,\lambda,a) \in \mathcal{E}$ the associated $w(z)$ solves $\Lambda_\rho w = f$ in Ω if and only if it corresponds to a fixed point for the map

$$\begin{aligned} &\mathcal{E} \to \mathcal{E} \\ &(h,\lambda,a) \to \mathcal{L}_\rho^{-1} f - \mathcal{L}_\rho^{-1} (\Lambda_\rho - \mathcal{L}_\rho) w. \end{aligned}$$

Since

$$\|\mathcal{L}_\rho^{-1} (\Lambda_\rho - \mathcal{L}_\rho) w\|_{\mathcal{E}} \le C \|(\Lambda_\rho - \mathcal{L}_\rho) w)\|_Y \le C r^s \|(h,\lambda,a)\|_{\mathcal{E}},$$

there exists ρ_0 small such that for $0 < \rho < \rho_0$ such a map defines a contraction. Thus, for any $f \in Y$ there exists a unique $(h,\lambda,a) \in \mathcal{E}$ solving $\Lambda_\rho w = f$ in Ω with $\|(h,\lambda,a)\|_{\mathcal{E}} \le C\|\mathcal{L}_\rho^{-1} f\|_{\mathcal{E}}$.

We rewrite the solution $w(z)$ in the form

$$w(z) = h'(z) + \sum_i \lambda_i \partial_{\lambda_i} v(\rho,0,0)(z) + 2(-a) \cdot \partial_{\overline{a}} v(\rho,0,0)(z)$$

with

$$\begin{aligned} h'(z) = h(z) &+ \chi(z-p)\lambda_1 \left(\partial_\tau v_{\rho,\tau_1}(z-p) - \partial_{\lambda_1} v(\rho,0,0)(z)\right) \\ &+ \chi(z-q)\lambda_2 \left(\partial_\tau v_{\rho,\tau_2,\gamma}(z-q) - \partial_{\lambda_2} v(\rho,0,0)(z)\right) \\ &+ 2\left(1-\chi(z-q)\right) a \cdot \partial_{\overline{a}} v(\rho,0,0)(z) \\ &+ 2\chi(z-q)\, a \cdot \left(\partial_{\bar{z}} v_{\rho,\tau_2,\gamma}(z-q) + \partial_{\overline{a}} v(\rho,0,0)(z)\right), \end{aligned}$$

where we have taken into account that $(1-\chi(z-p))\,\partial_{\lambda_1} v(\rho,0,0)(z)$ and $(1-\chi(z-q))\partial_{\lambda_2} v(\rho,0,0)(z)$ are identically zero. Let us compute the derivatives of $v(\rho,\lambda,a)$:

$$\begin{aligned} \partial_{\lambda_1} v(\rho,0,0)(z) &= \chi\left(\frac{z-p}{r_1}\right) \partial_\tau v_{\rho,\tau_1}(z-p), \\ \partial_{\lambda_2} v(\rho,0,0)(z) &= \chi\left(\frac{z-q}{r_2}\right) \partial_\tau v_{\rho,\tau_2,\gamma}(z-q), \end{aligned}$$

and

$$\partial_{\overline{a}} v(\rho,0,0)(z) = \begin{cases} \left(1-\chi(\frac{z-p}{r_1})\right) 8\pi \partial_{\bar{z}'} G(z,q) & \text{in } B(p,1), \\ -\partial_{\bar{z}} v_{\rho,\tau_2,\gamma}(z-q) - \partial_{\overline{a}} P_0(z) & \\ +\frac{1}{r_2}\partial_{\bar{z}}\chi(\frac{z-q}{r_2}) \left(8\pi(1+\alpha)G(z,p) + 8\pi G(z,q) - U_\rho^2(z)\right) & \\ +\left(1-\chi(\frac{z-q}{r_2})\right)(8\pi \partial_{\bar{z}'} G(z,q) & \\ \qquad +\partial_{\bar{z}} v_{\rho,\tau_2,\gamma}(z-q) + \partial_{\overline{a}} P_0(z)) & \text{in } B(q,1), \\ 8\pi \partial_{\bar{z}'} G(z,q) \quad \text{in } \tilde{\Omega}. & \end{cases}$$

Using again that $\Delta \ln f(q) = 0$, we get $\partial_{\overline{a}} P_0(z) = O(|z-q|^2)$, and so

$$\begin{aligned} \|\chi(z-p)\left(\partial_\tau v_{\rho,\tau_1}(z-p) - \partial_{\lambda_1} v(\rho,0,0)(z)\right)\|_{2,\beta,\nu_1,B(p,1)} &\le C(r_1)^{-\nu_1}, \\ \|\chi(z-q)\left(\partial_\tau v_{\rho,\tau_2,\gamma}(z-q) - \partial_{\lambda_2} v(\rho,0,0)(z)\right)\|_{2,\beta,\nu_2,B(q,1)} &\le C(r_2)^{-\nu_2}, \\ \|\chi(z-q)\left(\partial_{\bar{z}} v_{\rho,\tau_2,\gamma}(z-q) + \partial_{\overline{a}} v(\rho,0,0)(z)\right)\|_{2,\beta,\nu_2,B(q,1)} &\le C(r_2)^{-\nu_2}, \end{aligned}$$

the last estimate being valid in view of the fact that

$$8\pi \partial_{\bar{z}'} G(z,q) + \partial_{\bar{z}} v_{\rho,\tau_2,\gamma}(z-q) = 2\frac{z-q}{|z-q|^2} + O(1) - 2\frac{z-q}{|z-q|^2} + O\left(\frac{\rho^2}{|z-q|^3}\right) = O(1)$$

for any $z \in B(q,1) \setminus B(q,r_2)$. Hence $\|h'\|_{X'} \leq C\|\mathcal{L}_\rho^{-1}(f)\|_{\mathcal{E}}$ for some uniform constant $C > 0$. Thus, we have proved the following result.

THEOREM 4.12. *There exists $\rho_0 > 0$ small such that for any $\rho \in (0, \rho_0)$, we have that for any $f \in Y$ there exists a unique solution $(h, \lambda, a) \in \mathcal{E}'$ satisfying*

$$\begin{cases} \Lambda_\rho w = f \quad \text{in } \Omega, \\ w = 0 \quad \text{on } \partial\Omega, \\ w(z) = h(z) + \sum_i \lambda_i \partial_{\lambda_i} v(\rho,0,0)(z) + 2a \cdot \partial_{\overline{a}} v(\rho,0,0)(z) \end{cases}$$

with $\|(h,\lambda,a)\|_{\mathcal{E}'} \leq C\|\mathcal{L}_\rho^{-1}(f)\|_{\mathcal{E}}$ for some uniform constant $C > 0$.

Let us recall now the definition of $L_{(0,0,0)} : \mathcal{E}' \to Y$ (see section 3): for any $(h,\sigma,b) \in \mathcal{E}'$ we set

$$\begin{aligned} L_{(0,0,0)}(h,\sigma,b) = \Lambda_\rho \Big(h + \sum_i \sigma_i \partial_{\lambda_i} v(\rho,0,0) + 2b \cdot \partial_{\overline{a}} v(\rho,0,0) \Big) \\ + 2\partial_{\overline{z}} \big[\triangle v(\rho,0,0) + \rho^2 |z-p|^{2\alpha} f(z) e^{v(\rho,0,0)} \big] \\ \cdot \big[\big(b\partial_a + \overline{b}\partial_{\overline{a}} \big) \Psi(0,\cdot)^{-1} \big] . \end{aligned}$$

We have to estimate in Y the term

$$\partial_{\overline{z}} \big[\Delta v(\rho,0,0) + \rho^2 |z-p|^{2\alpha} f(z) e^{v(\rho,0,0)} \big] \cdot \big[\big(b\partial_a + \overline{b}\partial_{\overline{a}} \big) \Psi(0,\cdot)^{-1} \big] .$$

Since $\Psi(a,z) \equiv z$ for $z \in \Omega \setminus B(q,2)$, we have that $\big(b\partial_a + \overline{b}\partial_{\overline{a}} \big) \Psi(0,\cdot)^{-1} \equiv 0$ in $\Omega \setminus B(q,2)$. In view of (39) and (40) we get

$$\Big\| \partial_{\overline{z}} \big[\Delta v(\rho,0,0) + \rho^2 |z-p|^{2\alpha} f(z) e^{v(\rho,0,0)} \big] \cdot \big[\big(b\partial_a + \overline{b}\partial_{\overline{a}} \big) \Psi(0,\cdot)^{-1} \big] \Big\|_Y = o(1)|b|,$$

and so using a perturbation argument by Theorem 4.12 we derive the following result.

THEOREM 4.13. *There exists $\rho_0 > 0$ small such that for any $\rho \in (0, \rho_0)$ and $f \in Y$ there exists a unique solution $(h,\sigma,b) \in \mathcal{E}'$ satisfying*

$$\begin{cases} L_{(0,0,0)} w = f \quad \text{in } \Omega, \\ w = 0 \quad \text{on } \partial\Omega, \\ w(z) = h(z) + \sum_i \sigma_i \partial_{\lambda_i} v(\rho,0,0)(z) + 2b \cdot \partial_{\overline{a}} v(\rho,0,0)(z) \end{cases}$$

such that $\|L^{-1}_{(0,0,0)}(f)\|_{\mathcal{E}'} \leq C\|\mathcal{L}_\rho^{-1}(f)\|_{\mathcal{E}}$ for some uniform constant $C > 0$.

5. Some estimates. In order to apply a fixed point argument to K, we used in a crucial way the fact that $K : \mathcal{E}' \to \mathcal{E}'$ maps a suitable small ball into itself; see Step 4 of section 3. To obtain such information we need the estimate contained in (37) below. For this end, first we estimate the preimage through $\mathcal{L}_\rho$ of the error term $\eta = \Delta v(\rho,0,0) + \rho^2 |z-p|^{2\alpha} f(z) e^{v(\rho,0,0)}$. In $\tilde{\Omega} = \Omega \setminus B$, $v(\rho,0,0)$ is a harmonic function and hence

$$|\eta(z)| = |\rho^2 |z-p|^{2\alpha} f(z) \exp\left(8\pi(1+\alpha)G(z,p) + 8\pi G(z,q)\right)| = O(\rho^2)$$

in $\tilde{\Omega}$. In $B(p,1)$ we have that

$$\begin{aligned}\eta(z) = {} & \frac{1}{r_1^2}\Delta\chi\left(\frac{z-p}{r_1}\right)[v_{\rho,\tau_1}(z-p) - \ln f(p) - 8\pi(1+\alpha)G(z,p) - 8\pi G(z,q)] \\ & + \frac{8}{r_1}\partial_{\bar z}\chi\left(\frac{z-p}{r_1}\right)\cdot\partial_{\bar z}\left[v_{\rho,\tau_1}(z-p) - \ln f(p) - 8\pi(1+\alpha)G(z,p) - 8\pi G(z,q)\right] \\ & + \rho^2|z-p|^{2\alpha}e^{v_{\rho,\tau_1}(z-p)}\Big\{ - \chi\left(\frac{z-p}{r_1}\right) + f(z)e^{-\ln f(p)} \\ & \times \exp\left[\left(1-\chi\left(\frac{z-p}{r_1}\right)\right)(8\pi(1+\alpha)G(z,p) + 8\pi G(z,q)\right. \\ & \left. - v_{\rho,\tau_1}(z-p) + \ln f(p))\right]\Big\}\end{aligned}$$

and in $B(q,1)$

$$\begin{aligned}\eta(z) = {} & \frac{1}{r_2^2}\Delta\chi\left(\frac{z-q}{r_2}\right)[v_{\rho,\tau_2,\gamma}(z-q) - P_0(z) - 8\pi(1+\alpha)G(z,p) - 8\pi G(z,q)] \\ & + \frac{8}{r_2}\partial_{\bar z}\chi\left(\frac{z-q}{r_2}\right)\cdot\partial_{\bar z}\left[v_{\rho,\tau_2,\gamma}(z-q) - P_0(z) - 8\pi(1+\alpha)G(z,p) - 8\pi G(z,q)\right] \\ & + \rho^2 e^{v_{\rho,\tau_2,\gamma}(z-q)}\Big\{ - \chi\left(\frac{z-q}{r_2}\right) + |z-p|^{2\alpha}f(z)e^{-P_0(z)} \\ & \times \exp\left[\left(1-\chi\left(\frac{z-q}{r_2}\right)\right)(8\pi(1+\alpha)G(z,p) + 8\pi G(z,q)\right. \\ & \left. - v_{\rho,\tau_2,\gamma}(z-q) + P_0(z))\right]\Big\},\end{aligned}$$

where $P_a(z)$ is defined in section 2. In $B(q,1)$ there holds

$$\partial_{\bar z}\left(v_{\rho,\tau_2,\gamma}(z-q) - P_0(z) - 8\pi(1+\alpha)G(z,p) - 8\pi G(z,q)\right) = -\partial_{\bar z}\mathcal{F}_2(q) \tag{35}$$

$$-\overline{z-q}\,(\partial_{z\bar z}\mathcal{F}_2(q) - 2\overline{\gamma}) + O\left(|z-q|^2 + \frac{\tau_2^2\rho^2}{|z-q|^3}\right) = O\left(|z-q|^2 + \frac{\tau_2^2\rho^2}{|z-q|^3}\right)$$

in view of the fact that $\partial_z\mathcal{F}_2(q) = 0$ and $\gamma = \frac{1}{2}\partial_{zz}\mathcal{F}_2(q)$ (see section 2 for the definitions of $\mathcal{F}_2(z)$ and γ). Similarly, in $B(p,1)$ we get

$$\partial_{\bar z}\left(v_{\rho,\tau_1}(z-p) - \ln f(p) - 8\pi(1+\alpha)G(z,p) - 8\pi G(z,q)\right) = O\left(1 + \frac{\tau_1^2\rho^2}{|z-p|^{2\alpha+3}}\right).$$

As far as second derivatives are concerned, in $B(q,1)$ we have the estimate

$$\partial_{\bar z\bar z}^2\left(v_{\rho,\tau_2,\gamma}(z-q) - P_0(z) - 8\pi(1+\alpha)G(z,p) - 8\pi G(z,q)\right) = O\left(|z-q| + \frac{\tau_2^2\rho^2}{|z-q|^4}\right). \tag{36}$$

Since $\frac{\rho^2}{r_1^{2\alpha+5}} + \frac{\rho^2}{r_2^5} = O(1)$, recalling (11), (12), and (13), for any $\nu \in (0,2)$ we get the estimates

$$\|\eta\|_{0,\beta,\nu-2,B(p,1)} = O\left(r_1^{1-\nu}\right), \qquad \|\eta\|_{0,\beta,\nu-2,B(q,1)} = O\left(r_2^{3-\nu}\right).$$

Fix $0 < \delta < 1$ to be specified below. Following the notations of section 4.1, we find

$$|H^0_{\rho,\tau_1}(\eta\,|_{B(p,1)})| + \|\partial_{r_1} G_{\rho,\tau_1}(\eta\,|_{B(p,1)})\|_{1,\beta,\partial B(p,1)} = O(r_1^{1-\delta})$$

and

$$|H^0_{\rho,\tau_2}(\eta\,|_{B(q,1)})| + |H^1_{\rho,\tau_2}(\eta\,|_{B(q,1)})| + \|\partial_{r_2} G_{\rho,\tau_2}(\eta\,|_{B(q,1)})\|_{1,\beta,\partial B(q,1)} = O(r_2^{3-\delta}).$$

Moreover, choosing $\nu = \nu_1$ in $B(p,1)$ and $\nu = \nu_2$ in $B(q,1)$ we get

$$\|G_{\rho,\tau_1}(\eta\,|_{B(p,1)})\|_{2,\beta,\nu_1,B(p,1)} = O(r_1^{1-\nu_1})\,,\ \|G_{\rho,\tau_2,\gamma}(\eta\,|_{B(q,1)})\|_{2,\beta,\nu_2,B(q,1)} = O(r_2^{3-\nu_2}).$$

By Proposition 4.5 we have

$$|H^0_{\rho,\tau_2,\gamma}(\eta\,|_{B(q,1)})| + |H^1_{\rho,\tau_2,\gamma}(\eta\,|_{B(q,1)})| + \|\partial_{r_2} G_{\rho,\tau_2,\gamma}(\eta\,|_{B(q,1)})\|_{1,\beta,\partial B(q,1)} = O(r_2^{3-\delta}),$$

$$\|G_{\rho,\tau_2,\gamma}(\eta\,|_{B(q,1)})\|_{2,\beta,\nu_2,B(q,1)} = O(\|\eta\|_{0,\beta,\nu_2-2,B(q,1)}) = O(r_2^{3-\nu_2}).$$

Hence, following the notation and the construction of section 4.3, we obtain $\|w_{\rm ext}\|_{2,\beta,\tilde\Omega} \le C\|\eta\|_{0,\beta,\tilde\Omega} = O(\rho^2)$, $\|\partial_{r_1} w_{{\rm int},1}\|_{1,\beta,\partial B(p,1)} = O(r_1^{1-\delta})$, and $\|\partial_{r_2} w_{{\rm int},2}\|_{1,\beta,\partial B(q,1)} = O(r_2^{3-\delta})$.

Let $\Phi = -(S_\rho - T_\rho)^{-1}\left(\partial_{r_1}(w_{\rm ext} - w_{{\rm int},1})\ |_{\partial B(p,1)}, \partial_{r_2}(w_{\rm ext} - w_{{\rm int},2})\ |_{\partial B(q,1)}\right)$. So $\|\Phi\|_{C^{2,\beta}(\partial B)} = O(r_1^{1-\delta} + r_2^{3-\delta})$. Consequently, $\|w_{\rm ker}\|_{2,\beta,\tilde\Omega} + \|w_{\rm ker}\ |_{B(p,1)}\ \|_{\mathcal{E}_1} + \|w_{\rm ker}\ |_{B(q,1)}\ \|_{\mathcal{E}_2} = O(r_1^{1-\delta} + r_2^{3-\delta})$ which in turn implies $\|\mathcal{L}_\rho^{-1}\eta\|_{\mathcal{E}} = O(r_1^{1-\delta} + r_1^{1-\nu_1} + r_2^{3-\delta} + r_2^{3-\nu_2})$. We can choose $\nu_1 \in (0,1)$ and $\nu_2 \in (1,2)$ in such a way that $(\nu_1, 1-\nu_1) \cap (\nu_2 - 1, 2-\nu_2) \neq \emptyset$ and we can suppose that δ is fixed to belong in this set. Hence, by Theorem 4.13 we get

$$\|L^{-1}_{(0,0,0)}\eta\|_{\mathcal{E}'} = O\left(r_1^{1-\delta} + r_2^{2-\delta}\right). \tag{37}$$

Let us define $\sigma = \frac{4\alpha+5}{2\nu_1} + 1$ and choose $r_i = \rho^{\frac{1}{\sigma\nu_i}}$. In this way, we have $\frac{\rho^2}{r_1^{4\alpha+5}} = \rho^{2(1-\frac{4\alpha+5}{2\sigma\nu_1})} \to 0$ as $\rho \to 0$, $\frac{\rho^2}{r_2^5} = \rho^{2(1-\frac{5}{2\sigma\nu_2})} \to 0$ as $\rho \to 0$, and

$$\left(\sum_{i=1}^{2} r_i^{-\nu_i}\right)\left(r_1^{1-\delta} + r_2^{2-\delta}\right) = 2\left(\rho^{\frac{1-\delta-\nu_1}{\sigma\nu_1}} + \rho^{\frac{2-\delta-\nu_2}{\sigma\nu_2}}\right) \to 0 \tag{38}$$

as $\rho \to 0$.

Now, taking the derivative with respect to $\bar z$ in the expression for $\eta = \Delta v(\rho,0,0) + \rho^2|z-p|^{2\alpha} f(z) e^{v(\rho,0,0)}$ and using (13), (35), and (36), we can conclude

$$\left\|\partial_{\bar z}\left(\Delta v(\rho,0,0) + \rho^2|z-p|^{2\alpha} f(z) e^{v(\rho,0,0)}\right)\right\|_{0,\beta,B(q,2)\setminus B(q,1)} = O(\rho^2) \tag{39}$$

and

$$\left\|\partial_{\bar z}\left(\Delta v(\rho,0,0) + \rho^2|z-p|^{2\alpha} f(z) e^{v(\rho,0,0)}\right)\right\|_{0,\beta,\nu_2-2,B(q,1)} = O\left(r_2^{2-\nu_2}\right), \tag{40}$$

where we use in a crucial way the fact that $|z-p|^{2\alpha} f(z) e^{-P_0(z)} - 1 = O(|z-q|^3)$.

Acknowledgment. I express my deep gratitude to Gabriella Tarantello. I am indebted to her for the stimulating discussions and the helpful comments about the manuscript.

REFERENCES

[1] A. Bahri, Y. Y. Li, and O. Rey, *On a variational problem with lack of compactness: The topological effect of critical points at infinity*, Calc. Var. Partial Differential Equations, 3 (1995), pp. 67–93.

[2] S. Baraket, *Construction de limites singulières pour des problèmes elliptiques non linéaires en dimension deux*, C. R. Acad. Sci. Paris Sér. I Math., 323 (1996), pp. 609–614.

[3] S. Baraket, *Construction of singular limits for a strongly perturbed two-dimensional Dirichlet problem with exponential nonlinearity*, Bull. Sci. Math., 123 (1999), pp. 255–284.

[4] S. Baraket and F. Pacard, *Construction of singular limits for a semilinear elliptic equation in dimension* 2, Calc. Var. Partial Differential Equations, 6 (1998), pp. 1–38.

[5] S. Baraket and D. Ye, *Singular limit solutions for two-dimensional elliptic problems with exponentially dominated nonlinearity*, Chinese Ann. Math. Ser. B, 22 (2001), pp. 287–296.

[6] D. Bartolucci and G. Tarantello, *The Liouville equation with singular data: A concentration-compactness principle via a local representation formula*, J. Differential Equations, 185 (2002), pp. 161–180.

[7] D. Bartolucci and G. Tarantello, *Liouville type equations with singular data and their applications to periodic multivortices for the electroweak theory*, Comm. Math. Phys., 229 (2002), pp. 3–47.

[8] H. Brezis and F. Merle, *Uniform estimates and blow-up behavior for solutions of* $-\Delta u = V(x)e^u$ *in two dimensions*, Comm. Partial Differential Equations, 16 (1991), pp. 1223–1253.

[9] L. A. Caffarelli, R. Hardt, and L. Simon, *Minimal surfaces with isolated singularities*, Manuscripta Math., 48 (1984), pp. 1–18.

[10] L. A. Caffarelli and Y. Yang, *Vortex condensation in the Chern-Simons-Higgs model: An existence theorem*, Comm. Math. Phys., 168 (1995), pp. 321–336.

[11] E. Caglioti, P. L. Lions, C. Marchioro, and M. Pulvirenti, *A special class of stationary flows for two-dimensional Euler equations: A statistical mechanics description*, Comm. Math. Phys., 143 (1992), pp. 501–525.

[12] E. Caglioti, P. L. Lions, C. Marchioro, and M. Pulvirenti, *A special class of stationary flows for two-dimensional Euler equations: A statistical mechanics description.* II, Comm. Math. Phys., 174 (1995), pp. 229–260.

[13] D. Chae and O. Imanuvilov, *The existence of non-topological multivortex solutions in the relativistic self-dual Chern-Simons theory*, Comm. Math. Phys., 215 (2000), pp. 119–142.

[14] S. Chandrasekhar, *An Introduction to the Study of Stellar Structure*, Dover, New York, 1957.

[15] S. A. Chang and P. C. Yang, *A perturbation result in prescribing scalar curvature on* S^n, Duke Math. J., 64 (1991), pp. 27–69.

[16] W. Chen and C. Li, *Classification of solutions of some nonlinear elliptic equations*, Duke Math. J., 63 (1991), pp. 615–623.

[17] C. C. Chen and C. S. Lin, *On the symmetry of blowup solutions to a mean field equation*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 18 (2001), pp. 271–296.

[18] K. S. Chou and T. Y. H. Wan, *Asymptotic radial symmetry for solutions of* $\triangle u + \exp u$ *in a punctured disc*, Pacific J. Math., 163 (1994), pp. 269–276.

[19] M. G. Crandall and P. H. Rabinowitz, *Some continuation and variational methods for positive solutions of nonlinear elliptic eigenvalue problems*, Arch. Ration. Mech. Anal., 58 (1975), pp. 207–218.

[20] P. Esposito, *A Class of Liouville-Type Equations Arising in Chern–Simons Vortex Theory: Asymptotics and Construction of Blowing Up Solutions*, Ph.D. thesis, Universitá degli Studi Roma "Tor Vergata," Rome, Italy, 2003.

[21] I. M. Gelfand, *Some problems in the theory of quasilinear equations*, Amer. Math. Soc. Transl. Ser. 2, 29 (1969), pp. 295–381.

[22] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin, 1983.

[23] M. H. K. Kiessling, *Statistical mechanics of classical particles with logarithmic interaction*, Comm. Pure Appl. Math., 46 (1993), pp. 27–56.

[24] J. Liouville, *Sur l'équation aud dérivées partielles* $\partial^2 \log \lambda/\partial u \partial v \pm 2\lambda a^2 = 0$, J. de Math., 18 (1853), pp. 71–72.

[25] F. Mignot, F. Murat, and J. P. Puel, *Variation d'un point de retournement par rapport au domaine*, Comm. Partial Differential Equations, 4 (1979), pp. 1263–1297.
[26] J. L. Moseley, *A two-dimensional Dirichlet problem with an exponential nonlinearity*, SIAM J. Math. Anal., 14 (1983), pp. 934–946.
[27] K. Nagasaki and T. Suzuki, *Asymptotic analysis for a two dimensional elliptic eigenvalue problem with exponentially dominated nonlinearity*, Asymptot. Anal., 3 (1990), pp. 173–188.
[28] W. M. Ni and J. Wei, *On the location and profile of spike-layer solutions to singularly perturbed semilinear Dirichlet problems*, Comm. Pure Appl. Math., 48 (1995), pp. 731–768.
[29] M. Nolasco, *Non-topological N-vortex condensates for the self-dual Chern-Simons theory*, Comm. Pure Appl. Math., 56 (2003), pp. 1752–1780.
[30] M. Nolasco and G. Tarantello, *Double vortex condensates in the Chern-Simons-Higgs theory*, Calc. Var. Partial Differential Equations, 9 (1999), pp. 31–91.
[31] J. Prajapat and G. Tarantello, *On a class of elliptic problem in $\mathbb{R}^2$: Symmetry and uniqueness results*, Proc. Roy. Soc. Edinburgh Sect. A, 131 (2001), pp. 967–985.
[32] O. Rey, *The role of Green's function in a nonlinear elliptic equation involving the critical Sobolev exponent*, J. Funct. Anal., 89 (1990), pp. 1–52.
[33] T. Ricciardi and G. Tarantello, *Vortices in the Maxwell-Chern-Simons theory*, Comm. Pure Appl. Math., 53 (2000), pp. 811–851.
[34] J. Spruck and Y. Yang, *On multivortices in the electroweak theory* I: *Existence of periodic solutions*, Comm. Math. Phys., 144 (1992), pp. 1–16.
[35] M. Struwe and G. Tarantello, *On multivortex solutions in Chern-Simons gauge theory*, Boll. UMI, 8 (1998), pp. 109–121.
[36] T. Suzuki, *Two dimensional Emden-Fowler equation with exponential nonlinearity*, in Nonlinear Diffusion Equations and Their Equilibrium States, Gregynog, 1989, Progr. Nonlinear Differential Equations Appl. 7, 1992, pp. 493–512.
[37] G. Tarantello, *Multiple condensate solutions for the Chern-Simons-Higgs theory*, J. Math. Phys., 37 (1996), pp. 3769–3796.
[38] V. H. Weston, *On the asymptotic solution of a partial differential equation with exponential nonlinearity*, SIAM J. Math. Anal., 9 (1978), pp. 1030–1053.
[39] Y. Yang, *Solitons in Field Theory and Nonlinear Analysis*, Springer-Verlag, New York, 2001.

SIAM J. MATH. ANAL.
Vol. 36, No. 4, pp. 1346–1369

STABILITY CONDITIONS FOR STRONG RAREFACTION WAVES*

MARTA LEWICKA†

Abstract. In this paper we study a number of algebraic conditions connected with the stability of strictly hyperbolic $n \times n$ systems of conservation laws in one space dimension:

$$u_t + f(u)_x = 0.$$

Such conditions yield existence and continuity of the flow of solutions in the vicinity of the reference solution. Our main concern is a single rarefaction wave having arbitrarily large strength.

Key words. conservation laws, large data, rarefaction wave, stability conditions

AMS subject classifications. 35L65, 35L45

DOI. 10.1137/S0036141003429517

1. Introduction.

In this paper we study a number of algebraic conditions connected with the stability of strictly hyperbolic $n \times n$ systems of conservation laws in one space dimension:

$$u_t + f(u)_x = 0. \tag{1.1}$$

The well-posedness of (1.1) has been the subject of vast research in recent years; for an overview see [B, D, HR]. While most of the analysis (see [BLY] and more recently [BiB]) has been carried out in the setting of initial data

$$u(0,x) = \bar{u}(x) \tag{1.2}$$

having small total variation, at the same time examples in [BC, J] point out that for the stability of patterns containing large waves, extra assumptions are required, also when the large reference waves do not interact among themselves [BC, Scho, Le1, Le3]. These BV and L^1 stability conditions, in essence, aim at providing an estimate on the distance between a reference solution u_0 and another solution to (1.1) which is viewed as an infinitesimal perturbation of u_0. They refer to the existence of weights with respect to which the flow of the first order perturbation v generated by the linearized system

$$v_t + \mathrm{D}f(u_0)v_x + [\mathrm{D}^2 f(u_0) \cdot v] \cdot (u_0)_x = 0$$

becomes a contraction with respect to the BV or the L^1 norm, respectively, at states attained by u_0. Under these assumptions the existence of global solutions and their continuous dependence on initial data has been proven in the vicinity of patterns containing only noninteracting shocks [Le1] or being a single rarefaction wave [Le3]. The BV stability of general patterns containing shocks, contact discontinuities, and rarefaction waves was established in [Scho].

The objective of this paper is a more detailed study of the stability conditions arising when u_0 contains rarefactions. With respect to the case with only shocks

*Received by the editors June 10, 2003; accepted for publication (in revised form) March 19, 2004; published electronically March 17, 2005. This work was supported by NSF grant DMS-0306201.
http://www.siam.org/journals/sima/36-4/42951.html
†University of Chicago, Department of Mathematics, Eckhart Hall, 1118 E. 58th Street, Chicago, IL 60637 (lewicka@math.uchicago.edu).

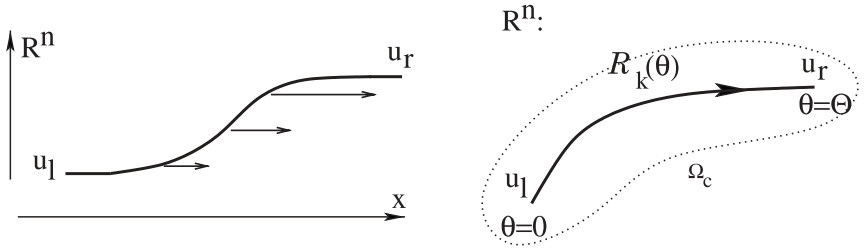

FIG. 1.1.

present [BC, Le2], the main difficulty here stems from the change of weights along rarefaction curves. This accounts for the change of location of perturbation waves of different characteristic families as they pass through each rarefaction fan. Hence we mainly focus on the case when u_0 is a single rarefaction wave of arbitrarily large strength. The stability conditions related to patterns with multiple (noninteracting) shocks and rarefaction waves are presented in section 8.

We now introduce the main hypothesis and set the notation.

(H1) The system (1.1) is strictly hyperbolic in a domain $\Omega \subset \mathbf{R}^n$ to be specified later. More precisely, for each $u \in \Omega$ the Jacobian matrix $\mathrm{D}f(u)$ of the smooth flux $f : \Omega \longrightarrow \mathbf{R}^n$ has n distinct and real eigenvalues: $\lambda_1(u) < \cdots < \lambda_n(u)$.

Let $\{r_i(u)\}_{i=1}^n$ be the basis of right eigenvectors of $\mathrm{D}f$ having unit length:

$$\mathrm{D}f(u)r_i(u) = \lambda_i(u)r_i(u), \quad ||r_i(u)|| = 1.$$

Call $\{l_i(u)\}_{i=1}^n$ the dual basis of left eigenvectors so that $\langle r_i(u), l_j(u)\rangle = \delta_{ij}$ for all $i, j : 1 \dots n$ and all $u \in \Omega$.

Fix $k : 1 \dots n$ and consider an integral curve $\mathcal{R}_k$ of the vector field r_k:

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}\theta}\mathcal{R}_k(\theta) = r_k(\mathcal{R}_k(\theta)), \\ &u_l = \mathcal{R}_k(0), \quad u_r = \mathcal{R}_k(\Theta), \qquad \Theta > 0. \end{aligned} \tag{1.3}$$

$\mathcal{R}_k$ is called the rarefaction curve joining the left and right states $u_l, u_r \in \Omega$ (see Figure 1.1). For a small $\epsilon > 0$ we define the domain

$$\Omega = \Omega_\epsilon = \{u \in \mathbf{R}^n : \quad ||u - \mathcal{R}_k(\theta)|| < \epsilon \quad \text{for some } \theta \in [0, \Theta]\}. \tag{1.4}$$

We further assume the following:

(H2) In Ω, each characteristic field $i : 1 \dots n$ either is linearly degenerate, $\langle \mathrm{D}\lambda_i, r_i\rangle \equiv 0$, or is genuinely nonlinear, which means that $\langle \mathrm{D}\lambda_i, r_i\rangle > 0$. The kth characteristic field is assumed to be genuinely nonlinear.

The piecewise smooth, self-similar function called the centered rarefaction wave is given by

$$u_0(t,x) = \begin{cases} u_l & \text{if } x < t\lambda_k(u_l), \\ \mathcal{R}_k(\theta) & \text{if } x = t\lambda_k(\mathcal{R}_k(\theta)), \quad \theta \in [0, \Theta], \\ u_r & \text{if } x > t\lambda_k(u_r) \end{cases} \tag{1.5}$$

and provides an entropy admissible solution of (1.1) [Sm, D].

The paper is constructed as follows. In section 2 we present the BV stability condition (BV) and the L^1 stability condition (L1). We also introduce a weaker condition which is sufficient for the solvability of Riemann problems in Ω. In section 3 we prove that our conditions are one stronger than the other, while sections 4, 5, and 6 gather their various properties. In particular, in section 5 we display an interesting connection between the weighted stability conditions and the Riccati equation in case $n = 3$. Section 7 contains examples complementing our work. In section 8 we restate some results of sections 2 and 3, in the context of a general pattern u_0 containing several strong shocks and rarefaction waves.

To appreciate the role of the studied conditions, we end this section by recalling the precise statements of the stability results.

THEOREM 1.1 (see [Le3]). *Assume that* (H1), (H2), *and the BV stability condition* (BV) *hold. For $c, \delta > 0$ let $\mathcal{E}_{c,\delta}$ denote the set of all continuous functions $\bar{u}$ satisfying the following:*

(i) $\bar{u}(x) \in \Omega_c$ *for all* $x \in \mathbf{R}$,
(ii) $\lim_{x\to-\infty} \bar{u}(x) = u_l$ *and* $\lim_{x\to\infty} \bar{u}(x) = u_r$,
(iii) $|TV(\bar{u}) - |\mathcal{R}_k|| < \delta$, *where* $|\mathcal{R}_k|$ *is the arc length of the rarefaction curve* $\mathcal{R}_k(\theta),\ \theta \in [0, \Theta]$.

There exists $c, \delta > 0$ such that for every $\bar{u} \in$ cl $\mathcal{E}_{c,\delta}$, where cl *denotes the closure in L^1_{loc}, the Cauchy problem* (1.1) (1.2) *has a global entropy admissible solution $u(t, x)$.*

THEOREM 1.2 (see [Le3]). *Assume that* (H1), (H2), *and the L^1 stability condition* (L1) *are satisfied. Then there exists a closed domain $\mathcal{D} \subset L^1_{loc}(\mathbf{R}, \Omega)$, containing all continuous functions $\bar{u}$ satisfying* (i), (ii), (iii) *in Theorem* 1.1, *for some $c, \delta > 0$, and there exists a semigroup $S : \mathcal{D} \times [0, \infty) \longrightarrow \mathcal{D}$ such that*

(i) $||S(\bar{u}, t) - S(\bar{v}, s)||_{L^1} \leq L \cdot (|t - s| + ||\bar{u} - \bar{v}||_{L^1})$ *for all $\bar{u}, \bar{v} \in \mathcal{D}$, all $t, s \geq 0$ and a uniform constant L, depending only on the system* (1.1),
(ii) *for all $\bar{u} \in \mathcal{D}$, the trajectory $t \mapsto S(\bar{u}, t)$ is the solution to* (1.1) (1.2) *given in Theorem* 1.1.

2. Stability conditions for strong rarefactions. Define the square $(n-1)$-dimensional production matrix function

$$\mathbf{P}(\theta) = [p_{ij}(\theta)]_{\substack{i,j:1\dots n,\\ i,j\neq k}} \quad \text{for } \theta \in [0, \Theta],$$

$$p_{ij}(\theta) = \begin{cases} |\langle l_j, [r_i, r_k]\rangle(\mathcal{R}_k(\theta))| & \text{if } i \neq j, \\ \operatorname{sgn}(k-i) \cdot \langle l_i, [r_i, r_k]\rangle(\mathcal{R}_k(\theta)) & \text{if } i = j, \end{cases} \tag{2.1}$$

where $[r_i, r_k] = \mathrm{D}r_i \cdot r_k - \mathrm{D}r_k \cdot r_i$ stands for the Lie bracket of the vector fields r_i and r_k. We then have the following:

(BV) *BV stability condition*: There exist positive smooth functions $w_1 \dots w_{k-1}, w_{k+1} \dots w_n : [0, \Theta] \to \mathbf{R}_+$ such that

$$\mathbf{P}(\theta) \cdot \begin{bmatrix} w_1(\theta) \\ \vdots \\ w_{k-1}(\theta) \\ w_{k+1}(\theta) \\ \vdots \\ w_n(\theta) \end{bmatrix} < \begin{bmatrix} w_1'(\theta) \\ \vdots \\ w_{k-1}'(\theta) \\ -w_{k+1}'(\theta) \\ \vdots \\ -w_n'(\theta) \end{bmatrix} \quad \text{for every } \theta \in (0, \Theta).$$

Here $w_i' = \mathrm{d}w_i/\mathrm{d}\theta$ and the above vector inequality holds componentwise.

Define the mass production matrix function

$$\mathbf{M}(\theta) = [m_{ij}(\theta)]_{\substack{i,j:1\dots n,\\ i,j\neq k}} \quad \text{for } \theta \in [0,\Theta],$$

$$m_{ij}(\theta) = \begin{cases} p_{ij}(\theta)\cdot \dfrac{|\lambda_j-\lambda_k|}{|\lambda_i-\lambda_k|}(\mathcal{R}_k(\theta)) & \text{if } i\neq j,\\ p_{ij}(\theta)+\dfrac{\mathrm{D}\lambda_i\cdot r_k}{|\lambda_i-\lambda_k|}(\mathcal{R}_k(\theta)) & \text{if } i=j. \end{cases} \tag{2.2}$$

Then, we have the following:

(L1) L^1 *stability condition*: There exist positive smooth functions $w_1\dots w_{k-1}, w_{k+1}\dots w_n : [0,\Theta]\to \mathbf{R}_+$ such that the inequality in (BV) is satisfied with $\mathbf{M}(\theta)$ replacing the matrix $\mathbf{P}(\theta)$.

A version of (L1), where all weights w_i are linear functions of the parameter θ, was introduced in [BM]. Condition (L1) is more general, as can be seen from Example 7.3; compare also Remark 7.4. On the other hand, (L1) holds if it is satisfied with constant and equal weights, for some rescaling of the coordinate system $\{r_i\}_{i=1}^n$ (see Corollary 4.2).

In section 3 we will prove that (L1) is stronger than the condition (BV). Below we introduce a third stability condition, guaranteeing the existence result of the type of Theorem 1.1, in the context of the Riemann initial data.

Define the $n\times n$ transport matrix function $\mathbf{T}(\theta)$ to be the solution of the following ODE system:

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}\theta}\mathbf{T}(\theta) = \mathrm{D}r_k(\mathcal{R}_k(\theta))\cdot\mathbf{T}(\theta), \quad \theta\in[0,\Theta],\\ \mathbf{T}(0) = \mathrm{Id}_n. \end{cases} \tag{2.3}$$

Also, for any $\theta_1,\theta_2\in[0,\Theta]$ with $\theta_1\le\theta_2$, let $F(\theta_1,\theta_2)$ be the $n\times n$ matrix whose columns $c_i(\theta_1,\theta_2)\in\mathbf{R}^n,\ i:1\dots n$ are given by

$$\begin{aligned} c_i(\theta_1,\theta_2) &= \mathbf{T}(\theta_2)\cdot\mathbf{T}(\theta_1)^{-1}\cdot r_i(\mathcal{R}_k(\theta_1)) \quad \text{for } i:1\dots k-1,\\ c_i(\theta_1,\theta_2) &= r_i(\mathcal{R}_k(\theta_2)) \quad \text{for } i:k\dots n. \end{aligned} \tag{2.4}$$

We may now set the following:

(F) *Finiteness condition*: For every $\theta_1,\theta_2\in[0,\Theta]$ with $\theta_1\le\theta_2$, the matrix $F(\theta_1,\theta_2)$ is invertible.

THEOREM 2.1. *Assume* (H1), (H2)*, and let the finiteness condition* (F) *hold. There exist* $\epsilon,\delta>0$ *such that for every* $u^-,u^+\in\Omega_\epsilon$ *with* $\lambda_k(u^+)-\lambda_k(u^-)>-\delta$*, the Riemann problem* (1.1) (1.2) *with*

$$\bar u = u(0,x) = \begin{cases} u^-, & x<0,\\ u^+, & x>0, \end{cases} \tag{2.5}$$

has the unique self-similar solution, attaining states inside Ω_ϵ*. The solution is composed of* $n-1$ *weak waves of families* $1\dots k-1, k+1\dots n$*, and a* k*th rarefaction wave or a weak* k*th shock.*

Proof. By a standard argument the assumptions (H1) and (H2) imply the assertion for $u^-,u^+\in\Omega_\epsilon$ such that $|\lambda_k(u^+)-\lambda_k(u^-)|<\delta$, if only δ and ϵ are small

[L, B]. We will prove that the invertibility of $F(0,\Theta)$ is sufficient for the solvability of (1.1) (2.5) whenever $||u^- - u_l|| < \delta$ and $||u^+ - u_r|| < \delta$ with a small $\delta > 0$. By a compactness argument, the proof will then be complete.

For each $i : 1 \dots n$ and $u \in \Omega$, call $\sigma \mapsto \mathcal{S}_i(u,\sigma)$ and $\sigma \mapsto \mathcal{R}_i(u,\sigma)$ the ith shock and the ith rarefaction curves through the point u [L, Sm]. In particular, by (1.3), we have $\mathcal{R}_k(u_l,\theta) = \mathcal{R}_k(\theta)$. Both curves are defined at least locally—that is, for $\sigma \in (-\epsilon,\epsilon)$—and have second order contact at $\sigma = 0$. The ith wave curve $\sigma \mapsto \mathcal{W}_i(u,\sigma)$ is obtained by taking the positive part of $\mathcal{R}_i$ ($\sigma \geq 0$) and the negative part of $\mathcal{S}_i$ ($\sigma < 0$).

Define an auxiliary $\mathcal{C}^2$ function $G(u^-,u^+,\sigma_1 \dots \sigma_n) \in \mathbf{R}^n$, whose arguments stay close to $u_l, u_r, \sigma_i = 0$ for $i \neq k$ and $\sigma_k = \Theta$, respectively:

$$G(u^-,u^+,\sigma_1 \dots \sigma_n) = \mathcal{W}_n(\sigma_n) \dots \circ \mathcal{W}_{k+1}(\sigma_{k+1}) \circ \mathcal{R}_k(\sigma_k) \\ \circ \mathcal{W}_{k-1}(\sigma_{k-1}) \dots \circ \mathcal{W}_1(u^-,\sigma_1) - u^+.$$

Notice that by (1.3) the function $\mathcal{R}_k(u,\sigma)$ is defined on $\Omega_\epsilon \times (-\epsilon, \Theta+\epsilon)$ for a small $\epsilon > 0$. We clearly have

$$\frac{\partial G}{\partial(\sigma_1 \dots \sigma_n)}(u_l, u_r, \sigma_i = 0 \text{ for } i \neq k \text{ and } \sigma_k = \Theta) = F(0,\Theta),$$

as $\mathrm{d}/\mathrm{d}\sigma \mathcal{W}_i(u,0) = r_i(u)$ and $\mathrm{d}/\mathrm{d}\sigma \mathcal{R}_k(u,0) = r_k(u)$ for every $u \in \Omega$. Since $F(0,\Theta)$ is invertible, by implicit function theorem we conclude the result. ∎

Remark 2.2. We have used the following property of the matrix $\mathbf{T}(\theta)$:

$$\mathbf{T}(\theta) \cdot r_i(u_l) = \lim_{\epsilon \to 0} \frac{\mathcal{R}_k(u_l + \epsilon r_i(u_l), \theta) - \mathcal{R}_k(\theta)}{\epsilon}. \tag{2.6}$$

For $i < k$, the left-hand side of (2.6) is equal to $c_i(0,\theta)$. Thus the first $k-1$ columns of the finiteness matrix $F(\theta_1,\theta_2)$ are equal to the eigenvectors at $\mathcal{R}_k(\theta_1)$ corresponding to characteristic families $i < k$ (slow modes), transported by the flow of the ODE (1.3) to the point $\mathcal{R}_k(\theta_2)$. The condition (F) simply says that this set of vectors can be completed by the remaining right eigenvectors at $\mathcal{R}_k(\theta_2)$ (that is, the eigenvectors corresponding to the fast modes $i \geq k$) to form a basis of $\mathbf{R}^n$. Obviously, the kth column c_k in (2.4) can be computed by any of the two formulae because the flow of (1.3) preserves the kth eigenvector: $\mathbf{T}(\theta_2) \cdot \mathbf{T}(\theta_1)^{-1} \cdot r_k(\mathcal{R}_k(\theta_1)) = r_k(\mathcal{R}_k(\theta_2))$.

We have shown that the invertibility of $F(0,\Theta)$ implies the solvability of any Riemann problem (1.1) (2.5) close to the initial data ($u^- = u_l, u^+ = u_r$). This condition is strictly weaker than (F), as shown by the Example 7.1. Also, it follows from Example 7.1 that (F) is a nontrivial condition.

3. A proof of (L1) ⇒ (BV) ⇒ (F). In this section we prove the basic relation among the three stability conditions from section 2. We first establish an abstract lemma on matrix analysis.

LEMMA 3.1. *Let $\tilde{\mathbf{P}}(\theta) = [\tilde{p}_{ij}(\theta)]_{i,j:1\dots n}$ be a continuous $n \times n$ matrix function, defined on an interval $[0,\Theta]$. Fix $k : 1 \dots n$ and define an associated matrix function $\hat{\mathbf{P}}(\theta) = [\hat{p}_{ij}(\theta)]_{i,j:1\dots n}$ by*

$$\hat{p}_{ij}(\theta) = \begin{cases} |\tilde{p}_{ij}(\theta)| & \text{if } i \neq j, \\ (\text{sgn } (i-k)) \cdot \tilde{p}_{ii}(\theta) & \text{if } i = j. \end{cases}$$

Assume that there exist positive smooth functions $w_1 \dots w_n : [0,\Theta] \to \mathbf{R}_+$ *such that the following vector inequality is satisfied componentwise:*

$$\hat{\mathbf{P}}(\theta)\cdot\begin{bmatrix} w_1(\theta)\\ \vdots\\ w_n(\theta)\end{bmatrix} < \begin{bmatrix} w_1'(\theta)\\ \vdots\\ w_{k-1}'(\theta)\\ -w_k'(\theta)\\ \vdots\\ -w_n'(\theta)\end{bmatrix} \quad \text{for every } \theta\in(0,\Theta). \tag{3.1}$$

Then we have the following:

(i) *Let* $b:[0,\Theta]\longrightarrow\mathbf{R}^n$, $b(\theta)=(b_1(\theta)\dots b_n(\theta))$ *satisfy*

$$\frac{\mathrm{d}}{\mathrm{d}\theta}b(\theta) = b(\theta)^t\cdot\tilde{\mathbf{P}}(\theta) \quad \text{for } \theta\in[0,\Theta], \tag{3.2}$$

$$\sum_{i=1}^{n}|b_i(0)|>0. \tag{3.3}$$

The above implies that

$$\sum_{i<k}\Big(|b_i(\Theta)|w_i(\Theta)-|b_i(0)|w_i(0)\Big) > \sum_{i\geq k}\Big(|b_i(\Theta)|w_i(\Theta)-|b_i(0)|w_i(0)\Big). \tag{3.4}$$

(ii) *Calling* B *the solution of the matrix differential equation,*

$$\begin{cases}\dfrac{\mathrm{d}}{\mathrm{d}\theta}B(\theta)=\tilde{\mathbf{P}}(\theta)\cdot B(\theta), \quad \theta\in[0,\Theta],\\ B(0)=\mathrm{Id}_n,\end{cases} \tag{3.5}$$

the $(k-1)\times(k-1)$ *principal minor of* $B(\Theta)$ *is invertible.*

Proof. (i). Using (3.2), (3.3), and (3.1) we obtain the following:

$$\begin{aligned}&\sum_{i<k}(\operatorname{sgn} b_i)\cdot(b_i\cdot w_i)' - \sum_{i\geq k}(\operatorname{sgn} b_i)\cdot(b_i\cdot w_i)'\\ &\quad> \sum_{i=1}^{n}\left((\operatorname{sgn} b_i)\cdot(\operatorname{sgn}(k-i))\cdot w_i\cdot\sum_{j=1}^{n}b_j\tilde{p}_{ji}\right)+\sum_{i=1}^{n}\left(|b_i|\cdot\sum_{j=1}^{n}w_j\hat{p}_{ij}\right)\\ &\quad= \left[\sum_{i\neq j}|b_i|w_j\hat{p}_{ij}+(\operatorname{sgn} b_j)(\operatorname{sgn}(k-j))\cdot b_iw_j\tilde{p}_{ij}\right]\\ &\qquad\qquad+\left[\sum_{i=1}^{n}|b_i|w_i\hat{p}_{ii}+(\operatorname{sgn}(k-i))\cdot|b_i|w_i\tilde{p}_{ii}\right]\\ &\quad\geq \left[\sum_{i\neq j}|b_iw_j\hat{p}_{ij}|-|b_iw_j\tilde{p}_{ij}|\right]+\left[\sum_{i=1}^{n}|b_i|w_i(\hat{p}_{ii}+(\operatorname{sgn}(k-i))\cdot\tilde{p}_{ii})\right].\end{aligned} \tag{3.6}$$

Since $\hat{p}_{ii}=-(\operatorname{sgn}(k-i))\tilde{p}_{ii}$ for every $i:1\dots n$, and $|\tilde{p}_{ij}|=|\hat{p}_{ij}|$ for $i\neq j$, we conclude that the right-hand side of (3.6) is nonnegative, and thus,

$$\forall\,\theta\in[0,\Theta]\quad\sum_{i<k}(\operatorname{sgn} b_i)(\theta)\cdot(b_i\cdot w_i)'(\theta) > \sum_{i\geq k}(\operatorname{sgn} b_i)(\theta)\cdot(b_i\cdot w_i)'(\theta). \tag{3.7}$$

Applying $\int_0^\Theta \mathrm{d}\theta$ to both sides of (3.7) we now arrive at (3.4).

(ii). We fix $k > 1$ and argue by contradiction. If the $(k-1) \times (k-1)$ principal minor of $B(\Theta)$ was singular, then there would exist $b : [0, \Theta] \longrightarrow \mathbf{R}^n$ satisfying (3.2), (3.3) together with

$$\forall\ i \geq k \quad b_i(0) = 0 \quad \text{and} \quad \forall\ i < k \quad b_i(\Theta) = 0. \tag{3.8}$$

In view of (3.4), the condition (3.8) now implies

$$-\sum_{i<k} |b_i(0)| w_i(0) > \sum_{i \geq k} |b_i(\Theta)| w_i(\Theta),$$

which is clearly a contradiction, as the weights $\{w_i\}$ are all positive functions. ∎

THEOREM 3.2. (BV) ⇒ (F).

Proof. It suffices to show that the existence of positive weights in (BV) implies the invertibility of the matrix $F(0, \Theta)$.

For $\theta \in [0, \Theta]$, let $R(\theta)$ denote the $n \times n$ matrix whose columns are the right eigenvectors of the matrix $\mathrm{D}f(\mathcal{R}_k(\theta))$. Obviously $R(\theta)$ is nonsingular and the rows of its inverse $R(\theta)^{-1}$ provide the basis of left eigenvectors $\{l_i(\mathcal{R}_k(\theta))\}$. It is easily seen that the invertibility of $F(0, \Theta)$ is equivalent to the invertibility of the product $R(\Theta)^{-1} \cdot F(0, \Theta)$, which is in turn equivalent to the following condition:

$$\text{The } (k-1)\times(k-1) \text{ principal minor of } R(\Theta)^{-1} \cdot \mathbf{T}(\Theta) \cdot R(0) \text{ is invertible.} \tag{3.9}$$

Recall that the transport matrix function $\mathbf{T}$ is defined in (2.3).

Let $\tilde{\mathbf{P}}(\theta) = [\tilde{p}_{ij}(\theta)]_{i,j:1\ldots n}$ be the $n \times n$ matrix function, with its coefficients given by

$$\tilde{p}_{ij}(\theta) = \langle l_j, [r_k, r_i] \rangle (\mathcal{R}_k(\theta)), \quad \theta \in [0, \Theta].$$

Let

$$B(\theta) = R(\theta)^{-1} \cdot \mathbf{T}(\theta) \cdot R(0). \tag{3.10}$$

We will show that B satisfies (3.5) on $[0, \Theta]$. Indeed, one has

$$B(0) = R(0)^{-1} \cdot \mathbf{T}(0) \cdot R(0) = R(0)^{-1} \cdot R(0) = \mathrm{Id}_n.$$

Using (3.10) and (2.3), we calculate

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\theta} B(\theta) &= \left\{ \frac{\mathrm{d}}{\mathrm{d}\theta} \left[R(\theta)^{-1} \right] \cdot \mathbf{T}(\theta) + R(\theta)^{-1} \cdot \frac{\mathrm{d}}{\mathrm{d}\theta} \mathbf{T}(\theta) \right\} \cdot R(0) \\ &= \left\{ -R(\theta)^{-1} \cdot \frac{\mathrm{d}}{\mathrm{d}\theta} \left[R(\theta) \right] \cdot R(\theta)^{-1} \cdot \mathbf{T}(\theta) + R(\theta)^{-1} \cdot \mathrm{D}r_k(\theta) \cdot \mathbf{T}(\theta) \right\} \cdot R(0) \\ &= \left\{ -R(\theta)^{-1} \cdot \frac{\mathrm{d}}{\mathrm{d}\theta} \left[R(\theta) \right] + R(\theta)^{-1} \cdot \mathrm{D}r_k(\theta) \cdot R(\theta) \right\} \cdot R(\theta)^{-1} \cdot \mathbf{T}(\theta) \cdot R(0). \end{aligned} \tag{3.11}$$

Since it is clear that

$$\tilde{\mathbf{P}}(\theta) = R(\theta)^{-1} \cdot \left[\mathrm{D}r_k(\theta) \cdot R(\theta) - \frac{\mathrm{d}}{\mathrm{d}\theta} R(\theta) \right],$$

we conclude in view of (3.11) and (3.10) that B satisfies the differential equation in (3.5).

On account of (3.9), it remains thus to prove that the condition (BV) implies the following:

(3.12) The $(k-1)\times(k-1)$ principal minor of $B(\Theta)$ is invertible.

Let $\hat{\mathbf{P}}(\theta) = [\hat{p}_{ij}(\theta)]_{i,j:1\dots n}$ be given by the formula in (2.1) for every $\theta \in [0,\Theta]$. Note that the kth row of $\hat{\mathbf{P}}(\theta)$ contains only zero elements. It is then easy to see that the condition (BV) is equivalent to the existence of positive smooth weights $w_1 \dots w_n : [0,\Theta] \longrightarrow \mathbf{R}_+$ such that (3.1) holds. Indeed, one implication is trivial, and the converse one is obtained by taking

$$w_k(\theta) = \epsilon \cdot (\Theta + 1 - \theta),$$

with $\epsilon > 0$ small enough. Now (3.1) implies (3.12) by Lemma 3.1 and our proof is complete. ∎

Remark 3.3. The implication (F) ⇒ (BV) is not true, as shown by Example 7.5.

We end this section by an easy observation.

THEOREM 3.4. (L1) ⇒ (BV).

Proof. Assume that (L1) holds. For $i \neq k$ define

$$\tilde{w}_i(\theta) = |\lambda_i(\theta) - \lambda_k(\theta)| \cdot w_i(\theta), \quad \theta \in [0,\Theta]. \tag{3.13}$$

We claim that (BV) is satisfied with weights $\{\tilde{w}_i\}_{i\neq k}$ as in (3.13). Indeed, for every $i \neq k$ we have

$$\begin{aligned}
&\left(\sum_{j\neq k} p_{ij}\tilde{w}_j\right) - (\operatorname{sgn}(k-i)) \cdot \tilde{w}_i' \\
&\quad = \left(\sum_{j\neq i,k} p_{ij} \cdot |\lambda_j - \lambda_k| \cdot \tilde{w}_j\right) + p_{ii} \cdot |\lambda_i - \lambda_k| \cdot \tilde{w}_i \\
&\qquad\qquad - \Big(\langle \mathrm{D}\lambda_k, r_k\rangle w_i - \langle \mathrm{D}\lambda_i, r_k\rangle w_i + (\lambda_k - \lambda_i) w_i'\Big) \\
&\quad = |\lambda_i - \lambda_k| \cdot \left\{ \left(\sum_{j\neq i,k} p_{ij} \cdot \frac{|\lambda_j - \lambda_k|}{|\lambda_i - \lambda_k|} \cdot \tilde{w}_j\right) + p_{ii} w_i + \frac{\langle \mathrm{D}\lambda_i, r_k\rangle}{|\lambda_i - \lambda_k|} \cdot w_i \right\} \\
&\qquad\qquad - \langle \mathrm{D}\lambda_k, r_k\rangle w_i \\
&\quad = |\lambda_i - \lambda_k| \cdot \left\{ \left(\sum_{j\neq i,k} m_{ij} w_j\right) + m_{ii} w_i - (\operatorname{sgn}(k-i)) \cdot w_i' \right\} \\
&\qquad\qquad - \langle \mathrm{D}\lambda_k, r_k\rangle w_i,
\end{aligned} \tag{3.14}$$

the last equality being a consequence of (2.2). The right-hand side of (3.14) is clearly negative in view of (L1) and the genuine nonlinearity of the kth characteristic field. This proves the theorem. ∎

4. Miscellaneous properties of (BV) and (L1). In this section we gather several useful properties of the BV and L^1 stability conditions. We mainly focus on (BV) because (L1) has the same structure, and consequently, results on (BV) can be easily translated for (L1) (see Theorem 4.6).

The next theorem states that the condition (BV) is independent of the scaling of eigenvectors $\{r_i\}_{i=1}^n$ in Ω.

THEOREM 4.1. *For every $i : 1 \dots n$ and $u \in \Omega$, define*

$$\tilde{r}_i(u) = \alpha_i(u) \cdot r_i(u),$$

where each rescaling function $\alpha_i : \Omega \longrightarrow \mathbf{R}_+$ is positive and smooth. Call $\{\tilde{l}_i\}_{i=1}^n$ the dual basis to $\{\tilde{r}_i\}_{i=1}^n$ and let $\tilde{\mathcal{R}}_k$ be the corresponding reparametrization of $\mathcal{R}_k$:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}s}\tilde{\mathcal{R}}_k(s) &= \tilde{r}_k(\tilde{\mathcal{R}}_k(s)), \\ u_l = \tilde{\mathcal{R}}_k(0), \quad u_r &= \tilde{\mathcal{R}}_k(S), \qquad S > 0. \end{aligned}$$

*Then (*BV*) holds iff there exists smooth positive weights $\{\tilde{w}_i(s)\}_{i\neq k}$, defined along the reparametrized rarefaction; $s \in [0, S]$, such that the appropriate vector inequality as in (*BV*) holds.*

Proof. Fix $s \in [0, S]$ and let $\theta \in [0, \Theta]$ be such that $\mathcal{R}_k(\theta) = \tilde{\mathcal{R}}_k(s)$. For every $i, j \neq k$ we have

$$\begin{aligned} \langle \tilde{l}_j, [\tilde{r}_i, \tilde{r}_i]\rangle(\tilde{\mathcal{R}}_k(s)) &= \Big\langle \frac{1}{\alpha_j} l_j,\ \alpha_i\alpha_k \cdot \mathrm{D}r_i \cdot r_k + \alpha_k \cdot \langle \mathrm{D}\alpha_i, r_k\rangle \cdot r_i \\ &\qquad - \alpha_i\alpha_k \cdot \mathrm{D}r_k \cdot r_i - \alpha_i \cdot \langle \mathrm{D}\alpha_k, r_i\rangle \cdot r_k \Big\rangle (\mathcal{R}_k(\theta)) \\ &= \left\{ \frac{\alpha_i}{\alpha_j}\alpha_k \cdot \langle l_j, [r_i, r_i]\rangle + \delta_{ij}\frac{\alpha_k}{\alpha_j} \cdot \langle \mathrm{D}\alpha_i, r_k\rangle \right\} (\mathcal{R}_k(\theta)). \end{aligned} \tag{4.1}$$

Define

$$\tilde{w}_i(s) = \alpha_i(\mathcal{R}_k(\theta)) \cdot w_i(\theta). \tag{4.2}$$

Since $\mathrm{d}\theta/\mathrm{d}s = \alpha_k(\tilde{\mathcal{R}}_k(s))$, by (4.1), (4.2), and (2.1) it follows for every $i \neq k$ that

$$\begin{aligned} &\left(\sum_{j\neq i,k} \tilde{w}_j(s) \cdot |\langle \tilde{l}_j, [\tilde{r}_i, \tilde{r}_k]\rangle(\tilde{\mathcal{R}}_k(s))| \right) \\ &\qquad + \tilde{w}_i(s) \cdot (\mathrm{sgn}\ (k-i)) \cdot \langle \tilde{l}_i, [\tilde{r}_i, \tilde{r}_k]\rangle(\tilde{\mathcal{R}}_k(s)) - (\mathrm{sgn}\ (k-i)) \cdot \tilde{w}_i'(s) \\ &= \left(\sum_{j\neq i,k} w_j(\theta) \cdot |\alpha_i\alpha_k \cdot \langle l_j, [r_i, r_k]\rangle|(\mathcal{R}_k(\theta)) \right) \\ &\qquad\qquad + w_i(\theta) \cdot (\mathrm{sgn}\ (k-i)) \cdot (\alpha_i\alpha_k\langle l_j, [r_i, r_k]\rangle)\,(\mathcal{R}_k(\theta)) \\ &\qquad\qquad + w_i(\theta) \cdot (\mathrm{sgn}\ (k-i)) \cdot (\alpha_k\langle \mathrm{D}\alpha_i, r_k\rangle)\,(\mathcal{R}_k(\theta)) \\ &\qquad - (\mathrm{sgn}\ (k-i)) \cdot \Big\{ w_i'(\theta) \cdot (\alpha_i\alpha_k)(\mathcal{R}_k(\theta)) + w_i(\theta) \cdot (\alpha_k\langle \mathrm{D}\alpha_i, r_k\rangle)(\mathcal{R}_k(\theta)) \Big\} \\ &= (\alpha_i\alpha_k)(\mathcal{R}_k(\theta)) \cdot \left\{ \left(\sum_{j\neq k} p_{ij}(\theta) w_j(\theta) \right) - (\mathrm{sgn}\ (k-i)) \cdot w_i'(\theta) \right\}. \end{aligned} \tag{4.3}$$

Recalling that all the rescalings α_i are positive, we obtain that the negativity of the left-hand side in (4.3) is equivalent to the inequality in (BV). This finishes the proof. ∎

COROLLARY 4.2. *The condition* (BV) *is equivalent to the following one. There exist smooth rescaling of eigenvectors* $\{r_i\}_{i\neq k}$ *along* $\mathcal{R}_k$*, given by functions* $\gamma_i : [0,\Theta] \longrightarrow \mathbf{R}_+$ *such that calling*

$$\tilde{r}_i(\mathcal{R}_k(\theta)) = \gamma_i(\theta)\cdot r_i(\mathcal{R}_k(\theta)) \text{ for } i\neq k \quad \text{and} \quad \tilde{r}_k = r_k,$$

one has for every $i\neq k$ *and every* $\theta\in[0,\Theta]$

$$\left(\sum_{j\neq k,i} |\langle \tilde{l}_j, [\tilde{r}_i,\tilde{r}_k]\rangle(\mathcal{R}_k(\theta))|\right) + (\text{sgn }(k-i))\cdot\langle \tilde{l}_i, [\tilde{r}_i,\tilde{r}_k]\rangle(\mathcal{R}_k(\theta)) < 0. \tag{4.4}$$

Above, the vectors $\{\tilde{l}_i(\mathcal{R}_k(\theta))\}_{i=1}^n$ *are the dual basis to* $\{\tilde{r}_i(\mathcal{R}_k(\theta))\}_{i=1}^n$.

Proof. If (BV) holds, then one may take

$$\gamma_i(\theta) = \frac{1}{w_i(\theta)} \qquad \text{for } i\neq k,\ \theta\in[0,\Theta].$$

On the other hand, if the functions γ_i are given, take $\alpha_i : \Omega \longrightarrow \mathbf{R}_+$ to be any smooth positive reparametrization such that

$$\alpha_i(\mathcal{R}_k(\theta)) = \gamma_i(\theta), \qquad \theta\in[0,\Theta].$$

Since the eigenvectors r_k are not to be rescaled, both implications follow now from Theorem 4.1. ∎

THEOREM 4.3. *The stability condition* (BV) *is satisfied in either of the following cases.*

(i) $k = 1$ *or* n*, that is when the wave in* (1.5) *is of the extreme characteristic field.*

(ii) Θ *is sufficiently small, that is when the wave in* (1.5) *is weak.*

Proof. (i). To fix the ideas, assume that $k = n$. Let Z be any constant $(n-1)\times(n-1)$ matrix whose components are strictly bigger than those of the matrix $\mathbf{P}(\theta)$, for all $\theta\in[0,\Theta]$. Take $w = (w_1\dots w_{k-1}, w_{k+1}\dots w_n)$ to be the solution of

$$w' = Z\cdot w, \qquad w_i(0) = 1 \text{ for } i\neq k. \tag{4.5}$$

Since the fundamental solution of (4.5) has all its components positive, each w_i must be a positive function and consequently the inequality in (BV) holds.

(ii). Define $Z(\theta) = \mathbf{P}(\theta) + \mathrm{Id}_{n-1}$ for $\theta\in[0,\Theta]$. The initial-value problem

$$Z(\theta)\cdot\begin{bmatrix} w_1\\ \vdots\\ w_{k-1}\\ w_{k+1}\\ \vdots\\ w_n\end{bmatrix}(\theta) = \begin{bmatrix} w_1'\\ \vdots\\ w_{k-1}'\\ -w_{k+1}'\\ \vdots\\ -w_n'\end{bmatrix}(\theta), \qquad w_i(0) = 1\,\forall\, i\neq k,$$

has a local solution, remaining positive on some interval $[0,\epsilon]$, and therefore satisfying (BV). ∎

Recall that the system (1.1) is said to have a coordinate system of Riemann invariants [D, Sm, S] if there exist smooth functions $v_1\dots v_n : \Omega\longrightarrow\mathbf{R}$ such that

$$\langle \mathrm{D}v_i, r_j\rangle(u) \quad \begin{cases} = 0 & \text{if } i\neq j,\\ \neq 0 & \text{if } i = j\end{cases} \qquad \text{for every } u\in\Omega. \tag{4.6}$$

Using the Frobenius theorem, one can prove (see [D]) that (4.6) implies

$$[r_i, r_j](u) \in \text{span } \{r_i, r_j\} \qquad \forall\, i,j : 1 \dots n,\ u \in \Omega.$$

Hence the matrix $\mathbf{P}(\theta)$ is diagonal for every $\theta \in [0,\Theta]$ and the inequality in (BV) becomes decoupled. Notice now that for any continuous function $a : [0,\Theta] \longrightarrow \mathbf{R}$, the differential inequality $w'(\theta) \lesseqgtr a(\theta)w(\theta)$ admits a positive solution $w(\theta) = \exp[\int_0^\theta a(s)\mathrm{d}s \mp \theta]$.

We have thus proved the following theorem.

THEOREM 4.4. *If* (1.1) *admits a system of Riemann invariants, then* (BV) *is satisfied, for every* $k : 1 \dots n$.

Remark 4.5. It is well known that every 2×2 hyperbolic system of conservation laws has a coordinate system of Riemann invariants. Therefore any rarefaction wave in such systems satisfies (BV), which is obviously also a consequence of Theorem 4.3(i).

We now restate the results of this section in the context of condition (L1); the detailed verification is left to the reader.

THEOREM 4.6. *The following assertions are true.*

(i) *The* L^1 *stability condition is independent of the scaling of the eigenvectors* $\{r_i\}_{i=1}^n$ *in* Ω. *In particular, it is equivalent to the condition formulated as in Corollary* 4.2 *with the inequality* (4.4) *replaced by*

$$\left(\sum_{j\neq k,i} \Big|(\lambda_j - \lambda_k)\cdot\langle \tilde{l}_j, [\tilde{r}_i, \tilde{r}_k]\rangle\Big|(\mathcal{R}_k(\theta))\right) + \Big((\lambda_k - \lambda_i)\cdot\langle \tilde{l}_i, [\tilde{r}_i, \tilde{r}_k]\rangle\Big)(\mathcal{R}_k(\theta)) + \langle D\lambda_i, r_k\rangle(\mathcal{R}_k(\theta)) < 0.$$

(ii) *Any extreme field* ($k = 1$ *or* n) *rarefaction, or a weak* (Θ *small) rarefaction satisfies* (L1).

(iii) *If* (1.1) *has a coordinate system of Riemann invariants, then* (L1) *holds for every* $k : 1 \dots n$.

In [Le3], the proof of Theorem 1.2 used the form of the mass production coefficients as in (2.2). They may be simplified as follows.

LEMMA 4.7. *For all* $\theta \in [0,\Theta]$ *and all* $i \neq j$ *distinct from* k *there holds*

$$m_{ij}(\theta) = |\langle l_j, \mathrm{D}r_i \cdot r_k\rangle(\mathcal{R}_k(\theta))|, \tag{4.7}$$

$$m_{ii}(\theta) = \text{sgn }(k-i)\cdot\langle l_i, \mathrm{D}r_i \cdot r_k\rangle(\mathcal{R}_k(\theta)). \tag{4.8}$$

Proof. Recall the following useful identity (see [D], pg. 126):

$$\forall j,k \quad \langle \mathrm{D}\lambda_j, r_k\rangle \cdot r_j - \langle \mathrm{D}\lambda_k, r_j\rangle \cdot r_k = \mathrm{D}f\cdot[r_j, r_k] - \lambda_j \mathrm{D}r_j \cdot r_k + \lambda_k \mathrm{D}r_k \cdot r_j. \tag{4.9}$$

Multiplying (4.9) by a left eigenvector l_i we obtain

$$\forall i \notin \{j,k\} \quad (\lambda_i - \lambda_j)\cdot\langle l_i, \mathrm{D}r_j \cdot r_k\rangle = (\lambda_i - \lambda_k)\cdot\langle l_i, \mathrm{D}r_k \cdot r_j\rangle, \tag{4.10}$$

$$\forall j \neq k \quad \langle \mathrm{D}\lambda_j, r_k\rangle = (\lambda_k - \lambda_j)\cdot\langle l_j, \mathrm{D}r_k \cdot r_j\rangle. \tag{4.11}$$

Now (4.7) follows directly from (4.10) and (4.8) is a consequence of (4.11). ∎

5. Discussion of the case $n = 3$, $k = 2$. In view of Theorem 4.3(i), every rarefaction wave (1.3) in a solution to a 2×2 system (1.1), as well as both the slowest and the fastest waves in any $n \times n$ system, is BV (and L^1) stable. In this section

we focus on intermediate field rarefactions in 3×3 systems. In particular, we show the natural correspondence between the conditions in section 2 and the solvability of certain associated Riccati equations. Using this approach we derive several sufficient conditions for (BV) (or (L1)).

Our study relies on a number of abstract matrix analysis results.

LEMMA 5.1. *Let $a, b, c, d : [0, \Theta] \longrightarrow \mathbf{R}$ be continuous functions, b and c nonnegative. Then the vector inequality*

$$\begin{bmatrix} a(\theta) & b(\theta) \\ c(\theta) & d(\theta) \end{bmatrix} \cdot \begin{bmatrix} w_1(\theta) \\ w_2(\theta) \end{bmatrix} < \begin{bmatrix} w_1'(\theta) \\ -w_2'(\theta) \end{bmatrix}, \qquad \theta \in (0, \Theta), \tag{5.1}$$

has a positive solution $w_1, w_2 : [0, \Theta] \longrightarrow \mathbf{R}_+$ iff the Riccati equation

$$v'(\theta) = b(\theta) + [a(\theta) + d(\theta)] \cdot v(\theta) + c(\theta) \cdot v(\theta)^2, \qquad \theta \in (0, \Theta), \tag{5.2}$$

has a positive solution $v : [0, \Theta] \longrightarrow \mathbf{R}_+$.

Proof. 1. If (5.1) holds, then the positive function v can be defined as w_1/w_2. Hence,

$$v' = \frac{w_1'}{w_2} - v \cdot \frac{w_2'}{w_2} > \frac{a \cdot w_1 + b \cdot w_2}{w_2} + v \cdot \frac{c \cdot w_1 + d \cdot w_2}{w_2} = b + [a + d] \cdot v + c \cdot v^2.$$

2. On the other hand, if (5.2) is satisfied for some positive function v, then the inequality

$$w'(\theta) > \epsilon + b(\theta) + [a(\theta) + d(\theta)] \cdot w(\theta) + c(\theta) \cdot w(\theta)^2$$

also has a positive solution $w : [0, \Theta] \longrightarrow \mathbf{R}_+$ if $\epsilon > 0$ is small enough. Define

$$\begin{aligned} w_2(\theta) &= \exp\left(- \int_0^\theta \frac{\epsilon}{w(s)} + d(s) + c(s) w(s) \mathrm{d}s \right), \\ w_1(\theta) &= w(\theta) \cdot w_2(\theta). \end{aligned}$$

It follows that

$$\begin{aligned} w_1' - a w_1 - b w_2 &= w' w_2 + w w_2' - a w w_2 - b w_2 \\ &= w_2 \cdot (w' + w \cdot (\ln w_2)' - a w - b) \\ &= w_2 \cdot (w' - w \cdot (\epsilon / w + d + c w) - a w - b) \\ &= w_2 \cdot \left(w' - \epsilon - b - (a + d) \cdot w - c w^2 \right) > 0 \end{aligned}$$

and

$$w_2' + c w_1 + d w_2 = w_2 \cdot ((\ln w_2)' + c w + d) = - w_2 \cdot \epsilon / w < 0.$$

Therefore, (5.1) holds. ∎

Remark 5.2. In the setting of Lemma 5.1, one can see that $v : [0, \Theta] \longrightarrow \mathbf{R}$ satisfies (5.2) iff the function $w : [0, \Theta] \longrightarrow \mathbf{R}$ defined by

$$w(\theta) = v(\theta) \cdot \exp\left(- \int_0^\theta (a + d)(s) \mathrm{d}s \right)$$

is a solution of the Riccati equation

$$
\begin{aligned}
w'(\theta) =& b(\theta)\cdot\exp\left(-\int_0^\theta (a+d)(s)\mathrm{d}s\right)\\
&+ c(\theta)\cdot\exp\left(\int_0^\theta (a+d)(s)\mathrm{d}s\right)\cdot w(\theta)^2.
\end{aligned}
\tag{5.3}
$$

Thus conditions in Lemma 5.1 are both equivalent to the following one: The initial value problem (5.3) with $w(0)=0$ has the solution defined on $[0,\Theta]$.

LEMMA 5.3. *Let $b,c:[0,\Theta]\longrightarrow \mathbf{R}_+$ be continuous nonnegative functions. Assume that*

$$
\int_0^\Theta c(\theta)\int_0^\theta b(s)\mathrm{d}s\mathrm{d}\theta < 1. \tag{5.4}
$$

Then the initial value problem

$$
\begin{cases} w'(\theta) = b(\theta) + c(\theta)\cdot w(\theta)^2,\\ w(0)=0 \end{cases} \tag{5.5}
$$

has the solution w defined on the entire interval $[0,\Theta]$.

Proof. As in the proof of Lemma 5.1, it is easy to see that the solvability of (5.5) is equivalent to the existence of positive solutions $w_1,w_2:[0,\Theta]\longrightarrow\mathbf{R}_+$ of the following system of two ODEs:

$$
\begin{cases} w_1' = bw_2,\\ w_2' = -cw_1. \end{cases} \tag{5.6}
$$

Indeed, take z to be a positive solution of the equation in (5.5) and define $w_2(\theta)=\int_0^\theta c(s)z(s)\mathrm{d}s$, $w_1(\theta)=z(\theta)w_2(\theta)$. On the other hand, given w_1 and w_2, the function $z=w_1/w_2$ clearly satisfies the ODE in (5.5).

We will prove that assuming (5.4), the solution to (5.6) with initial data

$$
w_1(0)=1, \qquad w_2(0)=C \tag{5.7}
$$

satisfies $w_2(\theta)>0$ for all $\theta\in[0,\Theta]$ if only $C>0$ is large enough. Since, consequently, $w_1>0$, the proof will be complete. We have

$$
\begin{aligned}
w_2(\theta) &= C-\int_0^\theta c(s)w_1(s)\mathrm{d}s = C-\int_0^\theta c(s)\left[1+\int_0^s b(\tau)w_2(\tau)\mathrm{d}\tau\right]\mathrm{d}s\\
&= C-\int_0^\theta c(s)\mathrm{d}s-\int_0^\theta c(s)\int_0^s b(\tau)w_2(\tau)\mathrm{d}\tau\mathrm{d}s.
\end{aligned}
\tag{5.8}
$$

Take $\epsilon\in(0,1)$ and $C>0$ such that

$$
\int_0^\Theta c(\theta)\int_0^\theta b(s)\ \mathrm{d}s\ \mathrm{d}\theta\le\epsilon \quad \text{and} \quad C-\int_0^\Theta c(\theta)\mathrm{d}\theta>\epsilon C.
$$

To obtain a contradiction, suppose that

$$
\min_{[0,\Theta]} w_2 \le 0. \tag{5.9}
$$

Then, by (5.8),

$$\begin{aligned}\max_{[0,\Theta]} w_2 = w_2(\theta_{max}) &\le C - \int_0^{\theta_{max}} c(s)\mathrm{d}s \\ &\quad - \left(\min_{[0,\Theta]} w_2\right) \cdot \int_0^{\theta_{max}} c(s) \int_0^s b(\tau)\mathrm{d}\tau\mathrm{d}s \\ &\le C - \epsilon \cdot \min_{[0,\Theta]} w_2,\end{aligned} \tag{5.10}$$

$$\begin{aligned}\min_{[0,\Theta]} w_2 = w_2(\theta_{min}) &\ge C - \int_0^{\theta_{min}} c(s)\mathrm{d}s \\ &\quad - \left(\max_{[0,\Theta]} w_2\right) \cdot \int_0^{\theta_{min}} c(s) \int_0^s b(\tau)\mathrm{d}\tau\mathrm{d}s \\ &> \epsilon C - \epsilon \cdot \max_{[0,\Theta]} w_2.\end{aligned} \tag{5.11}$$

Combining (5.10) and (5.11) we arrive at

$$\max_{[0,\Theta]} w_2 < C - \epsilon \cdot \left(\epsilon C - \epsilon \cdot \max_{[0,\Theta]} w_2\right),$$

which is equivalent to

$$\max_{[0,\Theta]} w_2 < C.$$

This contradicts (5.7) and thus we see that (5.9) cannot hold, thereby ending the proof. □

By Lemma 5.1, Remark 5.2, and Lemma 5.3, we obtain the following.

THEOREM 5.4. *When $n = 3$ and $k = 2$, then we have the following:*

(i) *The stability condition* (BV) *is equivalent to the existence of a positive solution $v : [0, \Theta] \longrightarrow \mathbf{R}_+$ of the Riccati equation*

$$v'(\theta) = p_{13}(\theta) + (p_{11}(\theta) + p_{33}(\theta)) \cdot v(\theta) + p_{31}(\theta) \cdot v(\theta)^2. \tag{5.12}$$

(ii) *In particular,* (BV) *is satisfied if*

$$\int_0^{\Theta} \int_0^{\theta} e^{\int_s^{\theta} p_{11}+p_{33}} \cdot p_{13}(s) \cdot p_{31}(\theta)\mathrm{d}s\mathrm{d}\theta < 1. \tag{5.13}$$

Remark 5.5. Condition (5.13) is certainly satisfied if p_{13} or p_{31} are equal to 0. We also see that in this case (5.12) becomes the Bernoulli or the linear equation, respectively. On the other hand, in general (5.13) is strictly weaker than the condition postulated in Theorem 5.4(i). Indeed, when $p_{11} = p_{33} = 0$ and $p_{13}(\theta) = b > 0$, $p_{31}(\theta) = c > 0$ are constant functions, then the solution to (5.12) takes the form

$$v(\theta) = \sqrt{b/c} \cdot \mathrm{tg}\left(\sqrt{bc}\theta + \mathrm{arctg}\left(v(0)/\sqrt{b/c}\right)\right).$$

Therefore the condition in (i) is here equivalent to $\Theta\sqrt{bc} < \pi/2$, while (5.13) reduces to: $\Theta^2 \cdot bc/2 < 1$. The former inequality is obviously less restrictive than the latter one.

In view of the above analysis, determining the BV stability of intermediate rarefactions in 3×3 systems of conservation laws reduces to evaluating the position of the blow-up time of the solution to (5.5). In particular the inequality (5.4) provides a sufficient condition for the blow-up to occur after the time Θ. Another proof of this result has been communicated to me by professor Ray Redheffer [R2].

Using the analysis in [R1] one can find other interesting sufficient and necessary conditions in this line. For example [R2], if $c'(0) = 0$, then

$$bc + \frac{1}{2}\left(\frac{c'}{c}\right)' - \frac{1}{4}\left(\frac{c'}{c}\right)^2 < \frac{\pi^2}{4} \qquad \text{on } [0,1] \tag{5.14}$$

implies that the corresponding solution exists on $[0,1]$. On the other hand, if (5.14) holds with a converse inequality, then the blow-up occurs at some point $\theta \leq \Theta = 1$. It can be checked that the conditions (5.14) and (5.13) are independent.

As remarked in section 4, the respective results concerning the L^1 stability condition can be easily recovered. In particular, we have the following theorem.

THEOREM 5.6. *When $n = 3$ and $k = 2$, both assertions of Theorem* 5.4 *remain valid also for the condition* (L1) *if we replace the coefficients p_{ij} in* (5.12) *and* (5.13) *by the mass production matrix coefficients m_{ij} given in* (2.2).

6. A remark for the case $n > 3$. When $n = 3$, the numbers p_{11}, p_{33}, p_{13}, and $p_{31}(\theta)$, playing roles in various conditions derived in the previous section, can be seen (in view of (2.1) and standard Taylor estimates [Sm]) as transmission and reflection coefficients in the interactions of small perturbation of families 1 and 3 with parts of the rarefaction wave $\mathcal{R}_k$ (located at θ). In this section we present a generalization of Theorem 5.4(ii) to a particular case of $n \times n$ systems (1.1) in which both transmission matrices are zero.

LEMMA 6.1. *Let k, n be natural numbers and $1 < k < n$. Let $B(\theta)$ and $C(\theta)$ be two continuous matrix functions defined on $[0, \Theta]$, with all its entries nonnegative, and of dimensions $(n-k) \times (k-1)$ and $(k-1) \times (n-k)$, respectively. Assume that*

$$\left\| \int_0^\Theta \int_0^\theta B^t(s) \cdot C^t(\theta) \mathrm{d}s \mathrm{d}\theta \right\|_1 < 1, \tag{6.1}$$

where the norm of a $m \times m$ matrix $X = [x_{ij}]_{i,j:1\ldots m}$ is defined by

$$\| X \|_1 = \max_{j:1\ldots m} \sum_{i=1}^m |x_{ij}|.$$

Then there exist positive functions $w_1 \ldots w_{k-1}, w_{k+1} \ldots w_n : [0, \Theta] \longrightarrow \mathbf{R}_+$ such that

$$B(\theta) \cdot \begin{bmatrix} w_{k+1} \\ \vdots \\ w_n \end{bmatrix}(\theta) < \begin{bmatrix} w_1' \\ \vdots \\ w_{k-1}' \end{bmatrix}(\theta), \tag{6.2}$$

$$C(\theta) \cdot \begin{bmatrix} w_1 \\ \vdots \\ w_{k-1} \end{bmatrix}(\theta) < -\begin{bmatrix} w_{k+1}' \\ \vdots \\ w_n' \end{bmatrix}(\theta), \tag{6.3}$$

componentwise, for all $\theta \in (0, \Theta)$.

Proof. We will prove that under the condition (6.1), the system of ODEs obtained by replacing the inequalities signs in (6.2) (6.3) by equalities has a positive solution $w_1 \dots w_{k-1}, w_{k+1} \dots w_n$ on $[0, \Theta]$. This will clearly complete the proof since the inequality in (6.1) is strict.

Let $w_i(0) = 1$ for all $i < k$, and $w_i(0) = C$ for all $i > k$ and some constant $C > 0$. Notice that the positivity of $w_1 \dots w_{k-1}$ is now implied by the positivity of $w_{k+1} \dots w_n$. We have, for every $\theta \in [0, \Theta]$,

$$\begin{aligned}\begin{bmatrix} w_{k+1} \\ \vdots \\ w_n \end{bmatrix}(\theta) &= \begin{bmatrix} w_{k+1} \\ \vdots \\ w_n \end{bmatrix}(0) - \int_0^\theta C(s) \cdot \begin{bmatrix} w_1 \\ \vdots \\ w_{k-1} \end{bmatrix}(s)\mathrm{d}s \\ &= \begin{bmatrix} w_{k+1} \\ \vdots \\ w_n \end{bmatrix}(0) - \int_0^\theta C(s)\mathrm{d}s \cdot \begin{bmatrix} w_1 \\ \vdots \\ w_{k-1} \end{bmatrix}(0) \\ &\qquad - \int_0^\theta C(s) \int_0^s B(\tau) \cdot \begin{bmatrix} w_{k+1} \\ \vdots \\ w_n \end{bmatrix}(\tau)\mathrm{d}\tau\mathrm{d}s. \end{aligned} \tag{6.4}$$

To prove that $w_{k+1} \dots w_n$ remain positive we argue by contradiction. Assume there exists $\theta_0 \in [0, \Theta]$ such that

$$\forall \theta \in [0, \theta_0) \quad \forall i > k \quad w_i(\theta) > 0 \quad \text{and} \quad \exists s > k \quad w_s(\theta_0) = 0. \tag{6.5}$$

Then, for every $\theta \in [0, \theta_0)$ and every $i < k$ there holds $w_i(\theta) > 0$. Hence

$$\forall \theta \in [0, \theta_0] \quad \forall i > k \quad w_i(\theta) \le w_i(0) = C.$$

Consequently, by (6.4)

$$\begin{aligned} 0 = w_s(\theta_0) &\ge C - \int_0^\Theta \sum_{j=1}^{k-1} C_{ij}(s)\mathrm{d}s - C \cdot \int_0^{\theta_0} \int_0^s \sum_{j=1}^{k-1} \left(C(s) \cdot B(\tau)\right)_{ij} \mathrm{d}\tau\mathrm{d}s \\ &\ge C - \left|\left| \int_0^\Theta C^t(s)\mathrm{d}s \right|\right|_1 - C \cdot \left|\left| \int_0^\Theta \int_0^\theta B^t(s) \cdot C^t(\theta)\mathrm{d}s\mathrm{d}\theta \right|\right|_1. \end{aligned} \tag{6.6}$$

The right-hand side of (6.6) is strictly positive for a large constant C by (6.1). This contradiction proves that θ_0 in (6.5) does not exist and the lemma follows. ∎

Recall now the definition (2.1) and take

$$\begin{aligned} A &= [p_{ij}]_{i,j:1\dots k-1}, & B &= [p_{ij}]_{\substack{i:1\dots k-1,, \\ j:k+1\dots n}} \\ C &= [p_{ij}]_{\substack{i:k+1\dots n, \\ j:1\dots k-1}} & D &= [p_{ij}]_{i,j:k+1\dots n}. \end{aligned}$$

We see that if A and D are zero matrices, then the condition (6.1) clearly implies (BV). Both this condition and (5.13) were postulated in [Scho] to be sufficient for the existence result as in Theorem 1.1. Using Lemma 6.1 to appropriate blocks of the mass production matrix $\mathbf{M}$, it is also not difficult to find the respective condition implying the L^1 stability.

In the general case, when A and D are not necessarily zero, one expects the following condition to be sufficient for (BV) to hold:

(6.7)
$$\left\| \int_0^{\Theta} \int_0^{\theta} \left[X^D(\theta) \cdot C(\theta) \cdot \left(X^{-A}(\theta)\right)^{-1} \cdot X^{-A}(s) \cdot B(s) \cdot \left(X^D(s)\right)^{-1} \right]^t \mathrm{d}s\mathrm{d}\theta \right\|_1 < 1,$$

where X^{-A} and X^D are the fundamental solutions of the ODEs

$$\begin{cases} \left(X^{-A}\right)' = -X^{-A} \cdot A, \\ X^{-A}(0) = \mathrm{Id}_{k-1} \end{cases} \qquad \begin{cases} \left(X^{D}\right)' = X^{D} \cdot D, \\ X^{D}(0) = \mathrm{Id}_{n-k.} \end{cases}$$

By a change of variables, (6.7) becomes (6.1) (now with different matrices C and B) and Lemma 6.1 can be used to recover (BV) under additional assumptions. Namely, the integrand matrix in (6.7) should have nonnegative components and the fundamental matrix $\left(X^D(\theta)\right)^{-1}$ should have positive diagonal and nonnegative off-diagonal components for each θ. This is the case when, for example, the transmission matrices A and D are diagonal.

7. Examples. In this section we present a number of examples complementing the analysis in sections 2–6. We will usually define a strictly hyperbolic matrix $\mathcal{A}(u)$, for u in a neighborhood of $\mathcal{R}_k$ given by (1.3). We set $\Theta = 1$. The right and left eigenvectors $\{r_i\}_{i=1}^n$, $\{l_i\}_{i=1}^n$ of $\mathcal{A}(u)$ will be used to compute the coefficients in $\mathbf{P}(\theta)$ or $\mathbf{T}(\theta)$. We will not necessarily have $\mathcal{A}(u) = \mathrm{D}f(u)$ for some smooth flux f.

Example 7.1. $F(0,\Theta)$ is invertible but $F(\theta_1,\theta_2)$ is not, for some $0 < \theta_1 < \theta_2 < \Theta$. Thus, in particular, the condition (F) is not satisfied.

Let $n = 3, k = 2$. Set $\mathcal{A}$ to be any strictly hyperbolic 3×3 matrix with the eigenvectors given by

$$r_1(x,y,z) = [\cos 2y,\ 0,\ \sin 2y]^t, \qquad r_2(x,y,z) = [0,\ -1,\ 0]^t,$$
$$r_3(x,y,z) = [-\sin y,\ 0,\ \cos y]^t.$$

Take $\mathcal{R}_2(\theta) = (0, 1-\theta, 0)$. Obviously $\mathbf{T} = \mathrm{Id}_3$. Therefore the matrix $F(0,1) = [r_1(0,1,0), r_2, r_3(0,0,0)]$ is invertible, but $F(1-\pi/4, 1)$ is not as $r_1(0,\pi/4,0) = r_3(0,0,0) = [0,0,1]^t$. □

Remark 7.2. In Example 7.1 take $r_2(x,y,z) = [0,1,0]^t$. Consider the rarefaction $\mathcal{R}_2(\theta) = (0,\theta,0)$ defined on $[0,1]$ and joining the same states as before, but in the reverse order. Using the analysis in section 5 one can prove that the condition (BV) is now equivalent to the existence of the nonnegative solution to the problem

$$\begin{cases} v'(y) = \dfrac{2}{\cos y} - 3(\tan y)v(y) + \dfrac{1}{\cos y}v(y)^2, \qquad y \in [0,1], \\ v(0) = 0. \end{cases}$$

The author used Maple to check that the solution exists on the whole interval $[0,1]$. Thus, in particular, (F) is satisfied along the "inverse rarefaction curve" (with respect to Example 7.1) $\mathcal{R}_2(\theta)$.

Example 7.3. The condition (BV) is satisfied but the weights $\{w_i\}_{i=1}^n$ cannot be taken to be linear.

Indeed, if we requested the weights $\{w_i\}_{i\neq k}$ in (BV) to be linear, then the condition would no longer be invariant under rescalings of the eigenvector basis (compare

Theorem 4.1). Let $n = 2, k = 2$. Take $\mathcal{A}(u)$ to be any smooth strictly hyperbolic 2×2 matrix whose right eigenvectors r_1, r_2 satisfy

$$r_1(\theta, 0) = [\sqrt{1 - \exp(2\theta - 4)},\ \exp(\theta - 2)]^t, \qquad r_2(\theta, 0) = [1\ , 0]^t.$$

By Theorem 4.3 (i), the condition (BV) must be satisfied for any rarefaction in this system. Take $\mathcal{R}_2(\theta) = (\theta, 0)$ and calculate

$$\begin{aligned} p_{11}(\theta) &= \langle \mathrm{D}r_1(\theta, 0) \cdot r_2(\theta, 0), l_1(\theta, 0)\rangle \\ &= \left[\mathrm{d}\sqrt{1 - \exp(2\theta - 4)}/\mathrm{d}\theta,\ \exp(\theta - 2)\right] \cdot \begin{bmatrix} 0 \\ \exp(2 - \theta) \end{bmatrix} = 1. \end{aligned}$$

If $w_1 > 0$ in (BV) could be taken linear, we would then have

$$p_{11} \cdot (w_1(0) + w_1' \cdot \theta) < w_1'.$$

This inequality, however, fails to be true on the interval $[1 - w_1(0)/w_1', 1)$. □

Remark 7.4. Note that all elements of the production matrix in Example 7.3 are nonnegative. This shows that the condition (BV) is indeed stronger than the BV stability version of the L^1 stability condition (3.44) from [BM], where all the second order coefficients p_{ij} (including the diagonal elements p_{ii}) are taken in the absolute value, and the existence of a linear positive solution $\{w_i\}_{i=1}^n$ to the corresponding vector inequality is asked. On the other hand, the existence of linear weights satisfying the inequality in (BV) with a matrix $\mathbf{P}$ with bigger components clearly implies our BV stability condition, which thus can be seen as a generalization of the argument in [BM].

Example 7.5. The condition (F) is satisfied but (BV) is not.

Let $n = 3, k = 2$. Take $\mathcal{A}(u = (x, y, z))$ to be a smooth 3×3 strictly hyperbolic matrix whose eigenvectors are given by

$$r_1(x, y, z) = [1, 0, 0]^t, \quad r_2(x, y, z) = [az, 1, ax]^t, \quad r_3(x, y, z) = [0, 0, 1]^t,$$

with some $a > \pi/2$. Consider the rarefaction curve $\mathcal{R}_2(\theta) = (0, \theta, 0)$. It is easy to calculate that the production matrix $\mathbf{P}$ has the form

$$\mathbf{P}(\theta) = \begin{bmatrix} 0 & a \\ a & 0 \end{bmatrix}.$$

By Remark 5.5, the condition (BV) is thus equivalent to $|a| < \pi/2$ and so it is not satisfied.

We will show that (F) is satisfied, however. Since

$$\mathrm{D}r_2(\mathcal{R}_2(\theta)) = \begin{bmatrix} 0 & 0 & a \\ 0 & 0 & 0 \\ a & 0 & 0 \end{bmatrix},$$

we have

$$\mathbf{T}(\theta) = \exp(\theta \cdot \mathrm{D}r_2) = \begin{bmatrix} \cosh(a\theta) & 0 & \sinh(a\theta) \\ 0 & 1 & 0 \\ \sinh(a\theta) & 0 & \cosh(a\theta) \end{bmatrix}.$$

Fix $0 < \theta_1 < \theta_2 < 1$. Using a version of (3.9), we see that the matrix $F(\theta_1, \theta_2)$ is invertible iff the first row–first column element of $\mathbf{T}(\theta_1)^{-1} \cdot \mathbf{T}(\theta_2)$ is nonzero. Noting that $\det \mathbf{T}(\theta) = 1$, this element can be easily computed as

$$\cosh(a\theta_1)\cosh(a\theta_2) - \sinh(a\theta_1)\sinh(a\theta_2) = \cosh(a\theta_1 - a\theta_2) > 0. \qquad \square$$

Example 7.6. The study of plane waves in a half space occupied by a hyperelastic solid leads to the following 6×6 system of hyperbolic conservation laws [TT]:

$$\begin{cases} S_x - \rho_0 V_t = 0, \\ V_x - G \cdot S_t = 0. \end{cases} \tag{7.1}$$

Here $S = (s_1, s_2, s_3)$ and $V = (v_1, v_2, v_3)$ are unknown quantities whose evolution is governed by a symmetric 3×3 matrix G containing appropriate derivatives of a sufficiently regular constitutive function $W(\sigma = s_1,\ \tau^2 = s_2^2 + s_3^2)$. The constant ρ_0 is positive. The derivation of the system, its physical relevance, and the related details can be found in [TT]. We are merely interested in verifying the BV stability condition for the rarefaction waves generated from the four intermediate characteristic fields of (7.1). Taking

$$W(\sigma, \tau^2) = \frac{\alpha}{2}\sigma^2 + \frac{\beta}{6}\sigma^3 + \frac{\delta}{4}(\tau^2)^2 \tag{7.2}$$

after a number of calculations [Mu] one arrives at explicit forms of the production matrices $\mathbf{P}$, corresponding to different rarefaction curves (which may be bounded or unbounded, depending on the initial data and the parameters of the system). Although the matrices $\mathbf{P}$ are 5×5 and in general with nonconstant coefficients, by their specific structure the inequality in (BV) can be reduced to studying different Riccati equations of the form

$$v'(\theta) = \frac{A}{B \pm \theta} \cdot (a + bv(\theta) + cv^2(\theta)). \tag{7.3}$$

By a change of variable, (7.3) is equivalent to

$$v'(s) = (a + bv(s) + cv^2(s)). \tag{7.4}$$

Since in each case $a, c > 0,\ b < 0$, and $b^2 - 4ac \geq 0$, the right-hand side of (7.4) has a positive root. Thus (7.4) has a (trivial) positive solution existing for all s. Based on this observation one obtains the BV stability of all rarefaction waves in the model (7.1) with the constitutive function (7.2). Incorporating the term $\sigma\tau^2$ in W may lead to a more complicated analysis [Mu]. $\square$

8. Stability conditions for general patterns of noninteracting large waves. In section 2 we have shown that for a single k-rarefaction the invertibility of the matrix $F(0, \Theta)$ implies the assertion of Theorem 2.1 with (u^-, u^+) close to the extreme states of the reference pattern u_0 in (1.5). For a single k-shock the corresponding property follows from the Majda stability condition [M]. It turns out that in case of multiple waves an additional finiteness condition, accounting for the mutual influence of the strong waves in u_0 is required. The analysis related to the case with strong shocks was the contents of [Le1, Le2].

Below we study the similar problem for a general pattern u_0 of M shock and rarefaction waves of different characteristic families. We also state the respective BV stability condition and prove a useful generalization of Theorem 3.2.

Let $M+1$ (with $2 \le M \le n$) distinct states $\{u_0^q\}_{q=0}^M$ in $\mathbf{R}^n$ be given. Assume that the Riemann problem (u_0^0, u_0^M) for (1.1) has a self-similar solution composed of M (large) waves $\{u_0^{q-1}, u_0^q\}_{q=1}^M$. For each $q : 1 \dots M$, the qth wave joining states (u_0^{q-1}, u_0^q) is said to belong to i_qth characteristic family and all families $i_1 < i_2 < \dots < i_M$ are genuinely nonlinear. The waves can be of two types.

(i) Stable rarefaction waves, that is,

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}\theta}\mathcal{R}_{i_q}(\theta) = r_{i_q}\left(\mathcal{R}_{i_q}(\theta)\right), \\ &u_0^{q-1} = \mathcal{R}_{i_q}(0), \quad u_0^q = \mathcal{R}_{i_q}(\Theta_q), \quad \Theta_q > 0, \end{aligned} \tag{8.1}$$

and the matrix $F_q(0, \Theta_q)$, defined as in (2.4) (2.3) with the field number i_q replacing k, is invertible.

(ii) Lax compressive, Majda stable shocks [L, M]. That is, calling Λ^q the speed of the shock we have

$$\Lambda^q \cdot (u_0^q - u_0^{q-1}) = f(u_0^q) - f(u_0^{q-1}), \tag{8.2}$$

$$\lambda_{i_q-1}(u_0^{q-1}) < \Lambda^q < \lambda_{i_q}(u_0^{q-1}) \quad \text{and} \quad \lambda_{i_q}(u_0^q) < \Lambda^q < \lambda_{i_q+1}(u_0^q), \tag{8.3}$$

$$\det\left[r_1(u_0^{q-1}) \dots r_{i_q-1}(u_0^{q-1}),\ u_0^q - u_0^{q-1},\ r_{i_q+1}(u_0^q) \dots r_n(u_0^q)\right] \neq 0. \tag{8.4}$$

We moreover assume that in a sufficiently small neighborhood of the set of states in $\mathbf{R}^n$ attained by u_0, the system (1.1) is strictly hyperbolic, with each characteristic family genuinely nonlinear or linearly degenerate.

For each $q : 0 \dots M$, let Ω^q be an open neighborhood of the state u_0^q. According to [Le2], for each shock (u_0^{q-1}, u_0^q) conditions (8.2), (8.3), (8.4) imply (and by the shock compressibility are essentially equivalent to) the existence of a constitutive function $\Psi^q : \Omega^{q-1} \times \Omega^q \longrightarrow \mathbf{R}^{n-1}$ whose zero locus is composed of pairs of states that can be joined by a stable i_q shock. Moreover, the following $n-1$ vectors are linearly independent:

$$\left\{ \frac{\partial \Psi^q}{\partial u^{q-1}}(u_0^{q-1}, u_0^q) \cdot r_i(u_0^{q-1}) \right\}_{i=1}^{i_q-1} \cup \left\{ \frac{\partial \Psi^q}{\partial u^q}(u_0^{q-1}, u_0^q) \cdot r_i(u_0^q) \right\}_{i=i_q+1}^{n}. \tag{8.5}$$

In case (u_0^{q-1}, u_0^q) is a stable rarefaction wave as in (i), the corresponding function Ψ^q can be defined as

$$\Psi^q(u^{q-1}, u^q) = (\sigma_1 \dots \sigma_{k-1}, \sigma_{k+1} \dots \sigma_n), \tag{8.6}$$

where $\{\sigma_i\}_{i=1}^n$ stand for the strengths of the waves in the solution of the Riemann problem (u^{q-1}, u^q); compare Theorem 2.1 and its proof.

For each $q : 1 \dots M$ define a $(n-1) \times (n-1)$ matrix C_q whose negative first $i_q - 1$ columns, and last $n - i_q$ columns are the vectors in (8.5). Notice that for rarefactions $C_q = \mathrm{Id}_{n-1}$ and thus C_q is invertible for each q. Call

$$\begin{aligned} F_q^{left} &= -C_q^{-1} \cdot \frac{\partial \Psi^q}{\partial u^{q-1}}(u_0^{q-1}, u_0^q) \cdot \left[r_{i_q}(u_0^{q-1}) \dots r_n(u_0^{q-1})\right], \\ F_q^{right} &= C_q^{-1} \cdot \frac{\partial \Psi^q}{\partial u^q}(u_0^{q-1}, u_0^q) \cdot \left[r_1(u_0^q) \dots r_{i_q}(u_0^q)\right]. \end{aligned} \tag{8.7}$$

By an argument as in the proof of Theorem 2.1, we see that the $(n-1) \times i_q$ matrix F_q^{right} expresses strengths of the weak outgoing waves in terms of strengths of waves

perturbing the right state of the Riemann problem (u_0^{q-1}, u_0^q). Analogously, the $(n-1)\times(n-i_q+1)$ matrix F_q^{left} corresponds to perturbations of u_0^{q-1} in the same Riemann problem.

Now define the square $M\cdot(n-1)$ dimensional finiteness matrix $\mathbf{F}$:

$$\mathbf{F}=\begin{bmatrix} [\Theta] & F_1^{right} & & & \\ F_2^{left} & [\Theta] & F_2^{right} & & \\ & F_3^{left} & [\Theta] & F_3^{right} & \\ & & \ddots & \ddots & \\ & & & F_M^{left} & [\Theta] \end{bmatrix}, \tag{8.8}$$

where $[\Theta]$ stands for the $(n-1)\times(n-1)$ zero matrix. The following is a generalization of Theorem 2.1.

(8.9) *Finiteness condition*: 1 is not an eigenvalue of the matrix $\mathbf{F}$.

THEOREM 8.1. *In the above setting let the condition* (8.9) *hold. Then any Riemann problem* $(u^-, u^+)\in\Omega^0\times\Omega^M$ *for* (1.1) *has a unique self-similar solution attaining* $n+1$ *states, consecutively connected by* $(n-M)$ *weak waves and* M *strong waves (shocks or rarefactions) joining states in different sets* Ω^q.

Proof. Define an auxiliary function

$$\begin{aligned} G:\big(\Omega^0\times\Omega^1\times\ldots\times\Omega^M\big)&\\ \times I^{i_1-1}\times I^{i_2-i_1-1}\times I^{i_3-i_2-1}\times\cdots\times I^{i_M-i_{M-1}-1}\times I^{n-i_M}&\longrightarrow\mathbf{R}^{M\cdot(n-1)}, \end{aligned}$$

$$\begin{aligned} &G((u^-,u^1,u^2\ldots u^{M-1},u^+),\\ &\quad(\sigma_1,\sigma_2\ldots\sigma_{i_1-1}),(\sigma_{i_1+1}\ldots\sigma_{i_2-1})\ldots(\sigma_{i_M+1}\ldots\sigma_n))\\ &\qquad=\Psi^1(\mathcal{W}_{i_1-1}(\sigma_{i_1-1})\ldots\circ\mathcal{W}_1(u^-,\sigma_1),\ u^1),\\ &\qquad\quad\Psi^2(\mathcal{W}_{i_2-1}(\sigma_{i_2-1})\ldots\circ\mathcal{W}_{i_1+1}(u^1,\sigma_{i_1+1}),\ u^2),\\ &\qquad\quad\ldots\\ &\qquad\quad\Psi^M(\mathcal{W}_{i_M-1}(\sigma_{i_M-1})\ldots\circ\mathcal{W}_{i_{M-1}+1}(u^{M-1},\sigma_{i_{M-1}+1}),\ u^M), \end{aligned}$$

where

$$u^+=\mathcal{W}_n(\sigma_n)\ldots\circ\mathcal{W}_{i_M+1}(u^M,\sigma_{i_M+1})$$

and I denotes a small interval in $\mathbf{R}$, containing 0. Call A the $M\cdot(n-1)$ dimensional square matrix, that is, the derivative of G with respect to the variables $(u^1\ldots u^{M-1}),(\sigma_1\ldots\sigma_n)$ at the point $\big((u_0^0\ldots u_0^M),(0\ldots0)\big)$. We will show that A is invertible iff the condition (8.9) holds, which by implicit function theorem will complete the proof.

Note first that the invertibility of A is equivalent to the invertibility of the following matrix (which without loss of generality we also call A), of the same dimension:

$$A=\begin{bmatrix} A_1 & B_1^r & & & & \\ & B_1^l & A_2 & B_2^r & & \\ & & & B_2^l & & \\ & & & \ddots & \ddots & \\ & & & & A_M & \widetilde{A}_M \end{bmatrix}. \tag{8.10}$$

Here

$$A_q = \begin{cases} \dfrac{\partial \Psi^1}{\partial u^0}\ (u_0^0, u_0^1) \cdot \big[r_1(u_0^0) \dots r_{i_1-1}(u_0^0)\big] & \text{for } q = 1, \\ \dfrac{\partial \Psi^q}{\partial u^{q-1}}\ (u_0^{q-1}, u_0^q) \cdot \big[r_{i_{q-1}+1}(u_0^{q-1}) \dots r_{i_q-1}(u_0^{q-1})\big] & \text{for } q : 2 \dots M \end{cases}$$

and

$$\widetilde{A}_M = \frac{\partial \Psi^M}{\partial u^M}\ (u_0^{M-1}, u_0^M) \cdot \big[r_{i_M+1}(u_0^M) \dots r_n(u_0^M)\big],$$

$$B_q^l = \frac{\partial \Psi^q}{\partial u^{q-1}}\ (u_0^{q-1}, u_0^q) \cdot \big[r_1(u_0^{q-1}) \dots r_n(u_0^{q-1})\big],$$

$$B_q^r = \frac{\partial \Psi^q}{\partial u^q}\ (u_0^{q-1}, u_0^q) \cdot \big[r_1(u_0^{q-1}) \dots r_n(u_0^{q-1})\big].$$

Introducing (8.7) in (8.10) and permuting the columns of A we observe that A is invertible iff the following matrix (which we again denote by A) is invertible:

$$A = \begin{bmatrix} -C_1 & C_1 \cdot F_1^{right} & & & \\ C_2 \cdot F_2^{left} & -C_2 & C_2 \cdot F_2^{right} & & \\ & & \ddots & \ddots & \\ & & & C_M \cdot F_M^{left} & -C_M \end{bmatrix}. \tag{8.11}$$

Multiplying A by the square block matrix

$$\begin{bmatrix} C_1^{-1} & & & \\ & C_2^{-1} & & \\ & & \ddots & \\ & & & C_M^{-1} \end{bmatrix},$$

we conclude that the invertibility of A in (8.11) is equivalent to the invertibility of $\mathbf{F} - \mathrm{Id}_{M\cdot(n-1)}$ and hence equivalent to (8.9). □

Remark 8.2. Let (u_0^{q-1}, u_0^q) be a stable i_q- rarefaction wave. After neglecting the i_qth rows of the two matrices

$$\begin{aligned} &F(0, \Theta_q)^{-1} \cdot \mathbf{T}_q(\Theta_q) \cdot \big[r_{i_q}(u_0^{q-1}), r_{i_q+1}(u_0^{q-1}) \dots r_n(u_0^{q-1})\big], \\ &F(0, \Theta_q)^{-1} \cdot \big[r_1(u_0^q) \dots r_{i_q-1}(u_0^q), r_{i_q}(u_0^q)\big], \end{aligned} \tag{8.12}$$

they become, respectively, F_q^{left} and F_q^{right}.

We now formulate the following:

(8.13) *BV stability condition for the wave pattern* u_0.

There exist positive continuous weights $\{w_i(u)\}_{i=1}^n$ defined on the set of states u attained by the reference solution u_0 (that is, at the isolated endpoints of shocks and along the rarefaction curves), such that for every $q : 1 \dots M$ the following holds.

(i) If (u_0^{q-1}, u_0^q) is a shock, then

$$\left|F_q^{left}\right|^t \cdot \begin{bmatrix} w_1(u_0^{q-1}) \\ \vdots \\ w_{i_q-1}(u_0^{q-1}) \\ w_{i_q+1}(u_0^q) \\ \vdots \\ w_n(u_0^q) \end{bmatrix} < \begin{bmatrix} w_{i_q}(u_0^{q-1}) \\ \vdots \\ w_n(u_0^{q-1}) \end{bmatrix}$$

$$\text{and} \quad \left|F_q^{right}\right|^t \cdot \begin{bmatrix} w_1(u_0^{q-1}) \\ \vdots \\ w_{i_q-1}(u_0^{q-1}) \\ w_{i_q+1}(u_0^q) \\ \vdots \\ w_n(u_0^q) \end{bmatrix} < \begin{bmatrix} w_1(u_0^q) \\ \vdots \\ w_{i_q}(u_0^q) \end{bmatrix},$$

where the components of a matrix $|A|$ are meant to be absolute values of the components of A, and the above vector inequality is understood componentwise.

(ii) If (u_0^{q-1}, u_0^q) is a rarefaction, then the corresponding BV stability condition (BV) is satisfied, with the production matrix $\mathbf{P}_q$ defined by (2.1) along the rarefaction curve $\mathcal{R}_q$.

Based on the results of [BM, Le1, Le3], we conjecture that the condition (8.13) implies the BV stability of the pattern u_0, in the sense of Theorem 1.1. Also, a similar weighted L^1 stability condition can be easily formulated and will imply the existence of a continuous flow of solutions, as in Theorem 1.2. Our final result is the following theorem.

THEOREM 8.3. *In the above setting, the condition* (8.13) *implies the solvability of any Riemann problem in the vicinity of* $(u_0(1, x_1), u_0(1, x_2))$ *for any* $x_1 < x_2$.

Proof. In view of Theorem 8.1, it is enough to show that (8.13) implies (8.9). By Lemma 3.3 from [Le2] and Remark 8.2, this will be achieved provided we prove the inequalities in (8.13) (i) for each rarefaction (u_0^{q-1}, u_0^q). But this indeed follows from Lemma 3.1 (i), applied to the matrix $\tilde{\mathbf{P}}$ as in the proof of Theorem 3.2. □

Acknowledgments. I thank my colleagues at the University of Chicago for providing the stimulating atmosphere which enabled me to finish this paper. I thank professor Constantine Dafermos for his encouragement and for bringing to my attention the paper [TT]. Professor Alberto Bressan read the manuscript and pointed out a number of improvements.

REFERENCES

[BiB] S. BIANCHINI AND A. BRESSAN, *Vanishing viscosity solutions of nonlinear hyperbolic systems*, Ann. of Math., to appear.

[B] A. BRESSAN, *Hyperbolic Systems of Conservation Laws. The One-Dimensional Cauchy Problem*, Oxford University Press, Oxford, UK, 2000.

[BC] A. BRESSAN AND R. M. COLOMBO, *Unique solutions of* 2×2 *conservation laws with large data*, Indiana Univ. Math. J., 44 (1995), pp. 677–725.

[BLY] A. BRESSAN, T. P. LIU, AND T. YANG, *L^1 Stability estimates for $n \times n$ conservation laws*, Arch. Ration. Mech. Anal., 149 (1999), pp. 1–22.

[BM] A. Bressan and A. Marson, *A variational calculus for discontinuous solutions of systems of conservation laws*, Comm. Partial Differential Equations, 20 (1995), pp. 1491–1552.

[D] C. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Springer-Verlag, Berlin, 2000.

[HR] H. Holden and N. H. Risebro, *Front Tracking for Hyperbolic Conservation Laws*, Springer-Verlag, New York, 2002.

[J] H. K. Jenssen, *Blowup for systems of conservation laws*, SIAM J. Math. Anal., 31 (2000), pp. 894–908.

[L] P. Lax, *Hyperbolic systems of conservation laws*. II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.

[Le1] M. Lewicka, *L^1 stability of patterns of non-interacting large shock waves*, Indiana Univ. Math. J., 49 (2000), pp. 1515–1537.

[Le2] M. Lewicka, *Stability conditions for patterns of noninteracting large shock waves*, SIAM J. Math. Anal., 32 (2001), pp. 1094–1116.

[Le3] M. Lewicka, *Lyapunov functional for solutions of systems of conservation laws containing a strong rarefaction*, submitted.

[M] A. Majda, *The stability of multidimensional shock fronts*, Mem. Amer. Math. Soc., 41 (1983), iv + 95 pp.

[Mu] P. Mucha, *private communication*, 2003.

[R1] R. Redheffer and D. Port, *Differential equations: Theory and applications*, Jones and Bartlett Publishers, Boston, 1991.

[R2] R. Redheffer, *private communication*, 2002.

[S] D. Serre, *Systems of Conservation Laws*, Cambridge University Press, Cambridge, UK, 1999.

[Scho] S. Schochet, *Sufficient conditions for local existence via Glimm's scheme for large BV data*, J. Differential Equations, 89 (1991), pp. 317–354.

[Sm] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1994.

[TT] Z. Tang and T. C. T. Ting, *Wave curves for the Riemann problem of plane waves in isotropic elastic solids*, Internat. J. Engrg. Sci., 25 (1987), pp. 1343–1381.

SIAM J. MATH. ANAL.
Vol. 36, No. 5, pp. 1371–1399

LYAPUNOV FUNCTIONAL FOR SOLUTIONS OF SYSTEMS OF CONSERVATION LAWS CONTAINING A STRONG RAREFACTION*

MARTA LEWICKA†

Abstract. We study the Cauchy problem for a strictly hyperbolic $n \times n$ system of conservation laws in one space dimension:

$$u_t + f(u)_x = 0,$$

$$u(0,x) = \bar{u}(x).$$

The initial data $\bar{u}$ is a small BV perturbation of a single rarefaction wave with an arbitrary strength. All characteristic fields are assumed to be genuinely nonlinear or linearly degenerate in the vicinity of the reference rarefaction curve. We prove that a suitable BV stability condition yields uniform bounds on the total variation of perturbation, thus implying the existence of a global admissible solution. On the other hand, a stronger L^1 stability condition guarantees the existence of the Lipschitz continuous flow of solutions. Our proof relies on the construction of a Lyapunov functional which is almost decreasing in time and which is equivalent to the L^1 distance between the two solutions.

Key words. conservation laws, large data, rarefaction wave, stability conditions

AMS subject classifications. 35L65, 35L45

DOI. 10.1137/S0036141003429505

1. Introduction and statement of the main results.

The system of conservation laws in one space dimension is the following first order system of nonlinear PDEs:

$$u_t + f(u)_x = 0. \tag{1.1}$$

The well-posedness of (1.1) has been the objective of vast research in recent years; however, at a considerable level of generality it remains an open problem. A complete analysis of the issue has been carried out for strictly hyperbolic flux in (1.1) and initial data $\bar{u} \in BV$ having suitably small total variation:

$$u(0,x) = \bar{u}(x). \tag{1.2}$$

Namely, the entropy solutions to (1.1), (1.2) constitute a flow which is Lipschitz continuous with respect to time and initial data. As shown recently in [BiB], its trajectories are the limits of the solutions to the parabolic regularizations of (1.1), when the viscosity parameter vanishes to zero.

Another approach was implemented in a series of papers [BC, BCP, BLY]. It relies on building piecewise constant approximations of solutions to (1.1), (1.2) and then controlling the evolution of their BV or L^1 norm. The fundamental block in this construction is provided by solutions of the Riemann problems, that is, for initial data $\bar{u}$ consisting of a single discontinuity:

$$u(0,x) = \begin{cases} u^- & x < 0, \\ u^+ & x > 0. \end{cases} \tag{1.3}$$

*Received by the editors June 10, 2003; accepted for publication (in revised form) April 23, 2004; published electronically March 25, 2005. This research was supported by NSF grant DMS-0306201.
http://www.siam.org/journals/sima/36-5/42950.html

†Department of Mathematics, University of Chicago, Eckhart Hall, 1118 E. 58th Street, Chicago, IL 60637 (lewicka@math.uchicago.edu).

To analyze how much the condition of the smallness of initial data can be relaxed, one wishes to study the well-posedness of (1.1), (1.2) with $\bar{u}$ being a small perturbation of fixed Riemann data of arbitrarily large strength. We assume that the solution of the latter is given and that it consists of a number of waves of different characteristic families. More generally, we wish to study the stability of a reference pattern containing possibly strong but noninteracting waves. The above mentioned results indicate that the trivial pattern with no waves present is stable, as one can control the amount (measured in TV or in the L^1 norm) of initially small perturbation of this pattern.

An example in [BC] points out that this is no longer true in the presence of strong waves. Indeed, one has to account for the waves' mutual influence as well as for their interaction with the perturbation, and therefore extra stability conditions are necessary. These conditions in essence refer to the existence of weights with respect to which the flow generated by the associated linearized problem is a contraction; the linearization is taken at states attained by the reference solution [BM]. This approach was realized in a series of papers [BC, Scho, BM, LeT, Le1]. All these works, however, concentrate mainly on patterns with strong shocks or deal solely with the BV stability in the presence of rarefactions.

In [BC] the authors study systems of two equations and prove their BV and L^1 stability under the corresponding nonresonance conditions relating to two shocks. The presence of strong rarefaction waves is also admitted; however, their stability follows without any additional restrictions [Le3], since they belong to the extreme characteristic fields. More general $n \times n$ systems of conservation laws are studied in [Scho], and the BV stability of patterns, including strong shocks, rarefactions, and contact discontinuities, is established. In particular this yields the local-in-time existence of solutions to (1.1), (1.2) within the class of initial data with bounded variation. In [Le1] we established both the BV and the L^1 stability of patterns of noninteracting strong classical shocks in $n \times n$ systems. The crucial ingredient for proving the L^1 stability was the Lyapunov functional approach from [BLY]; let us anticipate that the same method will be used in the present article. The role of the stability conditions from [BM, Le1] and their relations to [BC, Scho] were explained in [Le2].

As a next step, this paper studies BV and L^1 stability of solutions to (1.1), (1.2) close to a reference pattern which is a single rarefaction wave of arbitrary strength. The results of this work combined with [Le1] thus yield the well-posedness analysis for patterns of noninteracting shock and rarefaction waves (compare also [Le3]). The stability conditions presented in this paper are studied in a complementary work [Le3].

We now state our basic hypotheses and set the notation:

(H1) The system (1.1) is strictly hyperbolic in a domain $\Omega \subset \mathbf{R}^n$ to be specified later. That is, for each $u \in \Omega$ the Jacobian matrix $\mathrm{D}f(u)$ of the smooth flux $f : \Omega \longrightarrow \mathbf{R}^n$ has n distinct and real eigenvalues: $\lambda_1(u) < \cdots < \lambda_n(u)$.

Let $\{r_i(u)\}_{i=1}^n$ be the basis of right eigenvectors of $\mathrm{D}f$; $\mathrm{D}f(u)r_i(u) = \lambda_i(u)r_i(u)$. Call $\{l_i(u)\}_{i=1}^n$ the dual basis of left eigenvectors so that $\langle r_i(u), l_j(u)\rangle = \delta_{ij}$ for all $i, j : 1 \dots n$ and all $u \in \Omega$.

Fix $k : 1 \dots n$ and consider an integral curve $\mathcal{R}_k$ of the vector field r_k joining states $u_l, u_r \in \Omega$:

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}\theta}\mathcal{R}_k(\theta) = r_k(\mathcal{R}_k(\theta)), \\ &u_l = \mathcal{R}_k(0), \quad u_r = \mathcal{R}_k(\Theta), \qquad \Theta > 0. \end{aligned} \tag{1.4}$$

$\mathcal{R}_k$ is called the rarefaction curve. For a small $c > 0$ we define the domain

$$\Omega = \Omega_c = \{u \in \mathbf{R}^n : \quad ||u - \mathcal{R}_k(\theta)|| < c \quad \text{for some } \theta \in [0, \Theta]\}; \tag{1.5}$$

all the subsequent reasoning will be restricted to this domain, with the parameter c appropriately small. We further assume that

(H2) In Ω, each characteristic field $i : 1 \dots n$ is either linearly degenerate ($\langle \mathrm{D}\lambda_i, r_i \rangle \equiv 0$), or it is genuinely nonlinear, which means that $\langle \mathrm{D}\lambda_i, r_i \rangle > 0$. The kth characteristic field is assumed to be genuinely nonlinear.

In the case of linearly degenerate fields we set $||r_i(u)|| = 1$, while when the ith field is genuinely nonlinear we choose the normalization of right eigenvectors $r_i(u)$ so that $\langle \mathrm{D}\lambda_i(u), r_i(u) \rangle = 1$ for all $u \in \Omega$. In particular we have

$$\langle \mathrm{D}\lambda_k(u), r_k(u) \rangle = 1 \quad \text{for all } u \in \Omega \tag{1.6}$$

and thus $\Theta = \lambda_k(u_r) - \lambda_k(u_l)$.

The piecewise smooth, self-similar function, called the centered rarefaction wave (see Figure 1.1), is given by

$$u_0(t,x) = \begin{cases} u_l & \text{if } x < t\lambda_k(u_l), \\ \mathcal{R}_k(\theta) & \text{if } x = t\lambda_k(\mathcal{R}_k(\theta)), \quad \theta \in [0, \Theta], \\ u_r & \text{if } x > t\lambda_k(u_r) \end{cases} \tag{1.7}$$

and provides an entropy admissible solution of (1.1) [Sm, D]. The objective of this paper is a study of the stability of u_0. Our main results are expressed in the following theorems.

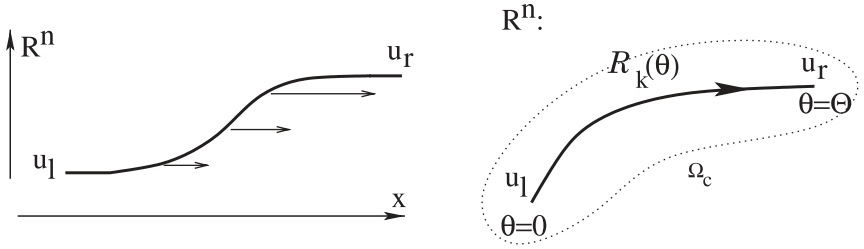

FIG. 1.1.

THEOREM 1. *Assume that* (H1), (H2), *and the BV stability condition* (2.6) *hold. For $c, \delta > 0$ let $\mathcal{E}_{c,\delta}$ denote the set of all continuous functions $\bar{u}$ satisfying*

(i) $\bar{u}(x) \in \Omega_c$ *for all* $x \in \mathbf{R}$,

(ii) $\lim_{x \to -\infty} \bar{u}(x) = u_l$ *and* $\lim_{x \to \infty} \bar{u}(x) = u_r$,

(iii) $|TV(\bar{u}) - |\mathcal{R}_k|| < \delta$, *where* $|\mathcal{R}_k| = TV(\mathcal{R}_k)$ *is the arc-length of the rarefaction curve* $\mathcal{R}_k(\theta)$, $\theta \in [0, \Theta]$.

There exist small parameters $c, \delta > 0$ such that for every $\bar{u} \in \mathrm{cl}\ \mathcal{E}_{c,\delta}$, where cl *denotes the closure in L^1_{loc}, the Cauchy problem* (1.1), (1.2) *has a global entropy admissible solution $u(t,x)$.*

THEOREM 2. *Assume that* (H1), (H2), *and the L^1 stability condition* (3.1) *are satisfied. Then there exists a closed domain $\mathcal{D} \subset L^1_{loc}(\mathbf{R}, \Omega)$, containing all continuous functions $\bar{u}$ satisfying* (i), (ii), (iii) *in Theorem* 1, *for some $c, \delta > 0$, and there exists a semigroup $S : \mathcal{D} \times [0, \infty) \longrightarrow \mathcal{D}$ such that*

(i) $||S(\bar{u},t)-S(\bar{v},s)||_{L^1} \leq L\cdot(|t-s|+||\bar{u}-\bar{v}||_{L^1})$ *for all* $\bar{u},\bar{v}\in\mathcal{D}$, *all* $t,s\geq 0$, *and a uniform constant* L, *depending only on the system* (1.1),

(ii) *for all* $\bar{u}\in\mathcal{D}$, *the trajectory* $t\mapsto S(\bar{u},t)$ *is the solution to* (1.1), (1.2) *given in Theorem* 1.

We now set other preliminaries. For each $i:1\dots n$ and $u\in\Omega$, call $\sigma\mapsto\mathcal{S}_i(u,\sigma)$ and $\sigma\mapsto\mathcal{R}_i(u,\sigma)$, the ith shock and the ith rarefaction curves through the point u [L, D]. In particular we have $\mathcal{R}_k(u_l,\theta)=\mathcal{R}_k(\theta)$. Both curves are defined at least locally, that is, for $\sigma\in(-c,c)$, and have second order contact at $\sigma=0$:

$$\mathcal{S}_i(u,\sigma)-\mathcal{R}_i(u,\sigma)=\mathcal{O}(1)|\sigma|^3. \tag{1.8}$$

The curves' parametrization is consistent with the normalization of the right eigenvectors r_i. That is, they are parametrized by arc-length if the ith characteristic field is linearly degenerate, and by the corresponding eigenvalue λ_i if the ith field is genuinely nonlinear:

$$\lambda_i(\mathcal{S}_i(u,\sigma))-\lambda_i(u)=\sigma=\lambda_i(\mathcal{R}_i(u,\sigma))-\lambda_i(u). \tag{1.9}$$

By this choice of parametrization we have

$$\mathcal{S}_i(\mathcal{S}_i(u,\sigma),-\sigma)=u. \tag{1.10}$$

The speed λ of a weak shock wave $(u^-,u^+=\mathcal{S}_i(u^-,\sigma))$ with strength $\sigma<0$ can be computed from the Rankine–Hugoniot identity:

$$f(u^+)-f(u^-)=\lambda\cdot(u^+-u^-). \tag{1.11}$$

Throughout the paper, by $\mathcal{O}(1)$ we mean any uniformly bounded function, depending only on the system (1.1). Any sufficiently small but positive constant is denoted by c. The Riemann data as in (1.3) is for simplicity denoted by (u^-,u^+).

The paper is constructed as follows. In sections 2 and 3 we present the stability conditions and their primary motivation. In section 4 we prove Theorem 1. The proof relies on the construction of approximate solutions by means of the wave front tracking algorithm [HR, BaJ], and applying the Glimm analysis in view of the BV stability condition. In section 9 we prove that the domain of applicability of these techniques actually contains the data with properties as in Theorem 1.

Toward the proof of Theorem 2, in section 6 we give the definition of the Lyapunov functional measuring the L^1 distance between the two approximate solutions constructed in section 4. The crucial observation for our construction is noting that in the initial time interval where the solutions are apart from each other, this distance decreases rapidly. A convenient tool to estimate the decrease is the first order rarefactions, introduced in section 5. For other times, the pointwise distance between solutions is calculated along shock curves, as in [BLY]. The decrease of the functional follows then from the assumed L^1 stability condition and the main concern of sections 7 and 8.

2. The weighted BV stability condition. In this section we discuss a stability condition guaranteeing the existence of solutions to the problem (1.1)(1.2) in the vicinity of the reference rarefaction wave (1.7). To motivate our approach we first recall the argument from [Le1, BM]. The stability conditions there were formulated in terms of the existence of a family of weights $w_i>0$, $i:1\dots n$, corresponding to different characteristic families of perturbation v, and depending on the location of

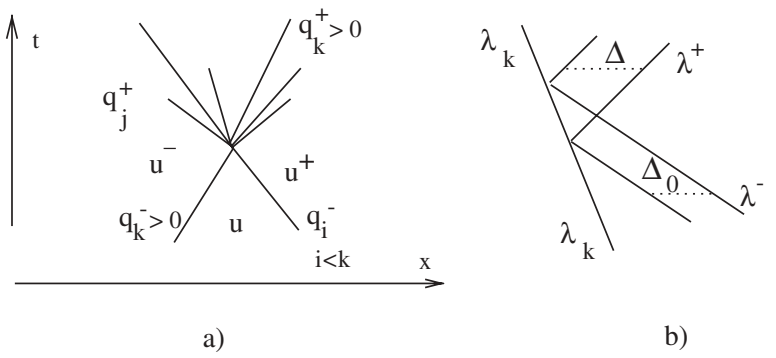

FIG. 2.1.

perturbing waves inside the reference pattern u_0. The conditions required that the weighted BV or L^1 norm of any solution of

$$v_t + Df(u_0)v_x + [D^2 f(u_0) \cdot v] \cdot (u_0)_x = 0$$

was nonincreasing in time.

Let $w_1 \dots w_{k-1}, w_{k+1} \dots w_n : (-c, \Theta + c) \longrightarrow \mathbf{R}_+$ be smooth, nonnegative functions defined along the rarefaction curve $\mathcal{R}_k$ in (1.4). We can extend this definition on the whole neighborhood Ω by setting

$$\forall i \neq k,\ \forall u \in \Omega \quad w_i(u) = w_i(\theta), \quad \text{where} \quad \lambda_k(u) = \lambda_k(\mathcal{R}_k(\theta)). \tag{2.1}$$

Consider an interaction of a weak ith wave with a small part of the rarefaction $\mathcal{R}_k$, located at the state $u = \mathcal{R}_k(\theta)$. To fix the ideas, assume that $i < k$ and call the strengths of the incoming waves and the states they join to u, respectively, $q_k^- > 0,\ q_i^-, u^-, u^+$ (as in Figure 2.1(a)). In particular, we have $u = \mathcal{R}_k(u^-, q_k^-)$ and $q_k^- = \theta - \lambda_k(u^-)$. The strengths of waves are computed in terms of change in the corresponding eigenvalue for genuinely nonlinear fields, or the arc-length of the rarefaction curve connecting the two states, for linearly degenerate fields. We thus remain consistent with the parametrization of the right eigenvectors, given in section 1. Now if q_k^- and q_i^- are small enough, the Riemann problem (u^-, u^+) has a self-similar solution composed of n outgoing waves having strengths $q_1^+ \dots q_n^+$. For the basic properties of this construction we refer to [L, Sm, B, D]. Assigning to each wave the weight w_i corresponding to its characteristic family and computed at the wave's left state, we now require that the weighted amount of perturbation decreases across the interaction, so that

$$\sum_{j \neq k} w_j^+ |q_j^+| < w_i^- |q_i^-|. \tag{2.2}$$

Recall the standard Taylor estimates [Sm]:

$$\begin{aligned} \forall j \neq k \qquad q_j^+ = \delta_{ij} \cdot q_i^- + \langle l_j(u), [r_i, r_k](u) \rangle \cdot q_i^- q_k^- \\ + \mathcal{O}(1) |q_i^- q_k^-| (|q_i^-| + |q_k^-|). \end{aligned} \tag{2.3}$$

Here $[r_i, r_k] = \mathrm{D}r_i \cdot r_k - \mathrm{D}r_k \cdot r_i$ stands for the Lie bracket of two vector fields, and δ_{ij} is the Kronecker delta.

In view of (2.3), we have

$$\begin{aligned}\forall j\neq k,i \qquad w_j^+|q_j^+| =& w_j(u)\cdot|\langle l_j(u),[r_i,r_k](u)\rangle|\cdot|q_i^-q_k^-|\\ &+\mathcal{O}(1)|q_i^-q_k^-|(|q_i^-|+|q_k^-|).\end{aligned}$$

On the other hand,

$$\begin{aligned}w_i^+q_i^+-w_i^-q_i^- =&(w_i^+-w_i^-)q_i^-+w_i^+(q_i^+-q_i^-)\\ =&-w_i'(\theta)\cdot q_i^-q_k^-+w_i(u)\cdot\langle l_i,[r_i,r_k]\rangle(u)\cdot q_i^-q_k^-\\ &+\mathcal{O}(1)|q_i^-q_k^-|(|q_i^-|+|q_k^-|).\end{aligned}$$

Hence,

$$\begin{aligned}w_i^+|q_i^+|-w_i^-|q_i^-| =&(\operatorname{sgn} q_i^-)\cdot(w_i^+q_i^+-w_i^-q_i^-)\\ =&\big\{w_i(u)\cdot\langle l_i,[r_i,r_k]\rangle(u)-w_i'(\theta)\big\}\cdot|q_i^-q_k^-|\\ &+\mathcal{O}(1)|q_i^-q_k^-|(|q_i^-|+|q_k^-|).\end{aligned}$$

Condition (2.2) is thus equivalent to

$$\left(\sum_{j\neq i,k}w_j(\theta)\cdot|\langle l_j,[r_i,r_k]\rangle(\mathcal{R}_k(\theta))|\right)+w_i(\theta)\cdot\langle l_i,[r_i,r_k]\rangle(\mathcal{R}_k(\theta))<w_i'(\theta). \tag{2.4}$$

Analogously, for $i>k$ one obtains

$$\left(\sum_{j\neq i,k}w_j(\theta)\cdot|\langle l_j,[r_k,r_i]\rangle(\mathcal{R}_k(\theta))|\right)+w_i(\theta)\cdot\langle l_i,[r_k,r_i]\rangle(\mathcal{R}_k(\theta))<-w_i'(\theta). \tag{2.5}$$

Define the $(n-1)\times(n-1)$ matrix function:

$$\begin{aligned}&\mathbf{P}(\theta)=[p_{ij}(\theta)]_{\substack{i,j:1\dots n,\\ i,j\neq k}} \quad \text{for } \theta\in[0,\Theta],\\ &p_{ij}(\theta)=\begin{cases}|\langle l_j,[r_i,r_k]\rangle(\mathcal{R}_k(\theta))| & \text{if } i\neq j,\\ \operatorname{sgn}(k-i)\cdot\langle l_i,[r_i,r_k]\rangle(\mathcal{R}_k(\theta)) & \text{if } i=j.\end{cases}\end{aligned}$$

Combining (2.4) and (2.5), we have proved the following.

LEMMA 2.1. *Condition* (2.2) *is equivalent to the following:*

$$\left[\begin{array}{l}\textit{BV stability condition: There exist positive smooth functions } w_1\dots w_{k-1},w_{k+1}\dots w_n:[0,\Theta]\to\mathbf{R}_+ \textit{ such that}\\ \\ \qquad\mathbf{P}(\theta)\cdot\begin{bmatrix}w_1(\theta)\\ \vdots\\ w_{k-1}(\theta)\\ w_{k+1}(\theta)\\ \vdots\\ w_n(\theta)\end{bmatrix}<\begin{bmatrix}w_1'(\theta)\\ \vdots\\ w_{k-1}'(\theta)\\ -w_{k+1}'(\theta)\\ \vdots\\ -w_n'(\theta)\end{bmatrix} \qquad \textit{for every } \theta\in(0,\Theta),\\ \\ \textit{where the above vector inequality is understood componentwise.}\end{array}\right. \tag{2.6}$$

Remark 2.2. Notice that because of the strict inequalities in (2.4) and (2.5), the condition (2.6) implies a stricter version of (2.2):

$$\sum_{j\neq k} w_j^+|q_j^+| < w_i^-|q_i^-| - c|q_i^- q_k^-|$$

for a small constant c.

Remark 2.3. The inequality in (2.6) is independent from rescaling $w_i \mapsto \alpha \cdot w_i$, for any $\alpha > 0$. Thus, in particular we may assume that

$$|w_i(u)| < 1 \quad \text{and} \quad ||\mathrm{D}w_i(u)|| < 1$$

for each i and every $u \in \Omega$.

Remark 2.4. If all $p_{ij}(\theta) \geq 0$, we can regard the quantity $w_i(\theta)$ as the measure of the amount of potential future interactions of the ith perturbation wave located at the state $\mathcal{R}_k(\theta)$. For $i < k$ each w_i is an increasing function of θ, and for $i > k$ each w_i is decreasing along the curve $\mathcal{R}_k$. Indeed, the slow waves ($\lambda_i < \lambda_k$ for $i < k$) travel in the direction of decreasing θ on the $t - x$ plane, and thus the bigger the parameter θ corresponding to their location is, the more potential contribution to the future amount of perturbation they create. The converse assertion is true for the fast waves of characteristic families $i > k$.

By an approximation argument, as the inequality in (2.6) is strict, we see that (2.2) also holds for any state $u \in \Omega_c$. For the more detailed discussion of condition (2.6) we refer to the paper [Le3]. In particular, we have the following.

LEMMA 2.5 [Le3]. *Let the condition* (2.6) *be satisfied. There exists* $c > 0$ *such that for every* $u^-, u^+ \in \Omega$ *with* $\lambda_k(u^+) - \lambda_k(u^-) > -c$, *the Riemann problem* (u^-, u^+) *for* (1.1) *has the unique self-similar solution attaining states in* Ω. *The solution is composed of* $n-1$ *weak waves of families* $1 \ldots k-1, k+1 \ldots n$ *and a* k*th wave which is either a weak shock or a rarefaction.*

Condition (2.6) is independent of the parametrization of the eigenvectors in Ω. The next lemma gathers several other properties of this condition.

LEMMA 2.6 [Le3]. *In any of the following cases* (2.6) *is satisfied:*

(i) *when the reference rarefaction is sufficiently weak, that is,* $0 < \Theta \ll 1$,

(ii) *when the reference rarefaction belongs to an extreme characteristic field (*$k = 1$ *or* n*),*

(iii) *when* (1.1) *has a coordinate system of Riemann invariants* [Sm, D, S].

In particular, any rarefaction wave in any 2×2 *system or the* 3×3 *system of Euler equations of gas dynamics* [D, Sm, Scho] *is BV stable.*

(iv) *For* $n = 3$ *and* $k = 2$, (2.6) *is equivalent to the existence of a positive solution* $v : [0, \Theta] \longrightarrow \mathbf{R}_+$ *to the Riccati equation:*

$$v'(\theta) = p_{12}(\theta) + [p_{11}(\theta) + p_{22}(\theta)] \cdot v(\theta) + p_{21}(\theta) \cdot v^2(\theta).$$

3. The weighted L^1 stability condition. The production matrix $\mathbf{P}$ in condition (2.6) accounts for the infinitesimal change of the strength of perturbation as it passes through the rarefaction fan (1.7). The elements of $\mathbf{P}(\theta)$ are second order coefficients in the Taylor expansion of the strength of waves produced through the interaction with a part of the large rarefaction $\mathcal{R}_k(\theta)$. In order to deal with the L^1 stability one is led to a "mass production" matrix $\mathbf{M}(\theta)$ whose components additionally account for the shifts in locations of the perturbing waves of different characteristic

families before and after the interaction. More precisely, define

$$\mathbf{M}(\theta) = [m_{ij}(\theta)]_{\substack{i,j:1\dots n,\\ i,j\neq k}} \quad \text{for } \theta \in [0,\Theta],$$

$$m_{ij}(\theta) = \begin{cases} p_{ij}(\theta)\cdot \dfrac{|\lambda_j-\lambda_k|}{|\lambda_i-\lambda_k|}(\mathcal{R}_k(\theta)) & \text{if } i\neq j,\\ p_{ij}(\theta) + \dfrac{\mathrm{D}\lambda_i\cdot r_k}{|\lambda_i-\lambda_k|}(\mathcal{R}_k(\theta)) & \text{if } i=j. \end{cases}$$

We have the following:

(3.1) L^1 *stability condition:* There exist positive smooth functions $w_1\dots w_{k-1}, w_{k+1}\dots w_n : [0,\Theta]\to \mathbf{R}_+$ such that the inequality in (2.6) is satisfied with $\mathbf{M}(\theta)$ replacing the matrix $\mathbf{P}(\theta)$.

Note that an observation as in Remark 2.3 remains valid.

A more restrictive version of (3.1), where all weights w_i are linear, was introduced in [BM] in the context of the well-posedness of the associated variational system.

LEMMA 3.1 [Le3]. *We have the following:*

(i) *Condition* (3.1) *is stronger than the BV stability condition* (2.6).

(ii) *The assertions of Lemma* 2.6 *hold in their respective versions.*

(iii) *For all* $i\neq j$ *and all* $\theta\in[0,\Theta]$ *there holds:* $m_{ij}(\theta) = |\langle l_j, \mathrm{D}r_i\cdot r_k\rangle(\mathcal{R}_k(\theta))|$ *and* $m_{ii}(\theta) = \operatorname{sgn}(k-i)\cdot\langle l_i, \mathrm{D}r_i\cdot r_k\rangle(\mathcal{R}_k(\theta))$.

We end this section by presenting a consequence of (3.1) which plays the same role as Lemma 2.1 and Remark 2.2 for the condition (2.6). Its proof will follow from the more general Lemma 8.2. To fix the ideas, let

$$\mathcal{S}_k(q_k^-)\circ\mathcal{S}_i(u,q_i^-) = \mathcal{S}_n(q_n^+)\circ\dots\mathcal{S}_1(u,q_1^+)$$

with $u\in\Omega$, $\{q_j^-\}_{j=i,k}$ small enough and $q_k^-\geq 0$. Then for a small uniform constant γ we have

$$\sum_{j\neq k} w_j^+|q_j^+|\cdot|\lambda_j^+-\lambda_k^+| < w_i^-|q_i^-|\cdot|\lambda_i^--\lambda_k^-| - \gamma|q_i^-q_k^-|.$$

Namely, the total weighted mass of perturbation decreases as it passes through the rarefaction wave (1.7). Recall [BM] that the ratio Δ/Δ_0 of shifts in the reflected or transmitted wave with respect to the shift in an incoming wave can be computed as $|\lambda^+-\lambda_k|/|\lambda^--\lambda_k|$. As in Figure 2.1(b), λ^- and λ^+ denote speeds of the modified waves before and after the interaction with a reference wave traveling with speed λ_k.

4. Existence of solutions: A proof of Theorem 1. Recall that given a Cauchy problem (1.1), (1.2) with $\bar u$ having small total variation, its solution can be obtained in the limit when $\epsilon\to 0$ of piecewise constant ϵ-approximations $u^\epsilon(t,x)$ constructed via the wave front tracking algorithm [BaJ, HR]. For the detailed description of the algorithm we refer to [B]. The crucial ingredient in proving the global existence of the approximate solutions and the compactness of its sequence is the Glimm functional [G] controlling the total variation of perturbation and the amount of the future interactions. Below we briefly discuss a natural modification of this standard construction, applicable when the reference pattern is a strong kth rarefaction $\mathcal{R}_k$ rather than a constant state. We then show that our Glimm-type functional Γ is indeed nonincreasing along any wave front tracking approximate solution, thanks to the BV stability condition (2.6).

DEFINITION 4.1. *Let* $\epsilon_0>0$. *By* $\mathcal{D}_{\epsilon_0}$ *we denote the set of piecewise constant functions* $v:\mathbf{R}\longrightarrow\mathbf{R}^n$ *enjoying the following properties:*

(i) $v(-\infty) = u_l,\ v(+\infty) = u_r$,
(ii) $v(x) \in \Omega$ *for all* $x \in \mathbf{R}$,
(iii) *all jumps in* v *have amplitudes smaller than* ϵ_0 *(and thus the corresponding Riemann problems admit the standard self-similar solution). We order the waves in these solutions according to their location and speed; for a wave* α, *we denote its characteristic family by* $i_\alpha : 1 \dots n$ *and its strength by* ϵ_α,
(iv) *setting* $\epsilon_\alpha^+ = \max(0, \epsilon_\alpha)$ *and* $\epsilon_\alpha^- = \max(0, -\epsilon_\alpha)$ *there holds*

$$\left| \left(\sum_{i_\alpha = k} \epsilon_\alpha^+ \right) - \Theta \right| + \left(\sum_{i_\alpha = k} \epsilon_\alpha^- \right) + \left(\sum_{i_\alpha \neq k} |\epsilon_\alpha| \right) \leq \epsilon_0. \tag{4.1}$$

Remark 4.2. *Let* v *satisfy* (i), (iii) *of Definition* 4.1 *and let the bound* (4.1) *hold with* ϵ_0 *exchanged by another parameter* δ. *Then if only* δ *is small enough with respect to* ϵ_0, *then* $v(x) \in \Omega_{2c}$ *for all* $x \in \mathbf{R}$ *implies* $v(x) \in \Omega_c$ *for all* $x \in \mathbf{R}$.

Take a function $u(0, \cdot) \in \mathcal{D}_{\epsilon_0}$ for some small $\epsilon_0 > 0$. Let $\epsilon \ll \epsilon_0$. Recall that the fundamental block for constructing the approximate solution $u^\epsilon(t, x)$ is provided by piecewise constant approximations of self-similar solutions to Riemann problems.

As customary, the nonphysical waves generated by the simplified Riemann solver are said to belong to the $(n+1)$th characteristic family. The simplified Riemann solver is used whenever one of the interacting waves is nonphysical or when the product of strengths of incoming waves is bigger than a threshold parameter $\rho(\epsilon)$. The details can be found in Chapter 7 of [B]. The associated nonphysical weight w_{n+1} is defined as follows:

$$w_{n+1}(u) = c \cdot \exp(-C \cdot \lambda_k(u)) \quad \text{for} \quad u \in \Omega, \tag{4.2}$$

for some suitable constants $c, C > 0$. Let w_k be a positive constant, strictly smaller than all other weights $w_i(u)$ defined in Ω by (2.6) and (2.1). Recall that given a weak ith wave, we associate with it the weight w_i computed at its left state.

DEFINITION 4.3. *Let* $u(0, \cdot) \in \mathcal{D}_{\epsilon_0}$, *with some small* $\epsilon_0 > 0$. *Let* u^ϵ *be the piecewise constant* ϵ*-approximate solution, given by the wave front tracking algorithm. Assume* t *is not an interaction time of fronts in* u^ϵ. *Using the notation of Definition* 4.1 *we set*

$$V(u^\epsilon(t, \cdot)) = \left| \left(\sum_{i_\alpha = k} \epsilon_\alpha^+ \right) - \Theta \right| + \left(\sum_{i_\alpha = k} \epsilon_\alpha^- \right) + \left(\sum_{i_\alpha \neq k} |\epsilon_\alpha| \right),$$

where the summations extend on all waves α *present in* $u^\epsilon(t, \cdot)$. *The quadratic interaction potential is defined:*

$$Q_0(u^\epsilon(t, \cdot)) = \sum_{(\alpha, \beta) \in \mathcal{A}} |\epsilon_\alpha \cdot \epsilon_\beta|,$$

with the set $\mathcal{A}$ *containing all couples of perturbation waves* (α, β) *in* $u^\epsilon(t, \cdot)$ *approaching each other. More precisely, assuming* $x_\alpha < x_\beta$, *we have* $(\alpha, \beta) \in \mathcal{A}$ *if and only if* $i_\alpha > i_\beta$ *or else* $i_\alpha = i_\beta$ *and at least one of the waves is a genuinely nonlinear shock. In both cases we require that none of the waves* α, β *is a positive* k*-wave. Finally, let*

$$Q_{large}(u^\epsilon(t, \cdot)) = \sum_{i_\alpha \neq k} w_{i_\alpha}(u^\epsilon(t, x_\alpha -)) \cdot |\epsilon_\alpha| + \sum_{i_\alpha = k} w_k \cdot \epsilon_\alpha^-,$$

$$Q = Q_0 + Q_{large}, \qquad \Gamma = V + \kappa \cdot Q,$$

for some large constant κ, to be determined later.

LEMMA 4.4. *Assume that the BV stability condition* (2.6) *holds. Then for some constants $c, \epsilon_0, \kappa > 0$ we have the following. Let $u(0,\cdot) \in \mathcal{D}_{\epsilon_0}$ and let u^ϵ be the corresponding piecewise constant approximate solution obtained through the wave front tracking algorithm. Then for any $t > 0$ when two wave fronts α and β interact, if $\Gamma(u^\epsilon(t-,\cdot)) \leq \epsilon_0$, then*

$$\begin{aligned} \Delta Q &= Q(u^\epsilon(t+,\cdot)) - Q(u^\epsilon(t-,\cdot)) \leq -c \cdot |\epsilon_\alpha \epsilon_\beta|, \\ \Delta\Gamma &= \Gamma(u^\epsilon(t+,\cdot)) - \Gamma(u^\epsilon(t-,\cdot)) \leq -c \cdot |\epsilon_\alpha \epsilon_\beta|. \end{aligned} \tag{4.3}$$

Proof. The proof consists of several cases, depending on whether the accurate or the simplified Riemann solver is used and whether the interaction involves a kth positive wave which we will view as a part of the reference rarefaction $\mathcal{R}_k$. We give only the main ideas; the detailed analysis is left to the reader.

Case 1. None of the interacting waves is a positive kth wave, and the interaction is solved by the accurate Riemann solver (Figure 4.1 (c)). By standard analysis [B] we have

$$\begin{aligned} \Delta V &= \mathcal{O}(1)|\epsilon_\alpha \epsilon_\beta|, \\ \Delta Q_0 &\leq -|\epsilon_\alpha \epsilon_\beta| + \mathcal{O}(1)\epsilon_0 \cdot |\epsilon_\alpha \epsilon_\beta|. \end{aligned}$$

Further,

$$\Delta Q_{large} \leq \left(\sum_{j \neq i_\alpha, i_\beta} w_j^{out} \cdot |\epsilon_j^{out}| \right) + \left(w_{i_\alpha}^{out} |\epsilon_\alpha^{out}| - w_{i_\alpha} |\epsilon_\alpha| \right) + \left(w_{i_\beta}^{out} |\epsilon_\beta^{out}| - w_{i_\beta} |\epsilon_\beta| \right).$$

Consequently, $\Delta Q_{large} \leq C \cdot |\epsilon_\alpha \epsilon_\beta|$, where the constant C depends linearly on the upper bound of the weights $\{w_i\}$ as well as their derivatives $\{Dw_i\}$. In view of Remark 2.3 and assuming ϵ to be small enough we thus obtain the first estimate in (4.3), which in turns yields the second one for large κ.

Case 2. Interaction of a wave of family $i_\beta \neq k$ with a kth positive wave ($i_\alpha = k,\ \epsilon_\alpha > 0$) solved by the accurate Riemann solver (Figure 4.1 (c)). As before, we obtain

$$\begin{aligned} \Delta V &= \mathcal{O}(1)|\epsilon_\alpha \epsilon_\beta|, \\ \Delta Q_0 &= \mathcal{O}(1)\epsilon_0 \cdot |\epsilon_\alpha \epsilon_\beta|. \end{aligned} \tag{4.4}$$

We view ΔQ_{large} as a function of the state $u \in \Omega$ attained by u^ϵ between the interacting fronts α and β and the strengths ϵ_α and ϵ_β:

$$\Delta Q_{large} = -w_{i_\alpha} \cdot |\epsilon_\alpha| + \sum_{j \neq k} w_j^{out} \cdot |\epsilon_j^{out}| = G(u, \epsilon_\alpha, \epsilon_\beta).$$

Choose $\theta \in [0, \Theta]$ such that $||u - \mathcal{R}_k(\theta)|| < \epsilon_0$. Since $G(u, \epsilon_\alpha, 0) = G(u, 0, \epsilon_\beta) = 0$, we have

$$\begin{aligned} &|G(u, \epsilon_\alpha, \epsilon_\beta) - G(\mathcal{R}_k(\theta), \epsilon_\alpha, \epsilon_\beta)| \\ &\qquad \leq |\epsilon_\alpha \epsilon_\beta| \cdot \int_0^1 \int_0^1 \left| \frac{\partial^2}{\partial \epsilon_\alpha \partial \epsilon_\beta} G(u, s\epsilon_\alpha, z\epsilon_\beta) - \frac{\partial^2}{\partial \epsilon_\alpha \partial \epsilon_\beta} G(\mathcal{R}_k(\theta), s\epsilon_\alpha, z\epsilon_\beta) \right| \mathrm{d}s\mathrm{d}z. \end{aligned}$$

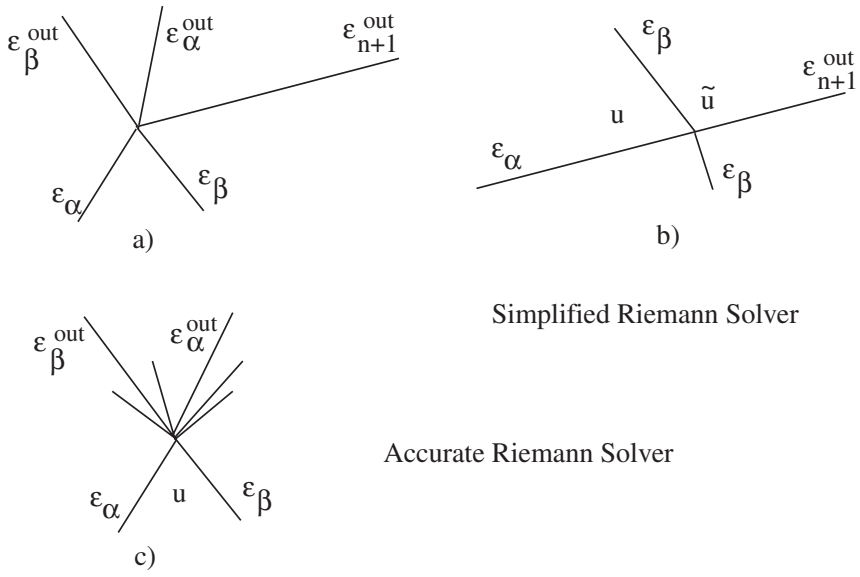

FIG. 4.1.

If only the constant c in the definition (1.5) of Ω is small enough, the integrand in the above estimate is as small as we wish. Thus in view of Remark 2.2 we obtain $\Delta Q_{large} \leq -c \cdot |\epsilon_\alpha \epsilon_\beta|$ for some different constant $c > 0$, taking w_k sufficiently small with respect to other weights. If ϵ is small enough and κ large, this implies (4.3).

We remark that if the interaction as in Case 2 is to be solved by the simplified Riemann solver (Figure 2.1 (a)), then (4.3) follows exactly as above provided we define ϵ_k^{out} to be equal to ϵ_k^{out} in the accurate solution and take the scaling constant c in (4.2) small with respect to other weights $w_i,\ i : 1 \dots n$.

Case 3. Interaction of a nonphysical front ($i_\alpha = n+1$) with a positive k-wave ($i_\beta = k,\ \epsilon_\beta > 0$) solved by the simplified Riemann solver (Figure 4.1 (b)). Again (4.4) is valid. Call u the left state of the wave α and call $\tilde{u}$ the state attained by u^ϵ between the two outgoing waves. Then

$$\begin{aligned}
\Delta Q_{large} &= w_{n+1}(\tilde{u}) \cdot |\epsilon_{n+1}^{out}| - w_{n+1}(u) \cdot |\epsilon_\alpha| \\
&\leq (w_{n+1}(\tilde{u}) - w_{n+1}(u)) \cdot |\epsilon_\alpha| + \mathcal{O}(1) w_{n+1}(\tilde{u}) \cdot |\epsilon_\alpha \epsilon_\beta| \\
&= c \cdot \exp(-C\lambda_k(\tilde{u})) \cdot [1 - \exp(-C\epsilon_\beta) + \mathcal{O}(1)\epsilon_\beta] \cdot |\epsilon_\alpha| \\
&= c \cdot \exp(-C\lambda_k(\tilde{u})) \cdot \left[\mathcal{O}(1) - \frac{\exp(C\epsilon_\beta) - 1}{\epsilon_\beta}\right] \cdot |\epsilon_\alpha \epsilon_\beta| \\
&\leq -c\frac{C}{4} \exp(-C\lambda_k(\tilde{u})) \cdot |\epsilon_\alpha \epsilon_\beta|
\end{aligned}$$

if only C in (4.2) is large enough. Taking ϵ_0 small and κ large, we conclude (4.3). ∎

Define now the domain

$$\bar{\mathcal{D}}_{\epsilon_0} = \text{cl}\,\{v \in \mathcal{D}_{\epsilon_0}, \quad \Gamma(v) \leq \epsilon_0\}, \tag{4.5}$$

where cl denotes the closure in L^1_{loc}. Relying on Lemma 4.4 and Remark 4.2, we obtain the following.

LEMMA 4.5. *In the setting of Lemma* 4.4, *an approximate solution* $u^\epsilon(t,x)$ *generated by the algorithm from initial data* $\bar u \in \bar{\mathcal{D}}_{\epsilon_0}$ *exists for all times* $t>0$ *and enjoys the following properties:*

(i) $||\bar u - u^\epsilon(0,\cdot)||_{L^1} \le \epsilon$,

(ii) u^ϵ *is piecewise constant, with jumps occurring along finitely many lines; jumps are of three types: shocks (and contact discontinuities), rarefaction fronts, and nonphysical waves; all jumps have strength* $<\epsilon_0$, *while all rarefaction fronts have strength* $<\epsilon$,

(iii) *along each shock or a rarefaction front not belonging to the* k*th family we have that its speed differs from the exact speed (Rankine–Hugoniot speed for shocks and the eigenvalue at the left state for rarefaction fronts) at most by* ϵ; *the speeds of all* k*-positive waves are exact (that is, equal to* λ_k *evaluated at the left state); all nonphysical waves travel with speed* $\hat\lambda$,

(iv) *at each time* $t\ge 0$ *the sum of strengths of nonphysical waves in* u^ϵ *is bounded by* ϵ,

(v) *for all* $t\ge 0$ *we have* $\Gamma(u^\epsilon(t,\cdot)) \le \epsilon_0$.

Now a standard argument yields that a subsequence of approximations u^ϵ converges to a solution of (1.1), (1.2) and that the domain $\bar{\mathcal{D}}_{\epsilon_0}$ is positively invariant with respect to the flow generated in this way. Again, all the details can be found in [B]. To prove Theorem 1 it thus suffices to show the following.

LEMMA 4.6. *Let* $\bar u \in \text{cl}\,\mathcal{E}_{c,\delta}$ *for sufficiently small* $c,\delta>0$, *as in Theorem* 1. *Then* $\bar u \in \bar{\mathcal{D}}_{\epsilon_0}$ *for some* $\epsilon_0=\epsilon_0(\delta)$ *and* $\lim_{\delta\to 0}\epsilon_0(\delta)=0$.

The proof will be given in section 9.

5. First order rarefactions. We call a positive kth wave located at y_0 at time $T>0$ a first order k-rarefaction wave if there exists a continuous curve $y(t)$ with $y(T)=y_0$ such that for almost all $t\in[0,T]$, $y(t)$ is the location of a positive kth wave. For each $t\in[0,+\infty)$ let $L^u(t)$ be the set of locations of first order k-rarefaction waves in u.

LEMMA 5.1. *Let* $u^\epsilon(t,x)$ *be as in Lemma* 4.5 *(in particular* $u^\epsilon(t,\cdot)\in\mathcal{D}_{\epsilon_0}$ *for all* $t\ge 0$*). Then*

$$\tilde V(t) := \left|\left(\sum_{x_\alpha\in L^u(t)}\epsilon_\alpha\right)-\Theta\right| + \left(\sum_{x_\alpha\notin L^u(t)}|\epsilon_\alpha|\right) = \mathcal{O}(1)\cdot\epsilon_0, \tag{5.1}$$

with the above summations extending on all waves α *present in* $u^\epsilon(t,\cdot)$. *Moreover, if* $y(t)$ *is continuous and* $y(t)\in L^u(t)$ *for almost all* $t\in[0,T]$, *then*

$$\forall t,s\in[0,T]\quad |\lambda_k(u^\epsilon(t,y(t)-)) - \lambda_k(u^\epsilon(s,y(s)-))| = \mathcal{O}(1)\cdot\epsilon_0. \tag{5.2}$$

Proof. Above $\tilde V(0)$ is understood as $\tilde V(t)$ for t close to 0. To prove (5.1) one defines new interaction potentials by the same formula as Q_0 and Q_{large} but treating positive kth waves located in $\mathbf{R}\setminus L^u(t)$ as perturbations. Then Lemma 4.4 and its proof are still valid, with V exchanged there to $\tilde V$. Thus the estimate in (5.1) follows.

In order to deduce (5.2) we may restrict our attention to the case $t=T$ and $s=0$. It is convenient to consider the evolution of the related functional

$$\tilde\Gamma(t) = |y'(t)-y'(0)| + \kappa\cdot\tilde V(t) + \kappa^2\cdot Q(t),$$

where $\tilde V(t)$ is defined as the sum of strengths of perturbation waves α in

$$\{x_\alpha < y(t) \text{ and } i_\alpha\ge k\}\cup\{x_\alpha>y(t) \text{ and } i_\alpha\le k\}$$

and $\kappa > 1$ is a large constant. We see that when $y(t)$ interacts with another wave α then $\Delta Q \le 0$, $\Delta y' = \mathcal{O}(1)|\epsilon_\alpha|$, and $\Delta \tilde{V} = -|\epsilon_\alpha|$. On the other hand, at any other time $\Delta y' = 0$ and $\Delta(\tilde{V} + \kappa Q) \le 0$. Thus $\tilde{\Gamma}$ is a nonincreasing function of t only if κ is large. Hence $|y'(T) - y'(0)| \le \tilde{\Gamma}(0) = \mathcal{O}(1)\epsilon_0$, and (5.2) follows since

$$y'(t) = \lambda_k(u^\epsilon(t, y(t)-))$$

for almost all $t \in [0, T]$. □

6. Lyapunov functional: A proof of Theorem 2. Toward a proof of Theorem 2, in this section we carry out the construction of the Lyapunov functional Φ. Following [LY, BLY], $\Phi(u, v)$ is supposed to control the L^1 distance between the two ϵ-approximate solutions $u, v : [0, \infty) \times \mathbf{R} \longrightarrow \mathbf{R}^n$ obtained by the wave front tracking algorithm and thus enjoying the properties in Lemma 4.5. Assuming the L^1 stability condition (3.1), the two crucial properties of Φ will be the following:

$$\Phi(u(t,\cdot), v(t,\cdot)) \le \Phi(u(s,\cdot), v(s,\cdot)) + C \cdot \epsilon \cdot (t - s), \tag{6.1}$$

$$\frac{1}{C} \cdot ||u(t,\cdot) - v(t,\cdot)||_{L^1} \le \Phi(u(t,\cdot), v(t,\cdot)) \le C \cdot ||u(t,\cdot) - v(t,\cdot)||_{L^1}, \tag{6.2}$$

for all $t > s \ge 0$ and a uniform constant $C > 0$ depending only on the system (1.1). In the remaining part of the article we will concentrate on proving (6.1), (6.2) for a functional Φ constructed below. Taking then $\mathcal{D} = \bar{\mathcal{D}}_{\epsilon_0}$, for a small $\epsilon_0 > 0$, the proof of Theorem 2 will follow by the already standard argument as in Chapter 8.3 of [B].

Fix a positive and small constant ν. Given piecewise constant functions u and v, let

$$T = \sup\Big\{t > 0; \quad \exists_x \quad |\lambda_k(u(t,x)) - \lambda_k(v(t,x))| > \nu\Big\}. \tag{6.3}$$

LEMMA 6.1. *T defined as above is finite.*

Proof. Notice that since the total strength of perturbation waves is of the order ϵ_0 at each time t, then taking $\epsilon \ll \epsilon_0$ we have

$$\sup_{t \ge 1, x} ||u(t,x) - \tilde{u}(t,x)|| + \sup_{t \ge 1, x} ||v(t,x) - \tilde{v}(t,x)|| = \mathcal{O}(1)\epsilon_0. \tag{6.4}$$

The functions $\tilde{u}$ and $\tilde{v} : [1, +\infty) \times \mathbf{R} \longrightarrow \mathbf{R}^n$ are smooth solutions to (1.1) with initial data

$$\tilde{u}(1,x) = u_0(1, \psi(x)), \quad \tilde{v}(1,x) = v_0(1, \phi(x)),$$

where ψ and $\phi : \mathbf{R} \longrightarrow \mathbf{R}$ are some increasing diffeomorphisms. We want to show that

$$\lim_{t \to +\infty} \sup |\lambda_k(\tilde{u}(t,x)) - \lambda_k(\tilde{v}(t,x))| = 0, \tag{6.5}$$

which in view of (6.4) and taking $\epsilon \ll \epsilon_0$ will imply that $T < +\infty$.

Notice that for each $t \ge 1$, $\tilde{u}$ is constant outside the interval

$$J_t^u = \big[\psi^{-1}(\lambda_k(u_l)) + \lambda_k(u_l) \cdot (t-1), \psi^{-1}(\lambda_k(u_r)) + \lambda_k(u_r) \cdot (t-1)\big]$$

and that it propagates along the straight lines—characteristics having slopes λ_k inside the region $\{(t,x);\ x\in J_t^u\}$. Consequently, one has

$$\sup_{x\in J_t^u\cap J_t^v}|\lambda_k(\tilde u(t,x))-\lambda_k(\tilde v(t,x))|\leq \frac{\max\limits_{w,z\in\{u_l,u_r\}}|\psi^{-1}(\lambda_k(w))-\phi^{-1}(\lambda_k(z))|}{t-1}, \tag{6.6}$$

where the interval J_t^v is defined as J_t^u, by means of the diffeomorphism ϕ. Obviously, the right-hand side of (6.6) vanishes as $t\to+\infty$. Likewise, $\sup_{x\notin J_t^u\cap J_t^v}|\lambda_k(\tilde u(t,x)-\lambda_k(\tilde v(t,x))|$ also converges to 0, because of the spreading of the rarefactions in $\tilde u$ and $\tilde v$. This establishes (6.5). □

The definition of the functional $\Phi(u,v)$ falls in two parts.

Case 1 *(the profiles* $\mathbf{u}$ *and* $\mathbf{v}$ *are apart from each other):* $\mathbf{t}\in[\mathbf{0},\mathbf{T}]$. Let $T>0$. Without loss of generality we may assume that for some x there holds $\lambda_k(u(t,x))>\lambda_k(v(t,x))+3\nu/4$ (the case of the opposite inequality may be treated similarily). Because of the estimate in (5.1) and taking $\epsilon_0\ll\nu$, there exists then a nonempty interval $I(T)=[z_0^-,z_0^+]$ such that $z_0^-\in L^u(T),\ z_0^+\in L^v(T)$ and

$$\forall x,y\in I(T)\quad \lambda_k(u(T,x))-\lambda_k(v(T,y))>\nu/2. \tag{6.7}$$

For $t\in[0,T]$ call $I(t)$ the space interval whose boundary is continuous polygonals $z^-(t)\in L^u(t),\ z^+(t)\in L^v(t)$ with $z^-(T)=z_0^-$ and $z^+(T)=z_0^+$. Notice that, taking ϵ_0 small enough, Lemma 5.1 yields

$$\forall t\in[0,T],\quad \forall x,y\in I(t)\quad \lambda_k(u(t,x))-\lambda_k(v(t,y))>\nu/3. \tag{6.8}$$

For all $t\in[0,T)$ the Lyapunov functional Φ is defined by the formula

$$\Phi(u,v)(t)=||u(t,\cdot)-v(t,\cdot)||_{L^1}+\kappa_1\cdot|I(t)|, \tag{6.9}$$

where $|I(t)|$ stands for the length of the interval $I(t)$ and κ_1 is a sufficiently large integer constant.

LEMMA 6.2. *If only* κ_1 *is large enough, then the functional* Φ *satisfies*

$$\Phi(u(t',\cdot),v(t',\cdot))\leq\Phi(u(t,\cdot),v(t,\cdot)), \tag{6.10}$$

$$||u(t,\cdot)-v(t,\cdot)||_{L^1}\leq\Phi(u(t,\cdot),v(t,\cdot))\leq C\cdot||u(t,\cdot)-v(t,\cdot)||_{L^1} \tag{6.11}$$

for all $0\leq t\leq t'\leq T$ *and a uniform constant* $C>0$.

Proof. The equivalence (6.11) of Φ with the L^1 distance follows in view of (6.8).

Denote by $\mathcal{J}(u)$ and $\mathcal{J}(v)$ the sets of all jumps in u and v, respectively. To prove (6.10) fix $t\in[0,T)$, which is not a time of interaction of any couple of fronts in u or v. We have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\Phi(u,v)(t)=&\\ \sum_{\alpha\in\mathcal{J}(u)\cup\mathcal{J}(v)}&\Big||u(x_\alpha+,t)-v(x_\alpha+,t)|-|u(x_\alpha-,t)-v(x_\alpha-,t)|\Big|\cdot\dot x_\alpha\\ &+\kappa_1\cdot\frac{\mathrm{d}}{\mathrm{d}t}|I(t)|.\end{aligned} \tag{6.12}$$

The first term in (6.12) is of the order of $\mathcal{O}(1)$ because of the finite speed of propagation, boundedness of $TV(u(t))$ and $TV(v(t))$, and

$$|u(x_\alpha+,t)-v(x_\alpha+,t)|-|u(x_\alpha-,t)-v(x_\alpha-,t)|=\mathcal{O}(1)|\epsilon_\alpha|.$$

On the other hand, in view of (6.8) we have $\mathrm{d}/\mathrm{d}t\ |I(t)|\le -\nu/4$. Thus if κ_1 is large with respect to the system constants and the prechosen ν, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi(u,v)(t)\le 0.$$

Integrating in time we conclude (6.10). □

Case 2 *(***u** *and* **v** *close):* $\mathbf{t}\ge\mathbf{T}$. The Lyapunov functional Φ is defined as in [BLY, B]:

$$\Phi(u,v)=\int_{-\infty}^{+\infty}\sum_{i=1}^{n}W_i(x)\cdot w_i(x)\cdot|q_i(x)|\,\mathrm{d}x. \tag{6.13}$$

The scalar quantities $q_i(x)$ are, roughly speaking, the curvilinear coordinates of the vector $v(x)-u(x)$, computed along combinations of shock curves in Ω. The precise definition of W_i and w_i will be our concern in what follows.

The coordinates $\{q_i(x)\}_{i=1}^n$ are implicitely defined by

$$v(x)=\mathcal{S}_n(q_n(x))\circ\ldots\mathcal{S}_k(q_k(x))\circ\ldots\mathcal{S}_1(u(x),q_1(x)). \tag{6.14}$$

Such decomposition exists if ν is small enough, as $|\lambda_k(u(x,t))-\lambda_k(v(x,t))|\le\nu$ for all x and $t\ge T$. The weights $w_i(x)$ are given by

$$w_i(x)=w_i\Big(\mathcal{S}_{i-1}(q_{i-1}(x))\circ\ldots\mathcal{S}_1(u(x),q_1(x))\Big), \tag{6.15}$$

where the w_i's in the right-hand side are given by (2.1) and the L^1 stability condition (3.1). We see that the weights $w_i(x)$ in (6.15) are computed at the left states of the corresponding waves. Recall that $w_k>0$ is constant in Ω.

We will now define the functional weights $W_i(x)$. Recall that $i_\alpha\in\{1\ldots n+1\}$ is the family of the jump located at x_α with strength ϵ_α. Also, by $\mathcal{J}(u)$ and $\mathcal{J}(v)$ we denote the sets of all jumps in u and v. Let $\mathcal{P}(u)$ and $\mathcal{P}(v)$ be the respective subsets of $\mathcal{J}(u)$ and $\mathcal{J}(v)$, containing those α for which $i_\alpha\ne n+1$ and either $i_\alpha\ne k$ or $i_\alpha=k$ and $\epsilon_\alpha<0$.

Define the quantities $A_i(x)$ measuring the total amount of physical perturbation waves in u and v which approach the ith wave $q_i(x)$ located at x [BLY]. More precisely, when the ith field is linearly degenerate we set

$$A_i(x)=\left[\sum_{\substack{\alpha\in\mathcal{P}(u)\cup\mathcal{P}(v)\\ x_\alpha<x,\ i_\alpha>i}}+\sum_{\substack{\alpha\in\mathcal{P}(u)\cup\mathcal{P}(v)\\ x_\alpha>x,\ i_\alpha<i}}\right]|\epsilon_\alpha|.$$

For a genuinely nonlinear ith field

$$A_i(x) = \left[\sum_{\substack{\alpha\in\mathcal{P}(u)\cup\mathcal{P}(v)\\ x_\alpha<x,\ i_\alpha>i}} + \sum_{\substack{\alpha\in\mathcal{P}(u)\cup\mathcal{P}(v)\\ x_\alpha>x,\ i_\alpha<i}}\right]|\epsilon_\alpha| + \begin{cases} \left[\sum_{\substack{\alpha\in\mathcal{P}(u)\\ x_\alpha<x,\ i_\alpha=i}} + \sum_{\substack{\alpha\in\mathcal{P}(v)\\ x_\alpha>x,\ i_\alpha=i}}\right]|\epsilon_\alpha| & \text{if } q_i(x)<0,\\ \left[\sum_{\substack{\alpha\in\mathcal{P}(v)\\ x_\alpha<x,\ i_\alpha=i}} + \sum_{\substack{\alpha\in\mathcal{P}(u)\\ x_\alpha>x,\ i_\alpha=i}}\right]|\epsilon_\alpha| & \text{if } q_i(x)\geq 0. \end{cases}$$

Define

$$\forall i: 1\dots n \quad W_i(x) = 1 + \kappa_2(Q(u)+Q(v)) + \kappa_3 A_i(x) + \delta_{ik}\cdot\kappa_4|q_k(x)|. \tag{6.16}$$

Here Q stands for the Glimm's interaction potential from Definition 4.3 and δ_{ik} is the Kronecker delta. The (large) constants $\kappa_2,\kappa_3,\kappa_4$ are to be determined later; we see that as soon as they have been assigned, we can impose a suitably small bound on the amount of perturbation in u and v (by taking ϵ_0 small in (4.5), or in particular δ small in Theorem 1) so that

$$1\leq W_i(x)\leq 4 \qquad \text{for all } i,x. \tag{6.17}$$

This ends the definition of the functional Φ.

LEMMA 6.3. *The functional Φ constructed above satisfies* (6.1) *and*

$$\frac{1}{C}||u(t,\cdot)-v(t,\cdot)||_{L^1}\leq\Phi(u(t,\cdot),v(t,\cdot))\leq||u(t,\cdot)-v(t,\cdot)||_{L^1} \tag{6.18}$$

for all $t'>t\geq T$ and a uniform constant $C>0$ depending only on the system (1.1).

Proof. The equivalence of Φ with the L^1 distance as in (6.18) follows from (6.17) if we take the weights $\{w_i\}_{i=1}^n$ small enough.

To prove the estimate in (6.1), define $\lambda_i(x)$ as the Rankine–Hugoniot speed of the shock/contact $q_i(x)$.

Recall that a direct calculation [BLY] gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi(u(t),v(t)) = \sum_{\alpha\in\mathcal{J}(u)\cup\mathcal{J}(v)}\sum_{i=1}^n E_{\alpha,i}, \tag{6.19}$$

with

$$\begin{aligned} E_{\alpha,i} = {} & \left(W_i\cdot w_i\cdot|q_i|\right)(x_\alpha+)\cdot(\lambda_i(x_\alpha+)-\dot{x}_\alpha)\\ & -\left(W_i\cdot w_i\cdot|q_i|\right)(x_\alpha-)\cdot(\lambda_i(x_\alpha-)-\dot{x}_\alpha). \end{aligned} \tag{6.20}$$

Above $\dot{x}_\alpha$ denotes the speed of propagation of the wave α located at x_α. We will prove that

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi(u(t),v(t))\leq\mathcal{O}(1)\epsilon \tag{6.21}$$

for every time $t \geq T$ where the fronts in u or v do not interact. Indeed, this will be the goal of the next section.

Next, let t be such that say fronts ϵ_α and ϵ_β in u interact. It is easy to notice that for every x and i we have

$$A_i(t+, x) - A_i(t-, x) \leq \mathcal{O}(1)|\epsilon_\alpha \epsilon_\beta|.$$

On the other hand, by Lemma 4.4, the quantity $Q(u)$ decreases by the same order of magnitude. Thus if κ_2 in (6.16) is large enough, all functional weights $W_i(x)$ must decrease across the time t. Consequently, the whole functional Φ decreases as well. Based on these two observations and integrating (6.21) in time, we conclude (6.1). □

7. Stability estimates. In this section we want to establish the inequality (6.21) by estimating local terms $E_{\alpha,i}$ in (6.20). All calculations refer to a fixed jump $\alpha \in \mathcal{J}(v)$, propagating with speed $\dot{x}_\alpha$ and belonging to a characteristic family $i_\alpha : 1 \dots n+1$. When $\alpha \in \mathcal{J}(u)$, only minimal and obvious modifications of our arguments are required and so we leave them to the reader.

We first focus on the case $i_\alpha = n+1$. We will prove that

$$\sum_{i=1}^{n} E_{\alpha,i} \leq \mathcal{O}(1)|\epsilon_\alpha|. \tag{7.1}$$

Indeed,

$$\forall\ i \neq k \quad |w_i^+ q_i^+ - w_i^- q_i^-| + |\lambda_i^+ - \lambda_i^-| = \mathcal{O}(1)|\epsilon_\alpha|.$$

Also, for $i \neq k$ and if $\operatorname{sgn} q_i^- = \operatorname{sgn} q_i^+$ we have $W_i^+ = W_i^-$. On the other hand, if $\operatorname{sgn} q_i^- \neq \operatorname{sgn} q_i^+$, then $|q_i^+| + |q_i^-| = \mathcal{O}(1)|\epsilon_\alpha|$ and, consequently,

$$\sum_{i \neq k} E_{\alpha,i} \leq \mathcal{O}(1)|\epsilon_\alpha|.$$

In a similar manner, $E_{\alpha,k} \leq \mathcal{O}(1)|\epsilon_\alpha|$ if $\operatorname{sgn} q_k^- \neq \operatorname{sgn} q_k^+$. The same is true if $\operatorname{sgn} q_k^- = \operatorname{sgn} q_k^+$ because then

$$\Delta W_k = \mathcal{O}(1)|\epsilon_\alpha|.$$

The bound (7.1) is thus proven. Now, recalling Lemma 4.5(iv), (7.1) yields

$$\sum_{i_\alpha = n+1} \sum_{i=1}^{n} E_{\alpha,i} \leq \mathcal{O}(1)\epsilon. \tag{7.2}$$

Let now $i_\alpha : 1 \dots n$. Our goal will be to prove that

$$\sum_{i=1}^{n} E_{\alpha,i} = \mathcal{O}(1)\epsilon|\epsilon_\alpha|. \tag{7.3}$$

Recall that by Lemma 4.5 $|\epsilon_\alpha| < \epsilon$, whenever α is a rarefaction wave. In view of (1.8) and the definition (6.20) we may thus without loss of generality replace each rarefaction wave α by a (possibly nonentropic) shock having the original strength ϵ_α and the speed $\dot{x}_\alpha = \lambda_k(v(x_\alpha -))$. We will prove that with this modification the same

estimate as in (7.3) holds. For simplicity, we write $W_i^+ = W_i(x_\alpha+)$, $q_i^- = q_i(x_\alpha-)$, etc.

The proof falls in several cases. Throughout the calculations, we often use the estimates from section 8. When α is a part of the rarefaction $\mathcal{R}_k$, our estimates rely on the stability condition (3.1); the parameters w_k and ν are chosen so that the negative term in (8.3) overcomes extra contributions which are not of the order $\mathcal{O}(1)\epsilon_\alpha^2$. When α is a perturbation wave, our argument is essentially a modification of the one from [BLY]. We again adjust ν appropriately and then take the constant κ_3 in (6.16) to be large with respect to other quantities in the derived estimates. The parameter ϵ_0, measuring the amount of perturbing waves present at any time in both approximate solutions u and v, is always set to be as small as needed, in particular $\epsilon_0 \ll \nu$.

Case 1. $i_\alpha = k$ and $\epsilon_\alpha > 0$. Recall that by Lemma 4.5 we have $|\epsilon_\alpha| < \epsilon$. We will prove

$$\sum_{i=1}^{n} E_{\alpha,i} = \mathcal{O}(1)\epsilon_\alpha^2, \tag{7.4}$$

which will clearly imply (7.3). We first estimate

$$\begin{aligned}\sum_{i\neq k} E_{\alpha,i} =& \sum_{i\neq k} (\Delta W_i)\cdot w_i^-|q_i^-|(\lambda_i^- - \dot{x}_\alpha)\\ &+ \sum_{i\neq k} W_i^+ \cdot \left[w_i^+|q_i^+|(\lambda_i^+ - \dot{x}_\alpha) - w_i^-|q_i^-|(\lambda_i^- - \dot{x}_\alpha)\right].\end{aligned} \tag{7.5}$$

Fix $i \neq k$. Notice that if $\operatorname{sgn} q_i^+ \neq \operatorname{sgn} q_i^-$, then

$$|q_i^-| \leq |q_i^+ - q_i^-| \quad \text{and} \quad \Delta W_i = \mathcal{O}(1)\kappa_3\epsilon_0. \tag{7.6}$$

On the other hand, if $\operatorname{sgn} q_i^+ = \operatorname{sgn} q_i^-$, then $\Delta W_i = 0$. Thus the first summand in (7.5) can be estimated using Lemma 8.1:

$$\begin{aligned}&\sum_{i\neq k} (\Delta W_i)\cdot w_i^-|q_i^-|(\lambda_i^- - \dot{x}_\alpha)\\ &\qquad \leq \mathcal{O}(1)\kappa_3\epsilon_0 \cdot \left[\epsilon_\alpha \cdot \left(\sum_{s>k}|q_s^-|\right) + \epsilon_\alpha\cdot|q_k^-|^2 + \epsilon_\alpha^2\right].\end{aligned} \tag{7.7}$$

In order to deal with the second summand in (7.5), we notice that if $\operatorname{sgn} q_i^+ \neq \operatorname{sgn} q_i^-$, then by (7.6) and Lemma 8.1, there holds

$$\begin{aligned}&\left|w_i^+|q_i^+|(\lambda_i^+ - \dot{x}_\alpha) - w_i^-|q_i^-|(\lambda_i^- - \dot{x}_\alpha)\right|\\ &\qquad \leq \mathcal{O}(1)\left[\epsilon_\alpha \cdot \left(\sum_{s>k}|q_s^-|\right) + \epsilon_\alpha\cdot|q_k^-|^2 + \epsilon_\alpha^2\right].\end{aligned} \tag{7.8}$$

The same is true when $\operatorname{sgn} q_i^+ = \operatorname{sgn} q_i^-$, as in this case the left-hand side of (7.8) equals $|w_i^+ q_i^+(\lambda_i^+ - \dot{x}_\alpha) - w_i^- q_i^-(\lambda_i^- - \dot{x}_\alpha)|$ and so one can again employ the estimates

of Lemma 8.1. In view of Remark 2.3, combining (7.5), (7.7), and (7.8) we obtain

$$\begin{aligned}
&\sum_{i\neq k} W_i^+ \cdot \left[w_i^+|q_i^+|(\lambda_i^+ - \dot{x}_\alpha) - w_i^-|q_i^-|(\lambda_i^- - \dot{x}_\alpha)\right] \\
&\leq [1+\kappa_2(Q(u)+Q(v))] \cdot \sum_{i\neq k} \left[w_i^+|q_i^+|(\lambda_i^+ - \dot{x}_\alpha) - w_i^-|q_i^-|(\lambda_i^- - \dot{x}_\alpha)\right] \\
&\quad + \mathcal{O}(1)\kappa_1\epsilon_0 \cdot \left[\epsilon_\alpha \cdot \left(\sum_{s>k}|q_s^-|\right) + \epsilon_\alpha \cdot |q_k^-|^2 + \epsilon_\alpha^2\right].
\end{aligned} \tag{7.9}$$

Estimating the first term in the right-hand side of (7.9) by Lemma 8.3 and noting (7.7), the quantity in (7.5) can be further bounded by

$$\sum_{i\neq k} E_{\alpha,i} \leq -\frac{\gamma_1}{2}\epsilon_\alpha \cdot \left(\sum_{s>k}|q_s^-|\right) + \mathcal{O}(1) \cdot \left[\epsilon_\alpha \cdot |q_k^-|^2 + \epsilon_\alpha^2\right] \tag{7.10}$$

if ϵ_0 is small enough.

We now aim at establishing (7.4) by estimating the remaining term $E_{\alpha,k}$. We distinguish two subcases.

Subcase 1.1. $\operatorname{sgn} q_k^+ \neq \operatorname{sgn} q_k^-$. Then

$$\Delta W_k = \mathcal{O}(1)\kappa_4\epsilon_\alpha + \mathcal{O}(1)\kappa_3\epsilon_0.$$

Therefore we have

$$\begin{aligned}
(\Delta W_k)w_k|q_k^-|(\lambda_k^- - \dot{x}_\alpha) &\leq \mathcal{O}(1)w_k\epsilon_\alpha\left(\kappa_4\epsilon_\alpha + \kappa_3\epsilon_0\right) \cdot \left(\epsilon_\alpha + \sum_{s>k}|q_s^-|\right) \\
&\leq \mathcal{O}(1)w_k\kappa_1\epsilon_0\epsilon_\alpha \cdot \left(\epsilon_\alpha + \sum_{s>k}|q_s^-|\right) + \mathcal{O}(1)\kappa_4\epsilon_\alpha^2.
\end{aligned} \tag{7.11}$$

On the other hand,

$$\begin{aligned}
&W_k^+ w_k \left[|q_k^+|(\lambda_k^+ - \dot{x}_\alpha) - |q_k^-|(\lambda_k^- - \dot{x}_\alpha)\right] \\
&\qquad \leq \mathcal{O}(1)w_k\epsilon_\alpha\left[|\lambda_k^+ - \dot{x}_\alpha| + |\lambda_k^- - \dot{x}_\alpha|\right] \leq \mathcal{O}(1)w_k\epsilon_\alpha \cdot \left(\sum_{s>k}|q_s^-|\right).
\end{aligned} \tag{7.12}$$

Summing (7.11) and (7.12) we obtain

$$E_{\alpha,k} = \mathcal{O}(1)w_k\epsilon_\alpha \cdot \left(\sum_{s>k}|q_s^-|\right) + \mathcal{O}(1)\kappa_4\epsilon_\alpha^2. \tag{7.13}$$

The bound (7.4) now follows by (7.13) and (7.10) if only w_k is chosen suitably small with respect to the constant γ_1 and for small ϵ_0.

Subcase 1.2. $\operatorname{sgn} q_k^+ = \operatorname{sgn} q_k^-$. By Lemma 8.1, we have

$$\Delta|q_k| = (\operatorname{sgn} q_k)\cdot\epsilon_\alpha + \mathcal{O}(1)\epsilon_\alpha\left(|q_k^-|^2 + \left(\sum_{s>k}|q_s^-|\right) + \epsilon_\alpha\right).$$

Thus, if only ϵ_0 and ν are small enough,

$$(\text{sgn } q_k)\cdot\Delta|q_k|\geq\epsilon_\alpha/2.$$

Moreover,

$$\lambda_k^- - \dot{x}_\alpha = \mathcal{O}(1)\left(\sum_{s>k}|q_s^-|\right) + (\text{sgn } q_k)\cdot\left(-\frac{|q_k^-|}{2}+\mathcal{O}(1)|q_k^-|^2\right)+\mathcal{O}(1)\epsilon_\alpha^2. \tag{7.14}$$

Recall that $\Delta W_k=\kappa_4\Delta|q_k|$. Hence,

$$\begin{aligned}(\Delta W_k)\cdot w_k|q_k^-|(\lambda_k^- - \dot{x}_\alpha) &= \kappa_4 w_k\cdot(\Delta|q_k|)\cdot|q_k^-|(\lambda_k^- - \dot{x}_\alpha)\\ &= w_k\kappa_4\cdot\left[\mathcal{O}(1)\epsilon_\alpha|q_k^-|\left(\sum_{s>k}|q_s^-|\right)+\mathcal{O}(1)\epsilon_\alpha|q_k^-|^3\right.\\ &\qquad\left.-\ \frac{1}{2}(\Delta|q_k|)(\text{sgn } q_k)|q_k^-|^2\right]+\mathcal{O}(1)\kappa_4\epsilon_\alpha^2\\ &\leq w_k\cdot\left[\mathcal{O}(1)\kappa_4\epsilon_\alpha|q_k^-|\left(\sum_{s>k}|q_s^-|\right)+\mathcal{O}(1)\epsilon_\alpha|q_k^-|^2\right]\\ &\quad-\ w_k\frac{\kappa_4}{4}\epsilon_\alpha|q_k^-|^2+\mathcal{O}(1)\kappa_4\epsilon_\alpha^2.\end{aligned} \tag{7.15}$$

Now, using (7.14) and Lemma 8.1 we obtain

$$(q_k^+ - q_k^-)(\lambda_k^- - \dot{x}_\alpha) = -\frac{q_k^-\epsilon_\alpha}{2}+\mathcal{O}(1)\epsilon_\alpha\left(|q_k^-|^2+\left(\sum_{s>k}|q_s^-|\right)+\epsilon_\alpha\right). \tag{7.16}$$

On the other hand, by Lemma 8.1

$$q_k^+(\lambda_k^+ - \lambda_k^-) = \frac{q_k^-\epsilon_\alpha}{2}+\mathcal{O}(1)\epsilon_\alpha\left(|q_k^-|^2+\left(\sum_{s>k}|q_s^-|\right)+\epsilon_\alpha\right).$$

Thus, in view of (7.16),

$$q_k^+(\lambda_k^+ - \dot{x}_\alpha) - q_k^-(\lambda_k^- - \dot{x}_\alpha) = \mathcal{O}(1)\epsilon_\alpha\left(|q_k^-|^2+\left(\sum_{s>k}|q_s^-|\right)+\epsilon_\alpha\right).$$

The above bound combined with (7.15) yields

$$\begin{aligned}E_{\alpha,k} = w_k\cdot\Bigg[&-\frac{\kappa_4}{5}\epsilon_\alpha|q_k^-|^2+\mathcal{O}(1)\kappa_4\epsilon_\alpha|q_k^-|\left(\sum_{s>k}|q_s^-|\right)\\ &+\mathcal{O}(1)\epsilon_\alpha\left(\sum_{s>k}|q_s^-|\right)\Bigg]+\mathcal{O}(1)\kappa_4\epsilon_\alpha^2,\end{aligned} \tag{7.17}$$

if only the constant κ_4 is larger than several independent quantities $\mathcal{O}(1)$ in the above series of estimates. Combining (7.17) and (7.10) we obtain (7.4) for w_k small and κ_4 large enough.

Case 2. $i_\alpha \neq k$. Note that for $i \neq k$ the quantities $E_{\alpha,i}$ can be estimated exactly as in [BLY]; see also [B] of Chapter 8.2. On the other hand, for $i = k$

$$\Delta W_k = \kappa_3 \cdot \operatorname{sgn}\,(i_\alpha - k) \cdot |\epsilon_\alpha| + \kappa_4 \cdot \Delta|q_k|$$

and

$$\Delta|q_k| = \mathcal{O}(1)|\epsilon_\alpha| \cdot \sum_{i=1}^{n} |q_i^-| = \mathcal{O}(1)|\epsilon_\alpha|(\epsilon_0 + \nu).$$

Thus the term in $E_{\alpha,k}$ containing ΔW_k can be estimated as follows:

$$\begin{aligned}(\Delta W_k) w_k |q_k^-|(\lambda_k^- - \dot{x}_\alpha) \leq & -\kappa_3 w_k \epsilon_\alpha |q_k^-||\lambda_k^- - \dot{x}_\alpha| \\ & + \mathcal{O}(1)\kappa_4 w_k \epsilon_\alpha(\epsilon_0 + \nu)|q_k^-||\lambda_k^- - \dot{x}_\alpha| \\ \leq & -\frac{\kappa_3}{2} w_k \epsilon_\alpha |q_k^-||\lambda_k^- - \dot{x}_\alpha|,\end{aligned}$$

if only $\epsilon_0 + \nu$ is small enough. The analysis in [BLY] can thus be applied to get (7.3).

Case 3. $i_\alpha = k$ and $\epsilon_\alpha < 0$. If $|\epsilon_\alpha| < \epsilon$ and $|q_k^-| \leq 2|\epsilon_\alpha|$, then recalling that $\Delta W_k \leq W_k^- + W_k^+ \leq 8$ by (6.17), and using (8.64) from [B], we conclude (7.3). The same argumentation as on page 167 of [B] yields (7.3) when $q_k^+ < 0 < q_k^-$.

We will now focus on the case when q_k^- and q_k^+ have the same sign. In view of the analysis of Lemma 8.3 we have

$$\begin{aligned}\Delta W_k &= \kappa_3(\operatorname{sgn} q_k)|\epsilon_\alpha| + \kappa_4|q_k^+ - q_k^-| = \kappa_3(\operatorname{sgn} q_k)|\epsilon_\alpha| \\ &+ \kappa_4(\operatorname{sgn} q_k) \cdot \left[-|\epsilon_\alpha| + \mathcal{O}(1)|\epsilon_\alpha||q_k^-|^2 + \mathcal{O}(1)|\epsilon_\alpha| \left(\sum_{s>k} |q_s^-| \right) + \mathcal{O}(1)\epsilon_\alpha^2 \right].\end{aligned} \tag{7.18}$$

Recalling the formula (8.50) from [B],

$$\dot{x}_\alpha - \lambda_k^- = \frac{q_k^- + \epsilon_\alpha}{2} + \mathcal{O}(1) \left[|q_k^- + \epsilon_\alpha|(|q_k^-| + |\epsilon_\alpha|) + \sum_{s \neq k} |q_s^-| \right],$$

the estimate (7.18) implies for κ_3 large (also $\kappa_3 > 2\kappa_4$) and ϵ_0 small that

$$\begin{aligned}(\Delta W_k) w_k |q_k^-|(\lambda_k^- - \dot{x}_\alpha) \leq & -\frac{\kappa_3}{3} w_k |\epsilon_\alpha||q_k^-||q_k^- + \epsilon_\alpha| \\ & + \mathcal{O}(1)\kappa_3 w_k |\epsilon_\alpha||q_k^-| \cdot \left(\sum_{s \neq k} |q_s^-| \right) + \mathcal{O}(1)\kappa_4 \epsilon_\alpha^2.\end{aligned} \tag{7.19}$$

Now, by the same reasoning as in Chapter 8.2, page 165 [B], we see that for ν small and some constant $c > 0$, there holds

$$\begin{aligned}W_k w_k \Delta[|q_k|(\lambda_k - \dot{x}_\alpha)] + \sum_{i \neq k} E_{\alpha,i} \leq & -c\kappa_3 |\epsilon_\alpha| \sum_{s \in \mathcal{I}} |q_s^-| \\ & + \mathcal{O}(1)|\epsilon_\alpha| \left(|q_k^-||q_k^- + \epsilon_\alpha| + \sum_{s \neq k} |q_s^-| \right),\end{aligned} \tag{7.20}$$

$$\sum_{i\neq k}|q_i^-| \leq |q_k^-||q_k^- + \epsilon_\alpha| + 2\sum_{s\in\mathcal{I}}|q_s^-|. \tag{7.21}$$

The index set $\mathcal{I}$ is defined as $\mathcal{I} = \{i : 1\dots n;\ i \neq k \text{ and } \operatorname{sgn} q_i^- = \operatorname{sgn} q_i^+\}$. Thus (7.19) becomes by (7.21)

$$\begin{aligned}(\Delta W_k)w_k|q_k^-|(\lambda_k^- - \dot{x}_\alpha) \leq -\ \frac{\kappa_3}{4}w_k|\epsilon_\alpha||q_k^-||q_k^- + \epsilon_\alpha| \\ +\ \mathcal{O}(1)\kappa_3 w_k|\epsilon_\alpha||q_k^-|\cdot\left(\sum_{s\in\mathcal{I}}|q_s^-|\right) + \mathcal{O}(1)\kappa_4\epsilon_\alpha^2\end{aligned}$$

if only ν is small enough. In view of (7.20), this implies

$$\begin{aligned}\sum_{i=1}^n E_{\alpha,i} \leq -\ c\kappa_3|\epsilon_\alpha|\left(\sum_{s\in\mathcal{I}}|q_s^-|\right) + \mathcal{O}(1)|\epsilon_\alpha|\left(|q_k^-||q_k^- + \epsilon_\alpha| + \sum_{s\in\mathcal{I}}|q_s^-|\right) \\ -\ \frac{\kappa_3}{4}w_k|\epsilon_\alpha||q_k^-||q_k^- + \epsilon_\alpha| + \mathcal{O}(1)\kappa_3 w_k|\epsilon_\alpha||q_k^-|\cdot\left(\sum_{s\in\mathcal{I}}|q_s^-|\right) + \mathcal{O}(1)\kappa_4\epsilon_\alpha^2,\end{aligned}$$

and consequently we obtain (7.4) for κ_3 large.

8. Technical lemmas.

LEMMA 8.1. *Let*

$$v = \mathcal{S}_n(q_n^-)\circ\dots\mathcal{S}_1(u,q_1^-), \qquad \mathcal{S}_k(v,\epsilon_\alpha) = \mathcal{S}_n(q_n^+)\circ\dots\mathcal{S}_1(u,q_1^+),$$

with $u\in\Omega$ and $\{q_i^-\}_{i=1}^n, \epsilon_\alpha$ small enough. For every $i : 1\dots n$, call $\lambda_i^\pm$ the speed of the shock wave $q_i^\pm$, as in (1.11). *Let* E *be any quantity satisfying the bound*

$$\mathrm{E} = \mathcal{O}(1)|\epsilon_\alpha|\left\{|q_k^-|^2 + \sum_{s>k}|q_s^-| + |\epsilon_\alpha|\right\}.$$

Then

(i) $|q_k^+ - q_k^- - \epsilon_\alpha| + \sum_{i\neq k}|q_i^+ - q_i^-| = \mathrm{E}$,

(ii) $\lambda_k^+ - \lambda_k^- = \epsilon_\alpha/2 + \mathrm{E}$,

(iii) *for all $i<k$ we have $\lambda_i^+ - \lambda_i^- = \mathrm{E}$, while for all $i>k$ there is $\lambda_i^+ - \lambda_i^- = \mathcal{O}(1)|\epsilon_\alpha| + \mathrm{E}$.*

Proof. We will prove only (i), the other assertions following in similar manner. For every $i : 1\dots n$, introduce an auxiliary function G_i:

$$G_i(u, q_1^-\dots q_n^-, \epsilon_\alpha) = q_i^+ - q_i^-.$$

We have

$$\begin{aligned}G_i = \epsilon_\alpha\cdot\left[\frac{\partial G_i}{\partial\epsilon_\alpha}(u, q_1^-\dots q_k^-, q_i^- = 0 \text{ for } i>k, \epsilon_\alpha = 0) + \mathcal{O}(1)\sum_{s>k}|q_s^-|\right] \\ +\ \mathcal{O}(1)\epsilon_\alpha^2.\end{aligned} \tag{8.1}$$

Moreover,

$$G_i(u, q_1^- \dots q_k^-, q_i^- = 0 \text{ for } i > k, \epsilon_\alpha = 0) - \delta_{ik} \cdot \epsilon_\alpha = \mathcal{O}(1)||G(u_{k-1}^-, q_k^-, \epsilon_\alpha)||, \tag{8.2}$$

where the quantity G is defined as

$$G(u_{k-1}^-, q_k^-, \epsilon_\alpha) = \mathcal{S}_k(u_{k-1}^-, q_k^- + \epsilon_\alpha) - \mathcal{S}_k(\mathcal{S}_k(u_{k-1}^-, q_k^-), \epsilon_\alpha)$$

for $u_{k-1}^- = \mathcal{S}_{k-1}(q_{k-1}^-) \circ \dots \mathcal{S}_1(u, q_1^-)$. Since

$$\begin{aligned} G(u_{k-1}^-, q_k^- = q, \epsilon_\alpha = -q) &= G(u_{k-1}^-, q_k^- = q, \epsilon_\alpha = 0) \\ &= G(u_{k-1}^-, q_k^- = 0, \epsilon_\alpha = q) = 0, \end{aligned}$$

consequently we obtain

$$\frac{\partial^2 G}{\partial \epsilon_\alpha \partial q_k^-}(u_{k-1}^-, q_k^- = 0, \epsilon_\alpha = 0) = 0.$$

Thus

$$G(u_{k-1}^-, q_k^-, \epsilon_\alpha) = \mathcal{O}(1)(|\epsilon_\alpha| \cdot |q_k^-|^2 + \epsilon_\alpha^2),$$

which in view of (8.1) and (8.2) implies (i). ∎

We now prove a generalization of the observation in section 3.

LEMMA 8.2. *Assume that the L^1 stability condition* (3.1) *is satisfied. There exists a constant $\gamma > 0$, depending only on the weights $\{w_i(\theta)\}_{i \neq k}$ such that the following holds. Let $u, v, \epsilon_\alpha, \{q_i^\pm\}$ be as in Lemma* 8.1 *with all $\{q_i^-\}_{i \leq k}$ equal to 0 and $\epsilon_\alpha \geq 0$. By $w_i^\pm$ we denote the weight associated to the shock wave $q_i^\pm$, computed at its left state, by means of* (2.1). *Then*

$$\begin{aligned} &\sum_{i>k} \left[w_i^+ |q_i^+| \cdot (\lambda_i^+ - \lambda_k(v)) - w_i^- |q_i^-| \cdot (\lambda_i^- - \lambda_k(v)) \right] \\ &\qquad + \sum_{i<k} w_i^+ |q_i^+| \cdot |\lambda_i^+ - \lambda_k(v)| \leq -\gamma \epsilon_\alpha \cdot \sum_{i>k} |q_i^-|. \end{aligned} \tag{8.3}$$

Analogously, if

$$\begin{aligned} &\mathcal{S}_k(q_k^+) \circ \mathcal{S}_{k-1}(q_{k-1}^-) \circ \dots \mathcal{S}_1(u, q_1^-) \\ &\qquad = \mathcal{S}_n(q_n^+) \circ \dots \mathcal{S}_{k+1}(q_{k+1}^+) \circ \mathcal{S}_{k-1}(q_{k-1}^+) \circ \dots \mathcal{S}_1(q_1^+) \circ \mathcal{S}_k(u, \epsilon_\alpha) \end{aligned}$$

for some $u \in \Omega$ and $\{q_i^-\}_{i<k}$ with $\epsilon_\alpha \geq 0$, then

$$\begin{aligned} &\sum_{i<k} \left[w_i^+ |q_i^+| \cdot (\lambda_i^+ - \lambda_k(u)) - w_i^- |q_i^-| \cdot |\lambda_i^- - \lambda_k(u)| \right] \\ &\qquad + \sum_{i>k} w_i^+ |q_i^+| \cdot |\lambda_i^+ - \lambda_k(u)| \leq -\gamma \epsilon_\alpha \cdot \sum_{i<k} |q_i^-|. \end{aligned}$$

Proof. We prove only the formula (8.3); the second part of the lemma follows by the same method. By standard interaction estimates [Sm] we have

$$\begin{aligned}\forall\ i>k \qquad |q_i^+|-|q_i^-| \leq & \sum_{s>k,\ s\neq i}\epsilon_\alpha|q_s^-|\cdot|\langle l_i,[r_k,r_s]\rangle(v)| \\ & +\ \epsilon_\alpha|q_i^-|\cdot\langle l_i,[r_k,r_i]\rangle(v) \\ & +\ \mathcal{O}(1)\epsilon_\alpha\left(\sum_{s>k}|q_s^-|\right)\left(\sum_{s\geq k}|q_s^-|\right),\end{aligned} \tag{8.4}$$

$$\begin{aligned}\forall\ i<k \qquad |q_i^+| \leq & \sum_{s>k}\epsilon_\alpha|q_s^-|\cdot|\langle l_i,[r_k,r_s]\rangle(v)| \\ & +\ \mathcal{O}(1)\epsilon_\alpha\left(\sum_{s>k}|q_s^-|\right)\left(\sum_{s\geq k}|q_s^-|\right).\end{aligned} \tag{8.5}$$

Also we have

$$\forall\ i>k \quad w_i^+-w_i^- = \epsilon_\alpha\cdot w_i'(\lambda_k(v))+\mathcal{O}(1)\left[\epsilon_\alpha\cdot\left(\sum_{s>k}|q_s^-|\right)+\epsilon_\alpha^2\right], \tag{8.6}$$

$$\forall\ i>k \quad \lambda_i^+-\lambda_i^- = \epsilon_\alpha\cdot\langle \mathrm{D}\lambda_i,r_k\rangle(v)+\mathcal{O}(1)\left[\epsilon_\alpha\cdot\left(\sum_{s>k}|q_s^-|\right)+\epsilon_\alpha^2\right]. \tag{8.7}$$

Thus

$$\begin{aligned}&\sum_{i>k}|q_i^-|(w_i^+-w_i^-)|\lambda_i^+-\lambda_k(v)|+\sum_{i>k}|q_i^-|w_i^-(\lambda_i^+-\lambda_i^-) \\ &\quad\leq\sum_{i>k}w_i'(\lambda_k(v))\cdot\epsilon_\alpha|q_i^-|\cdot|\lambda_i(v)-\lambda_k(v)|+\sum_{i>k}w_i(v)\epsilon_\alpha|q_i^-|\cdot\langle \mathrm{D}\lambda_i,r_k\rangle(v) \\ &\quad+\ \mathcal{O}(1)\left[\epsilon_\alpha^2\cdot\left(\sum_{s>k}|q_s^-|\right)+\epsilon_\alpha\cdot\left(\sum_{s>k}|q_s^-|\right)^2\right].\end{aligned} \tag{8.8}$$

Moreover, by (8.4) one arrives at

$$\begin{aligned}&\sum_{i>k}w_i^+\cdot\left(|q_i^+|-|q_i^-|\right)|\lambda_i^+-\lambda_k(v)| \\ &\quad\leq\sum_{i>k}\epsilon_\alpha|q_i^-|\cdot\Bigg(w_i(v)|\lambda_i(v)-\lambda_k(v)|\cdot\langle l_i,[r_k,r_i]\rangle(v) \\ &\qquad\qquad+\sum_{s>k,\ s\neq i}w_s(v)|\lambda_s(v)-\lambda_k(v)|\cdot|\langle l_s,[r_i,r_k]\rangle(v)|\Bigg) \\ &\quad+\ \mathcal{O}(1)\epsilon_\alpha\left(\sum_{s>k}|q_s^-|\right)\left(\epsilon_\alpha+\sum_{s>k}|q_s^-|\right).\end{aligned} \tag{8.9}$$

Adding (8.8) and (8.9), and noting (8.5), we see that the left-hand side of (8.3) can be estimated as follows:

$$
\begin{aligned}
&\epsilon_\alpha \cdot \sum_{i>k} |q_i^-| \cdot |\lambda_i(v) - \lambda_k(v)| \cdot \Bigg[w_i'(\lambda_k(v)) + w_i(v) \cdot \frac{\langle \mathrm{D}\lambda_i, r_k\rangle(v)}{|\lambda_i(v) - \lambda_k(v)|} \\
&\qquad + w_i(u) \cdot \langle l_i, [r_k, r_i]\rangle(v) + \sum_{i\neq k,i} w_s(v) \frac{|\lambda_s(v) - \lambda_k(v)|}{|\lambda_i(v) - \lambda_k(v)|} \cdot |\langle l_s, [r_i, r_k]\rangle(v)| \Bigg] \qquad (8.10) \\
&\quad + \mathcal{O}(1)\epsilon_\alpha \left(\sum_{s>k} |q_s^-| \right) \left(\epsilon_\alpha + \sum_{s>k} |q_s^-| \right).
\end{aligned}
$$

Applying the inequality (3.1) with $\theta \in (-c, \Theta + c)$ such that $\lambda_k(v) = \lambda_k(\mathcal{R}_k(\theta))$ and by a compactness argument, we obtain that (8.10) is bounded by the quantity in the right-hand side of (8.3). The proof is done. ∎

LEMMA 8.3. *Assume that the L^1 stability condition* (3.1) *is satisfied. Let u, v, ϵ_α, $\{q_i^\pm\}_{i=1}^n$ be as in Lemma* 8.1, *with $\epsilon_\alpha \geq 0$. Then*

$$
\begin{aligned}
&\sum_{i\neq k} \left[w_i^+ |q_i^+| (\lambda_i^+ - \lambda_k(v)) - w_i^- |q_i^-| (\lambda_i^- - \lambda_k(v)) \right] \\
&\qquad\qquad \leq -\gamma_1 \cdot \epsilon_\alpha \cdot \left(\sum_{s>k} |q_s^-| \right) + \mathcal{O}(1) \left[\epsilon_\alpha \cdot |q_k^-|^2 + \epsilon_\alpha^2 \right]
\end{aligned}
$$

for some constant $\gamma_1 > 0$, depending only on weights $\{w_i(\theta)\}_{i=1}^n$ and the uniform system bounds $\mathcal{O}(1)$.

Proof. Let Ξ denote the left-hand side of the desired inequality. We write $\{\tilde{q}_s\}_{s=1}^n$ and $\{\hat{q}_s\}_{s=1}^n$ for the quantities introduced implicitly by

$$
\mathcal{S}_n(\tilde{q}_n) \circ \ldots \mathcal{S}_{k+1}(\tilde{q}_{k+1}) \circ \mathcal{S}_{k-1}(\tilde{q}_{k-1}) \circ \ldots \mathcal{S}_1(\tilde{q}_1) \circ \mathcal{S}_k(u_k, \tilde{q}_k) = \mathcal{S}_k(v, \epsilon_\alpha),
$$

$$
\mathcal{S}_k(v, \epsilon_\alpha) = \mathcal{S}_n(\hat{q}_n) \circ \ldots \mathcal{S}_1(\hat{q}_1) \circ \mathcal{S}_k(u_{k-1}, q_{k-1}^- + \tilde{q}_k),
$$

$$
u_{k-1} = \mathcal{S}_{k-1}(q_{k-1}^-) \circ \ldots \mathcal{S}_1(u, q_1^-) \quad \text{and} \quad u_k = \mathcal{S}_k(u_{k-1}, q_k^-).
$$

By $\tilde{w}_s, \hat{w}_s$ and $\tilde{\lambda}_s, \hat{\lambda}_s$, we naturally denote weights and speeds corresponding to the waves $\tilde{q}_s$ and $\hat{q}_s$. We then have

$$
\begin{aligned}
\Xi = &\Bigg\{ \sum_{i>k} \left[\tilde{w}_i |\tilde{q}_i| (\tilde{\lambda}_i - \lambda_k(v)) - w_i^- |q_i^-| (\lambda_i^- - \lambda_k(v)) \right] \\
&\qquad\qquad - \sum_{i<k} \tilde{w}_i |\tilde{q}_i| (\tilde{\lambda}_i - \lambda_k(v)) \Bigg\} \qquad (8.11) \\
&+ \sum_{i\neq k} w_i^+ |q_i^+| (\lambda_i^+ - \lambda_k(v)) - \sum_{i<k} w_i^- |q_i^-| (\lambda_i^- - \lambda_k(v)) \\
&- \sum_{i>k} \tilde{w}_i |\tilde{q}_i| (\tilde{\lambda}_i - \lambda_k(v)) + \sum_{i<k} \tilde{w}_i |\tilde{q}_i| (\tilde{\lambda}_i - \lambda_k(v)).
\end{aligned}
$$

Observe that $\tilde{q}_k = \epsilon_\alpha + \mathcal{O}(1)\epsilon_\alpha \cdot \left(\sum_{s>k} |q_s^-|\right)$. Using the same arguments as in the proof of Lemma 8.1, we arrive at

$$\left(\sum_{i\neq k} |\tilde{q}_i - \hat{q}_i|\right) + |\hat{q}_k| \leq \mathcal{O}(1) \cdot \left[\epsilon_\alpha |q_k^-|^2 + \epsilon_\alpha^2\right]. \tag{8.12}$$

A similar bound is true for the corresponding differences of $\hat{\lambda}_i$ and $\tilde{\lambda}_i$, and $\hat{w}_i$ and $\tilde{w}_i$. Estimating the first term in (8.11) in view of Lemma 8.3, we obtain

$$\begin{aligned}
\Xi \leq & -\gamma\epsilon_\alpha \cdot \left(\sum_{s>k} |q_s^-|\right) + \mathcal{O}(1) \cdot \left[\epsilon_\alpha \cdot |q_k^-|^2 + \epsilon_\alpha^2\right] \\
& + \sum_{i>k} \left[w_i^+ |q_i^+| (\lambda_i^+ - \lambda_k(v)) - \hat{w}_i |\hat{q}_i| (\hat{\lambda}_i - \lambda_k(v))\right] \\
& + \sum_{i<k} \left[w_i^+ |q_i^+| (\lambda_i^+ - \lambda_k(v)) - w_i^- |q_i^-| (\lambda_i^- - \lambda_k(v))\right] \\
& + \sum_{i<k} \hat{w}_i |\hat{q}_i| (\hat{\lambda}_i - \lambda_k(v)).
\end{aligned} \tag{8.13}$$

Now, by standard interaction estimates [L], we have

$$\begin{aligned}
& \sum_{i<k} \left|q_i^+ - (q_i^- + \hat{q}_i)\right| + \sum_{i>k} \left|q_i^+ - \hat{q}_i\right| \\
& = \mathcal{O}(1) \cdot \left[\left(\sum_{i<k} |\hat{q}_i|\right) \cdot \left(\sum_{i<k} |q_i^-|\right) + \left|q_k^- + \epsilon_\alpha\right| \cdot \left(\sum_{i\leq k} |\hat{q}_i|\right)\right] \\
& = \mathcal{O}(1)\epsilon_\alpha \cdot \left[\left(\sum_{s>k} |q_s^-|\right) + |q_k^-|^2 + \epsilon_\alpha\right] \cdot \left[\left(\sum_{s<k} |q_s^-|\right) + |q_k^-| + \epsilon_\alpha\right] \\
& = \mathcal{O}(1) \cdot \epsilon_\alpha \cdot \left(\sum_{s>k} |q_s^-|\right) \left[\left(\sum_{s<k} |q_s^-|\right) + |q_k^-|\right] + \mathcal{O}(1) \left[\epsilon_\alpha |q_k^-|^2 + \epsilon_\alpha^2\right].
\end{aligned} \tag{8.14}$$

Noting that $\left(\sum_{s<k} |q_s^-|\right) + |q_k^-| = \mathcal{O}(1) \cdot (\epsilon_0 + \nu)$, we obtain

$$\begin{aligned}
& \sum_{i>k} \left[w_i^+ |q_i^+| (\lambda_i^+ - \lambda_k(v)) - \hat{w}_i |\hat{q}_i| (\hat{\lambda}_i - \lambda_k(v))\right] \\
& \qquad = \mathcal{O}(1) \cdot (\epsilon_0 + \nu)\epsilon_\alpha \cdot \left(\sum_{s<k} |q_s^-|\right) + \mathcal{O}(1) \cdot \epsilon_\alpha |q_k^-|^2 + \mathcal{O}(1)\epsilon_\alpha^2.
\end{aligned}$$

In view of (8.14), exactly the same bound as above is valid for the terms:

$$\sum_{i<k} \left[w_i^+ |q_i^+| (\lambda_i^+ - \lambda_k(v)) + \hat{w}_i |\hat{q}_i| (\hat{\lambda}_i - \lambda_k(v)) - w_i^- |q_i^-| (\lambda_i^- - \lambda_k(v))\right].$$

Hence by (8.13) the lemma follows, if only the constant ϵ_0 and ν are small enough. ∎

9. A sufficient condition for admissibility of initial data: A proof of Lemma 4.6.

LEMMA 4.6. *Let $\bar{u} \in \mathrm{cl}\ \mathcal{E}_{c,\delta}$ for some sufficiently small $c, \delta > 0$, as in Theorem 1. Then $\bar{u} \in \bar{\mathcal{D}}_{\epsilon_0}$, defined in (4.5), for some $\epsilon_0 = \epsilon_0(\delta)$ and $\lim_{\delta \to 0} \epsilon_0(\delta) = 0$.*

Proof. 1. Without loss of generality we may assume that $\bar{u}$ is piecewise constant, consecutively attaining N states $u_l = u^0, u^1 \dots u^N = u_r$ in $\mathbf{R}^n$, that for each $\alpha : 1 \dots N-1$ we have $||u^{\alpha+1} - u^\alpha|| < \delta$, and that (i), (ii), (iii) as in Theorem 1 are satisfied. For $\alpha : 0 \dots N-1$ and $i : 1 \dots n$ define

$$\gamma_\alpha^i = \langle l_i(u^\alpha), u^{\alpha+1} - u^\alpha \rangle.$$

Note that the self-similar solution of each Riemann problem $(u^\alpha, u^{\alpha+1})$ is composed of n waves having corresponding strengths $\epsilon_\alpha^1 \dots \epsilon_\alpha^n$ with the following obvious estimate:

$$\sum_{\alpha=0}^{N-1} \sum_{i=1}^{n} |\gamma_\alpha^i - \epsilon_\alpha^i| \le \sum_{\alpha=0}^{N-1} ||u^{\alpha+1} - u^\alpha||^2 < \delta \cdot \sum_{\alpha=0}^{N-1} ||u^{\alpha+1} - u^\alpha|| = \mathcal{O}(1)\delta.$$

To simplify the presentation we will assume that $||r_k(u)|| = 1$ for all $u \in \Omega$. In order to prove the lemma it is thus enough to show that

$$\left| \left(\sum_{\alpha=0}^{N-1} (\gamma_\alpha^k)^+ \right) - |\mathcal{R}_k| \right| < \epsilon_0 \quad \text{and} \quad \sum_{\alpha=0}^{N-1} \left((\gamma_\alpha^k)^- + \sum_{i \neq k} |\gamma_\alpha^i| \right) < \epsilon_0, \tag{9.1}$$

where $|\mathcal{R}_k|$ denotes the arc-length of the curve $\mathcal{R}_k(\theta),\ \theta \in [0, \Theta]$.

2. Fix a small constant $c > 0$ and divide the set of discontinuities in $\bar{u}$ into three subsets:

$$G = \left\{ \alpha : 0 \dots N-1, \quad \left\| \frac{u^{\alpha+1} - u^\alpha}{||u^{\alpha+1} - u^\alpha||} - r_k(u^\alpha) \right\| < c \right\},$$

$$B' = \left\{ \alpha : 0 \dots N-1, \quad \left\| \frac{u^{\alpha+1} - u^\alpha}{||u^{\alpha+1} - u^\alpha||} + r_k(u^\alpha) \right\| < c \right\},$$

$$B = \{0 \dots N-1\} \setminus (G \cup B').$$

It follows that for all $\alpha \in G$

$$\left| \frac{\gamma_\alpha^k}{||u^{\alpha+1} - u^\alpha||} - 1 \right| + \sum_{i \neq k} \left| \frac{\gamma_\alpha^i}{||u^{\alpha+1} - u^\alpha||} \right| = \mathcal{O}(1)c.$$

Thus

$$\left| \sum_{\alpha \in G} (\gamma_\alpha^k)^+ - \sum_{\alpha \in G} ||u^{\alpha+1} - u^\alpha|| \right| + \sum_{\alpha \in G} \left((\gamma_\alpha^k)^- + \sum_{i \neq k} |\gamma_\alpha^i| \right) = \mathcal{O}(1) \cdot c|\mathcal{R}_k|. \tag{9.2}$$

On the other hand, for all $\alpha \in B \cup B'$

$$\left| \frac{\gamma_\alpha^k}{||u^{\alpha+1} - u^\alpha||} - 1 \right| + \sum_{i \neq k} \left| \frac{\gamma_\alpha^i}{||u^{\alpha+1} - u^\alpha||} \right| = \mathcal{O}(1). \tag{9.3}$$

3. Let $\mathcal{P}:\Omega_\delta \longrightarrow \mathcal{R}_k$ be the orthogonal projection of Ω_δ onto $\mathcal{R}_k$. Note that if $u=\mathcal{R}_k(\theta)$ for some $\theta\in[0,\Theta]$, then $D\mathcal{P}(u)\cdot v=\langle v, r_k(u)\rangle\cdot r_k(u)$. We have

$$||u^{\alpha+1}-u^\alpha||-||\mathcal{P}(u^{\alpha+1})-\mathcal{P}(u^\alpha)||\geq \mathcal{O}(1)\delta\cdot||u^{\alpha+1}-u^\alpha||. \tag{9.4}$$

Also, for each $\alpha\in B$, the cosine of the angle between the vectors $u^{\alpha+1}-u^\alpha$ and $r_k(u^\alpha)$ satisfies

$$\left|\cos\angle\left(u^{\alpha+1}-u^\alpha, r_k(u^\alpha)\right)\right|\leq 1-c^2/2.$$

Thus, for $\alpha\in B$ we have

$$\begin{aligned}||\mathcal{P}(u^{\alpha+1})-\mathcal{P}(u^\alpha)||&\leq |\langle u^{\alpha+1}-u^\alpha, r_k(u^\alpha)\rangle|+\mathcal{O}(1)\cdot\delta||u^{\alpha+1}-u^\alpha||\\&\leq\left(1-\frac{c^2}{2}+\mathcal{O}(1)\delta\right)\cdot||u^{\alpha+1}-u^\alpha||,\end{aligned}$$

and, consequently,

$$||u^{\alpha+1}-u^\alpha||-||\mathcal{P}(u^{\alpha+1})-\mathcal{P}(u^\alpha)||\geq\left[\frac{c^2}{2}+\mathcal{O}(1)\delta\right]\cdot||u^{\alpha+1}-u^\alpha||. \tag{9.5}$$

By (9.4) and (9.5) we receive

$$\begin{aligned}&\sum_{\alpha=0}^{N-1}\left(||u^{\alpha+1}-u^\alpha||-||\mathcal{P}(u^{\alpha+1})-\mathcal{P}(u^\alpha)||\right)\\&\qquad\geq\frac{c^2}{2}\sum_{\alpha\in B}||u^{\alpha+1}-u^\alpha||+\mathcal{O}(1)\cdot\delta\sum_{\alpha}^{N-1}||u^{\alpha+1}-u^\alpha||.\end{aligned} \tag{9.6}$$

4. On the other hand, with $c\ll 1$ we have that $||\mathcal{P}(u^{\alpha+1})-\mathcal{P}(u^\alpha)||\geq 1/2\cdot||u^{\alpha+1}-u^\alpha||$ for all $\alpha\in G\cup B'$. Hence,

$$\begin{aligned}&\sum_{\alpha=0}^{N-1}||u^{\alpha+1}-u^\alpha||-\sum_{\alpha=0}^{N-1}||\mathcal{P}(u^{\alpha+1})-\mathcal{P}(u^\alpha)||\\&\qquad\leq|\mathcal{R}_k|+\delta-\left(|\mathcal{R}_k|+\mathcal{O}(1)\delta-2\cdot\sum_{\alpha\in B'}||\mathcal{P}(u^{\alpha+1})-\mathcal{P}(u^\alpha)||\right)\\&\qquad\leq\mathcal{O}(1)\delta-\sum_{\alpha\in B'}||u^{\alpha+1}-u^\alpha||.\end{aligned}$$

In view of (9.6) we thus obtain

$$c^2\cdot\sum_{\alpha\in B}||u^{\alpha+1}-u^\alpha||=\mathcal{O}(1)\delta. \tag{9.7}$$

The estimates (9.2), (9.3), and (9.7) yield

$$\begin{aligned}\left|\left(\sum_{\alpha=0}^{N-1}(\gamma_\alpha^k)^+\right)-|\mathcal{R}_k|\right|&\leq\left|\left(\sum_{\alpha\in G}(\gamma_\alpha^k)^+\right)-\left(\sum_{\alpha\in G}||u^{\alpha+1}-u^\alpha||\right)\right|\\&\quad+\left|\left(\sum_{\alpha\in B}(\gamma_\alpha^k)^+\right)-\left(\sum_{\alpha\in B}||u^{\alpha+1}-u^\alpha||\right)\right|+\mathcal{O}(1)\delta\\&\leq\mathcal{O}(1)c+\mathcal{O}(1)\sum_{\alpha\in B}||u^{\alpha+1}-u^\alpha||+\mathcal{O}(1)\delta\\&=\mathcal{O}(1)\cdot(c+\delta/c^2+\delta).\end{aligned}$$

Taking $c^2 = \sqrt{\delta}$, we receive the first estimate in (9.1) with $\epsilon_0 = \mathcal{O}(1)\delta^{1/4}$. The second estimate follows in the same manner. □

Acknowledgments. I am grateful to the referees for pointing out a mistake in an earlier version of the paper. I thank Professor Alberto Bressan for suggesting this problem to me. I thank Professor Constantine Dafermos for encouragement and his interest in this work. I also thank my colleagues at the University of Chicago for providing the stimulating atmosphere which allowed me to finish this project.

REFERENCES

[BaJ] P. Baiti and H. K. Jenssen, *On the front tracking algorithm*, J. Math. Anal. Appl., 217 (1998), pp. 395–404.

[BiB] S. Bianchini and A. Bressan, *Vanishing viscosity solutions of nonlinear hyperbolic systems*, Ann. of Math. (2), to appear.

[B] A. Bressan, *Hyperbolic Systems of Conservation Laws. The One-Dimensional Cauchy Problem*, Oxford Lecture Ser. Math. Appl. 20, Oxford University Press, Oxford, UK, 2000.

[BC] A. Bressan and R. M. Colombo, *Unique solutions of* 2×2 *conservation laws with large data*, Indiana Univ. Math. J., 44 (1995), pp. 677–725.

[BCP] A. Bressan, G. Crasta, and B. Piccoli, *Well posedness of the Cauchy problem for* $n \times n$ *systems of conservation laws*, Mem. Amer. Math. Soc., 146 (2000), No. 694.

[BLY] A. Bressan, T. P. Liu, and T. Yang, L^1 *stability estimates for* $n \times n$ *conservation laws*, Arch. Ration. Mech. Anal., 149 (1999), pp. 1–22.

[BM] A. Bressan and A. Marson, *A variational calculus for discontinuous solutions of systems of conservation laws*, Comm. Partial Differential Equations, 20 (1995), pp. 1491–1552.

[D] C. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Grundlehren Math. Wiss. 325, Springer-Verlag, Berlin, 2000.

[G] J. Glimm, *Solutions in the large for nonlinear hyperbolic systems of equations*, Comm. Pure Appl. Math., 18 (1965), pp. 697–715.

[HR] H. Holden and N. H. Risebro, *Front Tracking for Hyperbolic Conservation Laws*, Appl. Math. Sci. 152, Springer-Verlag, New York, 2002.

[L] P. Lax, *Hyperbolic systems of conservation laws* II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.

[Le1] M. Lewicka, L^1 *stability of patterns of non-interacting large shock waves*, Indiana Univ. Math. J., 49 (2000), pp. 1515–1537.

[Le2] M. Lewicka, *Stability conditions for patterns of noninteracting large shock waves*, SIAM J. Math. Anal., 32 (2001), pp. 1094–1116.

[Le3] M. Lewicka, *Stability conditions for strong rarefaction waves*, SIAM J. Math. Anal., 36 (2004), pp. 1346–1369.

[LeT] M. Lewicka and K. Trivisa, *On the* L^1 *well posedness of systems of conservation laws near solutions containing two large shocks*, J. Differential Equations, 179 (2002), pp. 133–177.

[LY] T. P. Liu and T. Yang, L_1 *stability of weak solutions for* 2×2 *systems of hyperbolic conservation laws*, J. Amer. Math. Soc., 12 (1999), pp. 729–774.

[S] D. Serre, *Systems of Conservation Laws*, Cambridge University Press, Cambridge, UK, 1999.

[Scho] S. Schochet, *Sufficient conditions for local existence via Glimm's scheme for large BV data*, J. Differential Equations, 89 (1991), pp. 317–354.

[Sm] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1994.

SIAM J. MATH. ANAL.
Vol. 36, No. 5, pp. 1400–1406

BOUNDS FOR THE STEADY-STATE SEL'KOV MODEL FOR ARBITRARY p IN ANY NUMBER OF DIMENSIONS*

GARY M. LIEBERMAN†

Abstract. The Sel'kov model is a system of two differential equations which describe various complex biological and chemical systems. In this system there is an exponent p which must be allowed to be an arbitrary number greater than one according to the underlying model but is usually restricted in mathematical analyses. We show that the restriction is not necessary by proving some a priori estimates for all solutions of the system. The techniques for proof include some maximum principle arguments and the weak Harnack inequality. In fact our techniques apply to a much more general class of problems, including the Brusselator model.

Key words. elliptic equations, a priori estimates, Sel'kov model

AMS subject classifications. 35J55, 35B45, 35B65, 35Q80, 92C15, 92C40

DOI. 10.1137/S003614100343651X

1. Introduction. In 1968, Sel'kov [13] introduced a model for glycolysis that has become a standard for various complex biological, chemical, and biochemical systems. The steady-state form of this model is the following boundary value problem: Find nonnegative functions u and v such that

$$\theta\Delta u = \lambda(uv^p - 1) \text{ in } \Omega, \tag{1.1a}$$

$$\Delta v = \lambda(v - uv^p) \text{ in } \Omega, \tag{1.1b}$$

$$\frac{\partial u}{\partial \nu} = \frac{\partial v}{\partial \nu} = 0 \text{ on } \partial\Omega \tag{1.1c}$$

for positive parameters p, θ, and λ. Here, Ω is a bounded domain in $\mathbb{R}^n$ ($n \geq 1$) and $\partial/\partial\nu$ denotes the inner normal derivative. We refer the reader to [2] and [14] for a more detailed bibliography and discussion of the significance of this system.

This problem has the obvious constant solution $u \equiv 1$, $v \equiv 1$ (and this is the only constant solution), but Eilbeck and Furter [3] used numerical bifurcation methods to show that the one-dimensional problems has nonconstant solutions for suitable ranges of the parameters. More recently, Davidson and Rynne [2] and Wang [14] proved corresponding results in two and three dimensions. (In fact, their arguments apply to any number of dimensions, but their results are only stated for two and three dimensions because of the physical applications.) A key step in their arguments is an a priori estimate for all such solutions, but, when $n \geq 3$, they require an upper bound on p, namely, $p < n/(n-2)$. We shall show that no restriction on p is needed for this estimate and that a simpler approach can be used, which is independent of dimension. Our interest came from the theoretical side of the problem, but there is a more practical reason to investigate such a result: In Sel'kov's original paper, this parameter should be allowed to take on arbitrary values greater than one for physical reasons. It should also be noted that the proofs in [2] and [14] of their bifurcation and existence results along with our more general estimates give their results without the

*Received by the editors October 23, 2003; accepted for publication (in revised form) August 13, 2004; published electronically March 25, 2005.
http://www.siam.org/journals/sima/36-5/43651.html
†Department of Mathematics, Iowa State University, Ames, IA 50011 (lieb@iastate.edu).

restriction on p. For example, we have the following result, in which μ_i for $i = 1, 2, \ldots$ denote the positive eigenvalues of the Laplacian operator with Neumann conditions.

THEOREM 1.1. *Let λ, θ, and p be positive parameters, and suppose there is at least one index i such that*

$$\lambda(\lambda + \mu_i) < \theta\mu_i[\lambda(p-1) - \mu_i], \tag{1.2}$$

but there are no indices j such that

$$\lambda(\lambda + \mu_j) = \theta\mu_j[\lambda(p-1) - \mu_j]. \tag{1.3}$$

If the sum of the dimensions of the eigenspaces corresponding to all eigenvalues μ_i satisfying (1.2) *is odd, then there is a nonconstant solution to* (1.1).

The proof of this theorem is the same as for [14, Theorem 6.2] once we have the appropriate estimate. (In [14], it was also assumed that $p < n/(n-2)$ if $n \geq 3$, but this assumption was used only to derive the a priori estimate.)

The key new element in our approach is the use of a weak Harnack inequality, described in section 2. We prove our estimate in section 3 with some generalizations in section 4.

2. Some general results for elliptic equations. Our new element is the following result which is well known as a local result for weak supersolutions of linear elliptic equations (see, for example, [5, Theorem 8.18]).

LEMMA 2.1. *Let Ω be a bounded Lipschitz domain in $\mathbb{R}^n$. Let Λ be a nonnegative constant and suppose that $v \in W^{1,2}$ is a nonnegative weak solution of the inequalities*

$$\Delta v - \Lambda v \leq 0 \text{ in } \Omega, \quad \frac{\partial v}{\partial \nu} \leq 0 \text{ on } \partial\Omega. \tag{2.1}$$

Then, for any $q \in [1, n/(n-2))$, there is a constant C_0, determined only by q, Λ, and Ω, such that

$$\|v\|_{q;\Omega} \leq C_0 \inf_{\Omega} v. \tag{2.2}$$

Proof. The proof of [5, Theorem 8.18] is easily adapted to handle the boundary condition by using the argument in [10, Theorem 6.40]. (See also [7] for further details.) □

Similarly, we have the following Harnack inequality for weak solutions, which is an analogue of [5, Theorem 8.16]. (See also [11, Lemma 4.3] for a proof in smooth domains.)

LEMMA 2.2. *Let Ω be a bounded Lipschitz domain in $\mathbb{R}^n$. Let $c \in L^q(\Omega)$ for some $q > n/2$, and suppose that $v \in W^{1,2}$ is a nonnegative weak solution of the boundary value problem*

$$\Delta v + cv = 0 \text{ in } \Omega, \quad \frac{\partial v}{\partial \nu} = 0 \text{ on } \partial\Omega. \tag{2.3}$$

Then there is a constant C_1, determined only by $\|c\|_q$, q, and Ω such that

$$\sup_{\Omega} v \leq C_1 \inf_{\Omega} v. \tag{2.4}$$

We also have the following analogue of [12, Proposition 2.2].

LEMMA 2.3. *Let Ω be a bounded Lipschitz domain and let $g \in C(\overline{\Omega} \times \mathbb{R})$. If $w \in W^{1,2}$ is a weak solution of the inequalities*

$$\Delta w + g(x,w) \geq 0 \text{ in } \Omega, \quad \frac{\partial w}{\partial \nu} \geq 0 \text{ on } \partial\Omega, \tag{2.5}$$

and if there is a constant K such that $g(x,z) < 0$ for $z > K$, then $w \leq K$ a.e. in Ω.

Proof. Use the test function $(w-K)^+$ to see that

$$\int_{\{w>K\}} |Dw|^2\,dx \leq \int_{\{w>K\}} g(x,z)(w-K)\,dx.$$

The integral on the left-hand side of this inequality is nonnegative, while the integrand of the integral on the right-hand side is negative, so $\{w > K\}$ must have zero measure, as required. ∎

3. The main a priori estimates. We now look at nonnegative solutions of the system (1.1).

Our a priori estimate has the following simple form.

THEOREM 3.1. *Let P, Θ, and Λ be positive constants, and suppose that $0 < p \leq P$, $0 < \theta \leq \Theta$, and $0 < \lambda \leq \Lambda$. Write C_2 for the value of C_0 corresponding to $q = 1$ and set $\varepsilon = |\Omega|/C_2$. Then any nonnegative $W^{1,2}$ solution to (1.1) satisfies the inequalities*

$$\varepsilon \leq v \leq \Theta\varepsilon^{-P} + 1, \tag{3.1a}$$

$$\frac{1}{(\Theta\varepsilon^{-P}+1)^P} \leq u \leq \varepsilon^{-P}. \tag{3.1b}$$

Proof. We begin by noting (see equations (2.1) and (2.2) in [2]) that

$$\int_\Omega uv^p\,dx = \int_\Omega v\,dx = |\Omega|. \tag{3.2}$$

Next, v satisfies $\Delta v \leq \lambda v \leq \Lambda v$ in Ω, so Lemma 2.1 implies that

$$\int_\Omega v\,dx \leq C_2 \inf_\Omega v,$$

so $v \geq \varepsilon$ in Ω. (Note that $\inf v \leq 1$, so $\varepsilon \leq 1$.)

We see that $v^p \geq \varepsilon^P$ in Ω, so $\Delta u \geq (\lambda/\theta)[\varepsilon^P u - 1]$ in Ω. It follows from Lemma 2.3 that $u \leq \varepsilon^{-P}$ in Ω.

Next, we set $w = \theta u + v$. Then $\Delta w = \lambda(v-1)$, and $v - 1 > 0$ wherever $w > \theta\varepsilon^{-P} + 1$, so Lemma 2.3 now implies that $w \leq \theta\varepsilon^{-P} + 1$. Since $u \geq 0$, it follows that $v \leq \theta\varepsilon^{-P} + 1 \leq \Theta\varepsilon^{-P} + 1$.

Finally, we note that, at a minimum of u, we must have $uv^p - 1 \geq 0$, so $\inf u \geq 1/(\sup v)^p \geq 1/(\sup v)^P$. Hence, $u \geq (\Theta\varepsilon^{-P}+1)^{-P}$. ∎

When $p = 0$ (and $\lambda > 0$ and $\theta > 0$), it follows from (1.1a) and (1.1c) that $u \equiv 1$ and then (1.1b) and (1.1c) imply also that $v \equiv 1$. Similarly, if $\theta = 0$ (and $\lambda > 0$), we can rewrite (1.1a) as $uv^p - 1 = 0$, so (1.1b) becomes $\Delta v = \lambda[v-1]$, so $v \equiv 1$ and hence $u \equiv 1$ also. Hence, the estimate (3.1) is uniform over the ranges $0 \leq \theta \leq \Theta$ and $0 \leq p \leq P$. However, the restriction $\lambda > 0$ is clearly necessary, because, for $\lambda = 0$, any positive constants are solutions of the resulting system (unless we impose the additional assumptions (3.2), in which case the only solution is $u \equiv v \equiv 1$). On

the other hand, for $\lambda > 0$, the only constant solution is $u \equiv v \equiv 1$. The existence of nonconstant solutions is the focus of [2, section 4] and [14].

In fact, Wang [14, Theorem 3.1] provides estimates depending only on a lower bound for θ if $P < (n+2)/(2(n-2))$. Our next step is to improve this result by showing that the estimates are independent of θ for the larger range $P < n/(n-2)$.

THEOREM 3.2. *If $P < n/(n-2)$, then there is a constant C_3, determined only by P, Λ, and Ω such that $v \le C_3$ and $u \ge C_3$ in Ω.*

Proof. Without loss of generality, we may assume that $P > 1$. From the proof of Theorem 3.1, we need only to prove an upper bound for v which is independent of θ using the lower bound for v and the upper bound for u (which are already independent of θ). To this end, we set $c = 1 - uv^{p-1}$. Because $\inf v \le 1$, we have

$$|c| \le 1 + \sup u (\inf v)^{p-P} v^{P-1} \le 1 + \varepsilon^{-2P} v^{P-1}.$$

In addition, there is a constant $q > n/2$ (determined only by P and n) such that $(P-1)q < n/(n-2)$. Now Lemma 2.1 provides a constant C_4 (determined only by P, Λ, and Ω) such that

$$\|v\|_{(P-1)q} \le C_4 \inf v.$$

Therefore $c \in L^q$ and we have an estimate for $\|c\|_q$. It follows that v satisfies the hypotheses of Lemma 2.2, which then provides the required upper bound for v. ∎

In addition, higher regularity of u and v is an immediate consequence of estimates for solutions of linear equations. For example, [6, Theorem X.2.1] (see also [10, Theorem 6.44]) implies Hölder estimates for u and v. From this regularity, it follows that u and v have Hölder continuous kth-order derivatives ($k \ge 1$) provided $\partial\Omega$ has the same smoothness. (For the case $k = 1$, see [6, Theorem X.2.1] or [8, Theorem 5.1] and for $k = 2$, see [5, Theorem 6.20].)

4. Generalizations. It should be noted that the method here applies to a much broader class of problems than those explicitly described. Here, we suggest a few possibilities.

The first way in which we generalize our hypotheses is to emphasize which properties of the right-hand sides of our differential equations were actually used.

THEOREM 4.1. *Let f and g be continuous functions defined on $\Omega \times \mathbb{R}^2$, and suppose $u \in W^{1,2}$ and $v \in W^{1,2}$ are nonnegative solutions of the system*

$$\Delta u = f(x,u,v), \qquad \Delta v = g(x,u,v) \text{ in } \Omega \tag{4.1}$$

with boundary condition (1.1c). *Suppose also that f and g satisfy the following conditions:*

1. *There are constants α_1 and β_1 and a function $\gamma \in L^1(\Omega)$ such that*

$$\alpha f(x,z_1,z_2) + \beta g(x,z_1,z_2) + \gamma(x) \le z_2 \tag{4.2}$$

for all $(x,z_1,z_2) \in \Omega \times \mathbb{R}^2$ and

$$\int_\Omega \gamma(x)\,dx > 0. \tag{4.3}$$

2. *There is a positive constant δ such that $g(x,z_1,z_2) \le \delta z_2$.*

3. *There is a function* $M\colon (0,\infty) \to (0,\infty)$ *such that, for any* $\varepsilon > 0$, *we have* $f(x, z_1, z_2) > 0$ *for* $z_2 \geq \varepsilon$ *and* $z_1 > M(\varepsilon)$.
4. *There are a constant* $\alpha_2 \geq 0$ *and a function* $N\colon (0,\infty) \to (0,\infty)$ *such that, for any* $\varepsilon > 0$, *we have* $\alpha_2 f(x, z_1, z_2) + g(x, z_1, z_2) > 0$ *for* $z_1 \leq \varepsilon$ *and* $z_2 > N(\varepsilon)$.
5. *There is a function* $M_1\colon (0,\infty) \to (0,\infty)$ *such that, for any* $\varepsilon > 0$, *we have* $f(x, z_1, z_2) < 0$ *for* $z_2 \leq \varepsilon$ *and* $z_1 < M_1(\varepsilon)$.

Then there are positive constants K_1 *and* K_2 *determined only by* α_2, $\int_\Omega \gamma\,dx$, δ, Ω, M, M_1, *and* N *such that* $K_1 \leq u, v \leq K_2$ *in* Ω.

In fact, this form of our estimate applies to several related models. First, we consider the two-variable system first used by Sel'kov [13, equation (5)]. In our notation, we have

$$f(x,u,v) = \frac{\lambda}{\theta^2}\left[\frac{uv^p}{1+v^p+uv^p} - \nu_1\right], \quad g(x,u,v) = \lambda\left[\kappa v - \frac{uv^p}{1+v^p+uv^p}\right]$$

for positive constants λ, θ, κ, ν_1, and p. In [13], the parameter ν_1 is assumed small. Here, we assume only that it is less than one. Noting that

$$\frac{\theta^2 f}{\lambda\kappa} + \frac{g}{\kappa\lambda} = -\frac{\nu_1}{\kappa} + v, \tag{4.4}$$

we see that condition 1 holds with $\alpha = \theta^2/(\lambda\kappa)$, $\beta = 1/(\kappa\lambda)$, and $\gamma = \nu_1/\kappa$. We also have condition 2 with $\delta = \lambda\kappa$. For condition 3, we note that

$$\frac{uv^p}{1+v^p+uv^p} = \frac{u}{v^{-p}+1+u} \geq \frac{u}{\varepsilon^{-p}+1+u},$$

and this fraction can be made larger than ν_1 by taking u sufficiently large. From (4.4), we infer condition 4 with $N(\varepsilon) \equiv \nu_1/\kappa$, and condition 5 follows with $M_1(\varepsilon) = \nu_1/\varepsilon^p$ once we note that

$$f(x,u,v) \leq \frac{\lambda}{\theta^2}[uv^p - \nu_1].$$

Similarly, we can handle the Brusselator model (see [4], [1], etc.), in which

$$f(x,u,v) = \frac{\lambda}{\theta^2}[v^2u - Bv], \quad g(x,u,v) = \lambda[-A + (B+1)v - v^2u]$$

for positive constants A and B. (Note that we have relabeled variables compared with [4].) In this case, condition 1 holds with $\alpha = \theta^2/\lambda$, $\beta = 1/\lambda$, and $\gamma = A$. Condition 2 holds with $\delta = \lambda(B+1)$, and condition 3 holds with $M(\varepsilon) = B/\varepsilon$. Condition 4 holds with $\alpha_2 = \theta^2$ and $M_1(\varepsilon) \equiv A$. Finally, we have condition 5 with $M_1(s) = B/s$. Note that our estimates improve those in [1, section 3].

As another generalization, we can replace the Laplace operator by any self-adjoint, uniformly elliptic operator provided we replace the normal boundary condition by the conormal boundary condition. Specifically, we replace Δ by the linear operator $L = D_i(a^{ij}D_j)$, where $[a^{ij}]$ is an $n \times n$ positive-definite matrix-valued function defined on Ω such that

$$\mu|\xi|^2 \leq a^{ij}(x)\xi_i\xi_j \leq \frac{1}{\mu}|\xi|^2$$

for all $\xi \in \mathbb{R}^n$, all $x \in \Omega$, and some $\mu \leq 1$. (Of course, we observe the Einstein summation convention, that repeated indices are summed from 1 to n.) The boundary condition is then

$$a^{ij} D_j u \nu_i = a^{ij} D_j v \nu_i = 0.$$

The proofs of Theorems 3.1 and 4.1 are unchanged except that C_0 and C_1 (and hence C_2, etc.) now depend also on μ.

We can also generalize our results by replacing the Laplacian by a quasi-linear operator Q of the form

$$Qu = D_i(A^i(x, Du)),$$

with A satisfying the hypotheses

$$\zeta \cdot A(x, \zeta) \geq |\zeta|^m, \qquad |A(x, \zeta)| \leq a_0 |\zeta|^{m-1}$$

for some constants $m > 1$ and $a_0 > 0$. In this case, the appropriate weak Harnack and Harnack inequalities are proved by slight modification of the ones already mentioned. (See, for example, [5, Theorem 15.7] for a brief description of the necessary modifications for the weak Harnack inequality.) The power function t^m can also be further generalized as in [9].

Finally, it is possible to deal with other boundary conditions. For the Dirichlet boundary condition $u = \varphi$ and $v = \psi$, we cannot expect positive lower bounds for u and v unless φ and ψ are strictly positive, but an upper bound for u follows from [5, Theorem 3.7] and then an upper bound for v follows by applying that same theorem to $\theta u + v$. On the other hand, for the Robin condition

$$\theta \frac{\partial u}{\partial \nu} + \gamma_1 u = 0, \quad \frac{\partial v}{\partial \nu} + \gamma_2 v = 0,$$

with γ_1 and γ_2 bounded nonpositive functions, the weak Harnack and Harnack inequalities (with C_0 and C_1 also depending on bounds for γ_1 and γ_2) and the maximum principle still hold, so we still obtain a lower bound for v and an upper bound for u. When $\gamma_1 \leq \gamma_2$ or $P < n/(n-2)$, our arguments immediately give an upper bound for v. In general, we note that the supremum of $w = \theta u + v$ can be estimated in terms of Θ, Λ, P, Ω, and $\int w \, dx$. (See, for example [5, Theorem 8.16] and [10, Theorem 6.41].) This observation provides an upper bound for v and then the lower bound for u follows as before.

REFERENCES

[1] K. J. Brown and F. A. Davidson, *Global bifurcation in the Brusselator system*, Nonlinear Anal., 24 (1995), pp. 1713–1725.

[2] F. A. Davidson and B. P. Rynne, *A priori bounds and global existence for solutions of the steady-state Sel'kov model*, Proc. Roy. Soc. Edinburgh Sect. A, 130 (2000), pp. 507–516.

[3] J. C. Eilbeck and J. E. Furter, *Analysis of bifurcations in reaction-diffusion systems with no flux boundary conditions: The Sel'kov model*, Proc. Roy. Soc. Edinburgh Sect. A, 125 (1995), pp. 413–438.

[4] T. Erneux and E. L. Reiss, *Brussellator isolas*, SIAM J. Appl. Math., 43 (1983), pp. 1240–1246.

[5] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 2001.

[6] O. A. Ladyzhenskaya and N. N. Ural'tseva, *Linear and Quasilinear Elliptic Equations*, Academic Press, New York, 1968.

[7] D. Le and H. L. Smith, *Strong positivity property of solutions to parabolic and elliptic equations on nonsmooth domains*, J. Math. Anal. Appl., 275 (2002), pp. 208–221.
[8] G. M. Lieberman, *Hölder continuity of the gradient of solutions of uniformly parabolic equations with conormal boundary conditions*, Ann. Mat. Pura Appl. (4), 148 (1987), pp. 77–99.
[9] G. M. Lieberman, *On the natural generalization of the natural conditions of Ladyzhenskaya and Ural'tseva for elliptic equations*, Comm. Partial Differential Equations, 16 (1991), pp. 311–361.
[10] G. M. Lieberman, *Second Order Parabolic Differential Equations*, World Scientific, Singapore, 1996.
[11] C. S. Lin, W. M. Ni, and I. Takagi, *Large amplitude stationary solutions to a chemotaxis system*, J. Differential Equations, 72 (1988), pp. 1–27.
[12] Y. Lou and W. M. Ni, *Diffusion, self-diffusion, and cross-diffusion*, J. Differential Equations, 131 (1996), pp. 79–131.
[13] E. E. Sel'kov, *Self-oscillations in glycolysis*, Eur. J. Biochem., 4 (1968), pp. 79–86.
[14] M. Wang, *Non-constant positive steady-states of the Sel'kov model*, J. Differential Equations, 190 (2003), pp. 600–620.

SIAM J. MATH. ANAL.
Vol. 36, No. 5, pp. 1407–1425

INITIAL BOUNDARY VALUE PROBLEMS FOR A QUASI-LINEAR PARABOLIC SYSTEM IN THREE-PHASE CAPILLARY FLOW IN POROUS MEDIA*

HERMANO FRID† AND VLADIMIR SHELUKHIN‡

Abstract. We study two types of initial boundary value problems for a quasi-linear parabolic system motivated by three-phase flows in porous media in the presence of capillarity effects. The first type of problem prescribes a mixed boundary condition, involving a combination of the value of the solution and its normal derivative at the boundary. The second type prescribes the value of the solution at the boundary, which is the so-called Dirichlet boundary condition. We prove the existence and uniqueness of smooth solution for the first type of initial boundary value problem, and we obtain the existence of a solution for the second one as a limit case of the first type. The main assumption about the diffusion matrix of the system is that it is triangular with strictly positive diagonal elements. Another interesting feature is concerned specifically with the application to three-phase capillary flow in a porous medium. Namely, we derive an important practical consequence of the assumption that the capillarity matrix is upper triangular, if we further impose that the second diagonal element depends only on the second variable, i.e., the second phase. We show that this mathematical assumption in turn provides an efficient method for the definition of the capillary pressures in the interior of the triangle of saturations through the solution of a well-posed boundary value problem for a linear hyperbolic system. Finally, as an example, we include the analysis of a special case of three-phase capillary flow model where the capillarity matrix is degenerate, but we are still able to solve it due to the particular form of the flux functions.

Key words. porous media, three-phase capillary flows, existence, uniqueness

AMS subject classifications. Primary, 35K55; Secondary, 35B65, 76S05, 76T05

DOI. 10.1137/S0036141003435333

1. Introduction.

We consider initial boundary value problems for 2×2 quasi-linear parabolic systems of the form

$$u_t + f(u)_x = (B(u)u_x)_x, \quad 0 < t < T, \quad x \in \Omega := (-1, 1), \tag{1.1}$$

motivated by one-dimensional three-phase capillary flows through a porous medium, say, an oil reservoir. Here,

$$u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \quad f = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}, \quad B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}.$$

We first explain our abstract results concerning (1.1) and then we discuss the application already mentioned.

We consider two initial boundary value problems for system (1.1):

$$\delta u_\eta + u = u_b \text{ at } |x| = 1, \ t > 0, \qquad u|_{t=0} = u_0(x), \ x \in \Omega \tag{1.2}$$

*Received by the editors September 30, 2003; accepted for publication (in revised form) July 9, 2004; published electronically March 25, 2005.
http://www.siam.org/journals/sima/36-5/43533.html

†Instituto de Matemática Pura e Aplicada—IMPA, Estrada Dona Castorina, 110, Rio de Janeiro, RJ 22460-320, Brazil (hermano@impa.br). This author's research was partially supported by CNPq through grants 352871/96-2, 46.5714/00-5, and 479416/2001-0, and FAPERJ through grant E-26/152.192/2002.

‡Lavrentyev Institute of Hydrodynamics, Lavrentyev pr. 15, Novosibirsk, 630090, Russia (shelukhin@hydro.nsc.ru). This author's research was partially supported by Russian Fund of Fundamental Researches grant 03-05-65299, INTAS grant 01-868, and CNPq grants 46.5714/00-5 and 479416/2001-0.

and

$$u = u_b \text{ at } |x| = 1,\ t > 0, \qquad u|_{t=0} = u_0(x),\ x \in \Omega, \tag{1.3}$$

where $\delta = \text{const} > 0$ and

$$u_\eta|_{x=\pm 1} = \pm u_x|_{x=\pm 1}, \quad u_b|_{x=\pm 1} = u_\pm(t).$$

Motivated by the application, the functions f_i, B_{ij}, $i, j = 1, 2$, are assumed to be defined and smooth over the closed triangle

$$\Delta := \{u \in \mathbb{R}^2 \,:\, 0 \le u_i,\ u_1 + u_2 \le 1\}, \tag{1.4}$$

and we take initial and boundary data satisfying

$$u_0(x), u_\pm(t) \in \Delta\ \forall x \in \Omega, t > 0. \tag{1.5}$$

We then seek for solutions to our problems (1.1), (1.2) and (1.1), (1.3) defined on $Q := \Omega \times (0, T)$ for any $T > 0$, which satisfy

$$u(t, x) \in \Delta\ \forall (x, t) \in Q. \tag{1.6}$$

For the mathematical analysis of the above problems, we need to assume that for certain constant $\nu > 0$ we have

$$B_{ii} \ge \nu, \quad i = 1, 2\ \forall u \in \Delta, \tag{1.7}$$

and that the matrix B is (upper) triangular, i.e.,

$$B_{21}(u) \equiv 0. \tag{1.8}$$

Concerning the behavior of the functions f_i and B_{ij} on the boundary of Δ, we assume the following:

$$f_1|_{u_1=0} = \text{const}, \quad f_2|_{u_2=0} = \text{const}, \quad (f_1 + f_2)|_{u_1+u_2=1} = \text{const}, \tag{1.9}$$

$$B_{12}|_{u_2=0} = 0, \quad (B_{11} - B_{12} - B_{22})|_{u_1+u_2=1} = 0. \tag{1.10}$$

As for the smoothness of the functions f_i, B_{ij} over Δ, we assume

$$f_i, \quad \frac{\partial f_i}{\partial u_j}, \quad B_{ij}, \quad \frac{\partial B_{ij}}{\partial u_k}, \quad \frac{\partial^2 B_{ij}}{\partial u_k \partial u_l} \in H^\beta(\Delta), \tag{1.11}$$

where $H^\beta(\Delta)$ is the space of Hölder continuous functions on Δ with $\beta \in (0, 1)$.

The initial and boundary data are also assumed to be in Hölder spaces:

$$u_0 \in H^{2+\beta}(\overline{\Omega}), \quad u_\pm(t) \in H^{1+\beta}([0, T]). \tag{1.12}$$

We can now state our result concerning the problem (1.1), (1.2).

THEOREM 1.1. *Let any $T > 0$ be given. Assume that the data $B(u)$, $f(u)$, $u_0(x)$, and $u_\pm(t)$ satisfy the conditions* (1.5), (1.7), (1.8), (1.9), (1.10), (1.11) *and* (1.12). *Suppose that the compatibility conditions*

$$\pm\delta u_0'(\pm 1) + u_0(\pm 1) = u_\pm(0)$$

are satisfied. Then problem (1.1), (1.2) *has a unique solution* $u(t,x)$ *such that* $u \in H^{2+\beta,1+\beta/2}(\overline{Q})$. *Moreover,* u *satisfies* (1.6).

The core of the proof is the application of Leray–Schauder's fixed-point theorem, which requires strong a priori estimates in Hölder spaces. This was also the approach followed in [4], where we considered the somewhat simpler case of periodic boundary conditions. It is, actually, the main point for dealing with nonlinear equations in the extensive theory developed by Ladyzhenskaya and Ural'ceva [7] and Ladyzhenskaya, Solonnikov, and Ural'ceva [6], which provide the basic tools of our approach. As is well known, the application of a fixed-point theorem to obtain a solution to an initial boundary value problem for a nonlinear PDE, in general, requires the definition of an operator in a suitable subset $\mathcal{U}$ of an appropriate space of functions. Such operator is usually given by means of the solution of a linear problem, that is, by associating with each element $v \in \mathcal{U}$ the solution $A(v)$ of a linear problem, and is such that its fixed points, that is, solutions of $A(u) = u$, solve our nonlinear problem. The Hölder spaces provide the standard setting in the context of smooth solutions. The a priori estimates in these spaces play then a twofold fundamental role. First, in order to provide bounds for the possible fixed points of the operator mentioned above, which will help us to conveniently define its domain. Second, these a priori estimates will also hold (with similar proof) for the solutions of the linear problems defining the mentioned operator, which in turn imply the properties required for the application of the fixed-point theorem. In our application of Leray–Schauder's theorem, as in [6], we need a priori estimates in the space of Hölder continuous functions, in space and time, for the solution u and its space derivative u_x. Roughly speaking, after obtaining the a priori estimate for u, to obtain the estimate for u_x, we differentiate the system with respect to x and regard the resulting equations as a linear system for $w = u_x$. In this way, the boundary condition (1.2) gives us the Hölder continuity of u_x at the boundary as a consequence of the regularity of u_b and the Hölder continuity of u in $\bar{Q}$, obtained in the first run. This is what turns problem (1.1), (1.2) easier to handle than problem (1.1), (1.3). The idea is then to obtain the solution of the latter as a limit case of the first one, when $\delta \to 0$. But, in doing that, we have to give up the context of classical solutions (up to the boundary) and content ourselves with a weaker notion of solution, which turns out to leave the question of uniqueness, for the moment, still open.

We now state our result concerning problem (1.1), (1.3).

THEOREM 1.2. *Let* $B(u)$ *and* $f(u)$ *be as in Theorem* 1.1. *Let the functions* $u_0(x)$ *and* $u_\pm(t)$ *satisfy* (1.5) *and* $u_0 \in L^2(\Omega)$, $u_\pm \in W^{1,1}(0,T)$. *Then problem* (1.1), (1.3) *has a solution* $u(t,x)$ *in the sense that* u *is a classical solution of* (1.1) *in* $Q := (0,T) \times (-1,1)$, *and* u *satisfies* (1.3) *in the sense of* $L^2(0,T)$, *for the boundary condition, and* $L^2(-1,1)$, *for the initial condition. Moreover, if* $u_\pm(t) \equiv 0$, $t \in [0,T]$, *and* $u_0 \in H^{2+\beta}(\overline{\Omega})$, *with* $u_0(\pm 1) = 0$, *then* $u(t,x)$ *is a classical solution of* (1.1) *in* Q, $u \in H^{\beta,\beta/2}(\overline{Q})$ *and the initial and boundary conditions are assumed in the usual sense for continuous functions. In any case the solution* u *satisfies* (1.6).

We prove Theorem 1.1 in sections 2 and 3. Theorem 1.2 is proved in section 4. Concerning correlated works on the theory of quasi-linear parabolic systems, we recall first that, in the well-known book of Ladyzhenskaya, Solonnikov, and Ural'ceva [6], the theory developed for linear and quasi-linear equations is applied to quasi-linear systems of the type (1.1) (in the multidimensional case, for any number of equations) where the diffusion matrix is a scalar multiple of the identity matrix. We also recall the results of Amann [2] which, in the case of 2×2 one-dimensional systems, require

$\frac{\partial f_2}{\partial u_1} \equiv 0$ and $\frac{\partial B_{22}}{\partial u_1} \equiv 0$, besides (1.7) and (1.8), and assume uniform boundedness of a certain Hölder norm of the local solution, in order to extent it to all values of time $t > 0$.

In section 5 we recall the basic facts about capillary multiphase flow in a porous medium. In this context one studies the evolution in time and space of the saturations of three immiscible fluid phases, say, oil, gas, and water, denoted by u_1, u_2, u_3, through a reservoir, which is assumed to be a (porous) solid cylinder with homogeneous cross sections. By saturation of a phase we mean the percentage of this phase in an infinitesimal pore, so that we must have $u_1 + u_2 + u_3 = 1$. In section 5, we describe how a system like (1.1) for the saturations u_1 and u_2 is derived from the mass conservation equations together with Darcy's basic law for flows in porous media. The dependent vector variable $u = (u_1, u_2)$ then must assume values in the triangle Δ defined in (1.4). By the formulas for the functions f_i and B_{ij}, derived from the mentioned basic facts, we easily check the validity of (1.9) and (1.10), when (1.8) holds. The physical diffusion matrix is diagonalizable with nonnegative eigenvalues everywhere in Δ, as shown in [1]. However, it may, in general, become singular at the boundary of Δ. So, to apply Theorems 1.1 and 1.2 one should, in general, add an artificial viscosity to the actual system.

Also in section 5, we discuss the mobility laws and capillary pressure laws which yield (1.8) and also the additional restriction

$$\frac{\partial B_{22}}{\partial u_1} = 0, \tag{1.13}$$

not needed in Theorems 1.1 and 1.2. The mobilities play a role in Darcy's law analogous to the coefficient of heat conduction in Fourier's law (see section 5). The capillary pressures, denoted here by P_1 and P_2, are defined by $P_1 = p_1 - p_3$ and $P_2 = p_2 - p_3$, where p_i is the pressure in the ith phase.

Actually, the main point of section 5 is a remarkable consequence of the mathematical assumptions (1.8) and (1.13) on the capillarity matrix. It is related with the problem of defining the capillary pressures in the interior of the triangle of saturations. As is well known by reservoir engineers, mobilities and capillary pressures can be plotted as functions of the saturations only in the case of flows where just two phases are present [10, 5], that is, when $u \in \partial\Delta$. While there are some widely accepted models which artificially prescribe the mobility functions in the interior of Δ, such as the one proposed by Stone [13], the same is no longer true for the functions representing the capillary pressures. Now, the constraints (1.8) and (1.13) amount to a linear hyperbolic system of partial differential equations for the capillary pressures P_1 and P_2, whose coefficients involve the mobility functions, for which we set the boundary conditions

$$P_i(u)|_{u_j=0} = \pi_i(u_i), \quad i, j = 1, 2, \quad j \neq i, \tag{1.14}$$

enforcing compatibility with the two-phase flow case, where the functions π_i can be obtained from two-phase flows experiments, as already mentioned. The result is a consistent recipe for defining the capillary pressures in the interior of the triangle of saturations, given the mobility functions everywhere defined therein! We call this procedure for determining P_1 and P_2 in the interior of Δ the *method of physical interpolation for capillary pressures*. As an example, we study in detail the case when the mobilities are linear functions of the corresponding phase saturation.

Finally, in section 6, the techniques developed in sections 2 to 5 are applied to the study of a particular degenerate reservoir fluid flow model. This model yields a

Dirichlet initial boundary value problem for a degenerate system of the form (1.1) where (1.8), (1.13), and yet $\frac{\partial f_2}{\partial u_1} = 0$ hold. In this case the second equation in (1.1) reads

$$\frac{\partial u_2}{\partial t} + \frac{\partial f_2(u_2)}{\partial x} = \frac{\partial}{\partial x}\left(B_{22}(u_2)\frac{\partial u_2}{\partial x}\right). \tag{1.15}$$

Since (1.6) must hold, (1.15) is not completely decoupled from the first equation in the corresponding system (1.1). Here, f and B in (1.1) have the form

$$f = \frac{1}{ku_2+1}\begin{pmatrix} u_1 \\ (1+k)u_2 \end{pmatrix}, \quad B = \begin{pmatrix} \alpha k_1(1-\xi)\pi_1'(\xi) & \xi(B_{11}-B_{22}) \\ 0 & \frac{k_1k_2u_2(1-u_2)\pi_2'(u_2)}{k_2u_2+k_1(1-u_2)} \end{pmatrix}, \tag{1.16}$$

where k_1, k_2 are positive constants, $k = \frac{k_2}{k_1} - 1$, $\xi = u_1(1-u_2)^{-1}$, and π_1, π_2 are given functions satisfying $\pi_1'(\xi) \geq 0$, $\pi_2'(u_2) \geq 0$.

The main difficulty here is that the matrix B is degenerate at $\partial\Delta$. The introduction of the variable ξ, so-called relative saturation, plays a decisive role in overcoming this difficulty. We then obtain the following result proved in section 6.

THEOREM 1.3. *Assume $\pi_1(\xi), \pi_2(u_2)$ are given smooth functions of $\xi, u_2 \in (0,1)$, with $\pi_1'(\xi) \geq 0$, $\pi_2'(u_2) \geq 0$, $0 < \xi < 1$, $0 < u_2 < 1$, and*

$$u_0 \in L^\infty(\Omega), \quad u_\pm \in W^{1,1}(0,T), \quad 0 < \delta \leq u_{2,\pm}(t) \leq 1-\delta, \quad \delta \leq \xi_\pm(t) \leq 1-\delta,$$

$$\delta \leq u_{2,0}(x) \leq 1-\delta, \quad \delta \leq \xi_0(x) \leq 1-\delta$$

for some $\delta \in (0,1)$, where

$$\xi_0 = \frac{u_{1,0}}{1-u_{2,0}}, \quad \xi_\pm = \frac{u_{1,\pm}}{1-u_{2,\pm}}.$$

Then, with f and B given by (1.16), *problem* (1.1), (1.3) *has a weak solution $u(t,x)$ such that*

$$u \in L^\infty(Q), \quad u_x \in L^2(Q), \quad u_t \in L^2(0,T;W^{-1,2}(\Omega)),$$

and the estimates

$$0 < \delta \leq u_2 \leq 1-\delta, \quad \delta \leq \xi \leq 1-\delta \tag{1.17}$$

hold.

2. A mixed initial boundary value problem. In this section, for given $\varepsilon, \delta > 0$, we consider the initial boundary value problem

$$u_t + f(u)_x = (B(u)u_x)_x + \varepsilon h, \quad (t,x) \in Q \equiv (0,T)\times\Omega, \quad \Omega = (-1,1), \tag{2.1}$$

$$\delta u_\eta + u = u_{b,\varepsilon} \quad \text{at} \quad |x| = 1, \quad u|_{t=0} = u_{0,\varepsilon}(x). \tag{2.2}$$

Here

$$u_\eta|_{x=\pm1} = \pm u_x, \quad u_b|_{x=\pm1} = u_\pm(t),$$
$$u_{i\pm,\varepsilon} = (1-\varepsilon)\left(\frac{\varepsilon}{2} + u_{i\pm}\right), \quad u_{i0,\varepsilon} = (1-\varepsilon)\left(\frac{\varepsilon}{2} + u_{i0}\right), \quad i = 1,2.$$

The function $h = (h_1, h_2)$ is any smooth map satisfying $h(\omega) \cdot \nu(\omega) < 0$, for all $\omega \in \partial\Delta$, with $\nu(\omega)$ the unit outer normal to $\partial\Delta$, defined everywhere in $\partial\Delta$, except at the vertices. For example, $h(u) = u_* - u$, where u_* is any point in the interior of Δ.

The problem (2.1), (2.2) should be considered as a perturbation of problem (1.1), (1.2). The inclusion of the perturbation term εh and the slight change in the initial and boundary data are necessary for the proof of the positive invariance of the triangle Δ in Lemma 2.1 below. However, we remark in advance that in all the remaining lemmas in this section, the constants c in the estimates do not depend on ε, which will easily allow to make $\varepsilon \to 0$ in the conclusion of the proof of Theorem 1.1, in the next section. In Lemmas 2.1 to 2.10 below, we omit the subscript ε when referring to the functions $u_{0,\varepsilon}(x)$, $u_{\pm,\varepsilon}(t)$.

LEMMA 2.1. *The solution u of* (2.1), (2.2) *takes values in* $\mathrm{int}(\Delta)$ *whenever u_0 and $u_\pm$ take values in Δ.*

Proof. We follow the method of positively invariant regions [12] (see also [11]). Let us denote $u_3 = 1 - u_1 - u_2$ and $G_i(u) = -u_i$, $i = 1, 2, 3$. Setting $z_i = G_i(u)$, we prove that $z_i < 0$ for each i. Clearly,

$$\max_{x\in\Omega} z_i(0, x) < 0, \quad i \in \{1, 2, 3\}.$$

Suppose there is a first time $t_1 > 0$ such that

$$\max_{x\in\Omega} z_i(t_1, x) = z_i(t_1, x_0) = 0$$

for some i. We consider separately the two cases $|x_0| < 1$ and $|x_0| = 1$. First, we observe that $x_0 = 1$ is impossible. Indeed, it follows from the equality

$$\delta z_{ix} + z_i = -u_{i,+\varepsilon} \tag{2.3}$$

that $z_{ix}(t_1, 1) < 0$, which gives a contradiction since $z_i(t_1, x) \le 0$, $x \in [-1, 1]$. Similarly, we prove that $x_0 = -1$ cannot hold either. So, the case $|x_0| = 1$ is ruled out.

Let us consider the case $|x_0| < 1$. Multiplying (2.1) by $\nabla_u G_i$ and using (1.9), (1.10), one arrives at

$$z_{it} + \alpha_i z_{ix} = (\mu_i z_{ix})_x + \varepsilon h \cdot \nabla_u G_i \quad \text{at} \quad (t_1, x_0), \tag{2.4}$$

where α_i and $\mu_i \ge 0$ are scalar functions of $u \in \{G_i(u) = 0\} \cap \Delta$. By the assumption,

$$z_i(t_1, x_0) = \max_{\substack{0\le\tau\le t_1 \\ |y|\le 1}} z_i(\tau, y).$$

Hence, we must have

$$z_{ix}(t_1, x_0) = 0, \quad z_{ixx}(t_1, x_0) \le 0, \quad z_{it}(t_1, x_0) \ge 0. \tag{2.5}$$

Due to the choice of h, we have

$$h \cdot \nabla_u G_i < 0 \quad \text{at} \quad (t_1, x_0).$$

Now, it follows from (2.4) that $z_{it}(t_1, x_0) < 0$, contradicting (2.5). ∎

LEMMA 2.2. *The estimate*

$$\|u_x\|_{L^2(Q)} + \delta \sum_{\pm} \int_0^T |u_x(t, \pm 1)|^2 dt \le c \tag{2.6}$$

holds with a constant c depending on ν and the norms $\|\dot{u}_\pm\|_{L^1(0,T)}$ and $\|h\|_{L^1(Q)}$. In particular, by (2.1)*, it follows that*

$$\|u_t\|_{L^2(0,T;W^{-1,2}(\Omega))} \le c \tag{2.7}$$

uniformly in ε and δ.

Proof. Denote

$$w = \frac{1-x}{2}u_- + \frac{1+x}{2}u_+ , \quad v = u - w.$$

Then we have from the second equation in (2.1) that

$$\frac{1}{2}\frac{d}{dt}\int_\Omega v_2^2 dx + \int_\Omega B_{22}|v_{2x}|^2 dx = v_2(B_{22}(v_{2x}+w_{2x}) - f_2)|_{-1}^{+1} + \int_\Omega v_{2x}(f_2 - B_{22}w_{2x}) - w_{2t}v_2 + \varepsilon h_2 v_2 \, dx.$$

Since

$$v|_{x=\pm 1} = \mp\delta(v_x + w_x)|_{x=\pm 1}$$

and $B_{22} \ge \nu$, estimate (2.6) for u_2 follows by the Cauchy inequality.

From the first equation in (2.1), we have

$$\frac{1}{2}\frac{d}{dt}\int_\Omega v_1^2 dx + \int_\Omega B_{11}|v_{1x}|^2 dx = v_1(B_{11}(v_{1x}+w_{1x}) + B_{12}u_{2x} - f_1)|_{-1}^{+1} - \int_\Omega v_{1x}(B_{11}w_{1x} + B_{12}u_{2x} - f_1) - w_{1t}v_1 + \varepsilon h_1 v_1 \, dx.$$

By the same argument, one can derive the claim of the lemma for the function u_1, using estimate (2.6) for u_2. ∎

Let $|u|_Q^{(\alpha)}$ denote the norm of the function $u(x,t)$ in the Hölder space $H^{\alpha,\alpha/2}(\overline{Q})$ (cf. [6]):

$$|u|_Q^{(\alpha)} = \sup_{\overline{Q}}|u(x,t)| + \sup_{x_1,x_2\in\overline{\Omega},t\in[0,T]}\frac{|u(x_1,t)-u(x_2,t)|}{|x_1-x_2|^\alpha} + \sup_{x\in\overline{\Omega},t_1t_2\in[0,T]}\frac{|u(x,t_1)-u(x,t_2)|}{|t_1-t_2|^{\alpha/2}}.$$

The following estimates depend, in general, on δ.

LEMMA 2.3. *There are constants c and $\alpha \in (0,1)$ such that*

$$|u_2|_Q^{(\alpha)} \le c. \tag{2.8}$$

Moreover, if $u_\pm \equiv 0$, the estimate (2.8) *holds uniformly in δ.*

Proof. Let $\zeta(t,x)$ be a test function with values between 0 and 1 and that is different from zero only for $x \in K_\rho$, the ball of radius ρ centered at $x^0 \in \overline{\Omega}$. Denote

$$\Omega_\rho = \overline{\Omega} \cap K_\rho = [x_-^0, x_+^0], \quad x_+^0 = \min\{1, x_0+\rho\}, \quad x_-^0 = \max\{-1, x_0-\rho\}.$$

Given $\delta' > 0$, we multiply the second equation in (2.1) by

$$\zeta^2 \max\{u_2 - k, 0\} \equiv \zeta^2 u_2^{(k)}, \quad k \geq -\delta',$$

and integrate over Ω_ρ. We have

$$\frac{1}{2}\frac{d}{dt}\int_{\Omega_\rho} \zeta^2 |u_2^{(k)}|^2\,dx + \int_{\Omega_\rho} \zeta^2 B_{22}|u_{2x}^{(k)}|^2\,dx = \zeta^2 B_{22} u_{2x} u_2^{(k)}|_{x_-^0}^{x_+^0} - \zeta^2 f_2 u_2^{(k)}|_{x_-^0}^{x_+^0}$$

$$- \int_{\Omega_\rho} 2\zeta\zeta_x B_{22} u_{2x} u_2^{(k)} - \zeta\zeta_t |u_2^{(k)}|^2 - f_2(2\zeta\zeta_x u_2^{(k)} + \zeta^2 u_{2x}^{(k)}) - \varepsilon h_2 \zeta^2 u_2^{(k)}\,dx.$$

Observe that

$$\delta u_x|_{x=\pm 1} = \pm(u_\pm - u)|_{x=\pm 1},$$

$$\zeta^2 B_{22} u_{2x} u_2^{(k)}|_{x_-^0}^{x_+^0} \leq \frac{1}{\delta}\zeta^2 B_{22} u_2^{(k)} u_{2+}|_{x=1} + \frac{1}{\delta}\zeta^2 B_{22} u_2^{(k)} u_{2-}|_{x=-1},$$

$$|\zeta^2 v^{(k)}|_{|x|=1} \leq |\int_{\Omega_\rho} \zeta^2 v_x^{(k)} + 2\zeta\zeta_x v^{(k)}\,dx| \tag{2.9}$$

for small ρ. Thus, using $B_{22} \geq \nu$,

$$\frac{1}{2}\frac{d}{dt}\int_{\Omega_\rho} \zeta^2 |u_2^{(k)}|^2\,dx + \nu\int_{\Omega_\rho} \zeta^2 |u_{2x}^{(k)}|^2\,dx \tag{2.10}$$
$$\leq \frac{\nu}{2}\int_{\Omega_\rho} \zeta^2 |u_{2x}^{(k)}|^2\,dx + c_1\int_{\Omega_\rho} |u_2^{(k)}|^2(|\zeta_x|^2 + |\zeta\zeta_t|) + \zeta^2 \mathbf{1}_{A_{k,\rho}(t)}\,dx,$$

where $A_{k,\rho}(t)$ is the intersection of the support of $u_2^{(k)}$ with K_ρ, and $\mathbf{1}_A$ is the characteristic function of the set A. Proceeding in an analogous way, we prove that (2.10) also holds with u_2 replaced by $-u_2$, for $k \leq 1 + \delta'$. These inequalities imply that u_2 belongs to a class $\mathcal{B}_2(Q \cup \Gamma, M, \gamma, r, \delta', \kappa)$ (see [6, Chapter II, sections 7 and 8]), with $\Gamma = \partial Q \setminus \{t = T\}$, $M = 1$, $r = 6$, $\kappa = 2$, and, hence, $u_2 \in H^{\alpha,\alpha/2}(\overline{Q})$ for some $\alpha \in (0, 1)$.

As for the last statement, indeed, in this case,

$$\zeta^2 B_{22} u_{2x} u_2^{(k)}\Big|_{x_-^0}^{x_+^0} = -\sum_{x=\pm 1} \frac{1}{\delta}\zeta^2 B_{22} u_2 u_2^{(k)} \leq 0,$$

and the constant c_1 in (2.10) does not depend on δ. ∎

LEMMA 2.4. *There is a constant c such that*

$$\max_{0\leq t\leq T}\left\{\int_\Omega u_{2x}^2\,dx + \delta\left(\sum_{x=\pm 1} u_{2x}^2\right)\right\} + \int_Q u_{2xx}^2 + u_{2x}^4 + u_{2t}^2\,dx\,dt \leq c. \tag{2.11}$$

Moreover, if $u_\pm \equiv 0$, the constant c in (2.11) *does not depend on δ.*

Proof. Let $\zeta(x)$ be the test function like above. We multiply the second equation in (2.1) by $(\zeta^2 u_{2x})_x$ and integrate over Ω_ρ. Using the equality

$$\delta\dot{u}_\eta + \dot{u} = \dot{u}_b \quad \text{at} \quad x = \pm 1,$$

and the Young inequality, we obtain that

$$\frac{1}{2}\frac{d}{dt}\left\{\int_{\Omega_\rho}\zeta^2 u_{2x}^2\,dx+\delta\left(\sum_{x=x_\pm^0}\zeta^2u_{2x}^2\right)\right\}+\nu\int_{\Omega_\rho}\zeta^2u_{2xx}^2\,dx\le J,$$

$$J\equiv\|\dot u_b\|_{C([0,T])}\times\sum_{x=x_\pm^0}|\zeta^2u_{2x}|+\frac{\nu}{4}\int_{\Omega_\rho}\zeta^2u_{2xx}^2\,dx$$
$$+c_*\int_{\Omega_\rho}\zeta^2u_{2x}^4dx+c\int_{\Omega_\rho}\zeta^2(u_{2x}^2+u_{1x}^2)+u_{2x}^2(\zeta_x^2+\zeta^2)+\zeta^2\,dx.$$

We estimate the first term in the right-hand side of the above identity using an inequality like (2.9) and, again, Young's inequality, in order to obtain

$$\|\dot u_b\|_{C([0,T])}\times\sum_{x=x_\pm^0}|\zeta^2u_{2x}|\le\frac{\nu}{4}\int_{\Omega_\rho}\zeta^2u_{2xx}^2\,dx+c_\nu\int_{\Omega_\rho}u_{2x}^2\zeta_x^2+\zeta^2\,dx.$$

Finally, we recall the inequality (see [6, Chapter II, Lemma 5.4])

$$\int_{K_\rho}\zeta^2v_x^4dx\le16\mathrm{osc}^2\{v,K_\rho\}\int_{K_\rho}2\zeta^2v_{xx}^2+\zeta^2v_x^2\,dx.\tag{2.12}$$

By Lemma 2.3,

$$\mathrm{osc}^2\{u_2,K_\rho\}\le c\rho^{\alpha_1},\quad\alpha_1<\alpha.$$

Now, the assertion of the lemma follows if we take ρ such that $32c_*\rho^\alpha<\nu/4$, where α is the constant from Lemma 2.3. □

LEMMA 2.5. *There are constants c and $\alpha\in(0,1)$ such that $|u_1|_Q^{(\alpha)}\le c$. Moreover, if $u_\pm\equiv0$, the constant c in Lemma 2.5 does not depend on δ.*

Proof. Let $\zeta(t,x)$ be a function like in Lemma 2.3. Then, given $\delta'>0$, for $k\ge-\delta'$,

$$\left(\frac{1}{2}\frac{d}{dt}\int_{\Omega_\rho}\zeta^2|u_1^{(k)}|^2\,dx+\int_{\Omega_\rho}\zeta^2B_{11}|u_{1x}^{(k)}|^2\,dx=J_1+J_2\ ,\quad J_1=\zeta^2B_{11}u_{1x}u_1^{(k)}\ \Big|_{x_-^0}^{x_+^0},\right.$$

$$J_2=\int_{\Omega_\rho}\zeta\zeta_tu_1^{(k)}-2\zeta\zeta_xB_{11}u_{1x}u_1^{(k)}+u_1^{(k)}\zeta^2\left[(B_{12}u_{2x})_x-\frac{\partial f_1}{\partial x}+\varepsilon h_1\right]dx.$$

We have

$$J_1\le\frac{1}{\delta}\sum_{x=\pm1}\zeta^2B_{11}u_{1,\pm}u_1^{(k)}\overset{(2.9)}{\le}\int_{\Omega_\rho}\frac{\nu}{4}\zeta^2|u_{1x}^{(k)}|^2+c\,\zeta_x^2|u_1^{(k)}|^2+\zeta^2\mathbf{1}_{A_{k,\rho}(t)}\,dx,$$

$$J_2\le\int_{\Omega_\rho}|u_1^{(k)}|^2(|\zeta\zeta_t|$$
$$+\zeta^2)\,dx+c\left(\int_{\Omega_\rho}\zeta\mathbf{1}_{A_{k,\rho}(t)}\,dx\right)^{1/2}\left(\int_{\Omega_\rho}\mathbf{1}_{A_{k,\rho}(t)}(1+u_{2x}^4+u_{2xx}^2\right)dx)^{1/2}.$$

As in Lemma 2.3, we can prove analogously that the above inequality also holds with u_1 replaced by $-u_1$ and $k\le1+\delta'$. Hence, (see [6, Chapter II, sections 7 and 8]) the

function u_1 belongs to the class $\mathcal{B}_2(Q\cup\Gamma, M, \gamma, r, \delta', \kappa)$, for some $\gamma > 0$, with $M = 1$, $r = 6$, $\kappa = 2$, and the lemma is proved. ∎

LEMMA 2.6. *There is a constant c such that*

$$\max_{0\le t\le T}\left\{\int_\Omega u_{1x}^2\,dx + \delta\sum_{x=\pm1}|u_{1x}|^2\right\} + \int_Q u_{1xx}^2 + u_{1x}^4 + u_{1t}^2\,dx\,dt \le c.$$

Moreover, if $u_\pm \equiv 0$, the constant c in the above lemma does not depend on δ.

Proof. Let $\zeta(x)$ be a test function like in Lemma 2.5. Then it follows from the first equation in (2.1) that

$$\frac{1}{2}\frac{d}{dt}\left\{\int_{\Omega_\rho}\zeta^2u_{1x}^2\,dx + \delta\left(\sum_{x=x_\pm^0}\zeta^2u_{1x}^2\right)\right\} + \nu\int_{\Omega_\rho}\zeta^2u_{1xx}^2\,dx \le cJ,$$

$$J = \|\dot u_b\|_{C([0,T])}\sum_{x=x_\pm^0}|\zeta^2u_{1x}| + \frac{\nu}{2}\int_{\Omega_\rho}\zeta^2u_{1xx}^2\,dx + \int_{\Omega_\rho}\zeta_x^2(u_{2x}^2 + u_{1x}^2 + u_{2xx}^2 + u_{2x}^4)\,dx$$
$$+ \int_{\Omega_\rho}\zeta^2(u_{1x}^4 + u_{2x}^4 + u_{1x}^2\zeta_x^2 + u_{2xx}^2 + u_{1x}^2 + 1 + u_{2x}^4\zeta_x^4 + u_{2x}^2\zeta_x^2).$$

Applying (2.12), one arrives at the conclusion of the lemma. ∎

LEMMA 2.7. *There is a constant c such that*

$$\int_Q |u_{ix}|^6\,dx\,dt \le c, \quad \int_Q |u_{ixx}|^3\,dx\,dt \le c, \quad \int_Q |u_{ix}u_{ixx}|^2\,dx\,dt \le c. \tag{2.13}$$

Proof. We start with the simple inequality

$$\int_Q |u_{ix}|^6\,dx\,dt \le \int_0^T \max_{x\in\Omega}|u_{ix}|^4\int_\Omega |u_{ix}|^2\,dx\,dt.$$

Observe that for any x and y,

$$v_x^2(x) - v_x^2(y) = 2\int_x^y v_{xx}v_x dz,$$

so

$$\max_{|x|<1} v_x^4 \le \|v_x\|_{L^2(\Omega)}^2 + 8\|v_{xx}\|_{L^2(\Omega)}^2\|v_x\|_{L^2(\Omega)}^2.$$

Hence,

$$\|u_{ix}\|_{L^6(Q)}^6 \le \|u_{ix}\|_{L^\infty(0,T;L^2(\Omega))}^4(1 + 8\|u_{ixx}\|_{L^2(Q)}^2) \le c,$$

and the first estimate of the lemma is proved.

Let us write the second equation in (2.1) as

$$\begin{cases} u_{2t} = B_{22}u_{2xx} + F, \\ F = \frac{\partial B_{22}}{\partial u_2}|u_{2x}|^2 - (\frac{\partial f_2}{\partial u_1}u_{1x} + \frac{\partial f_2}{\partial u_2}u_{2x}) + \frac{\partial B_{22}}{\partial u_1}u_{1x}u_{2x} + \varepsilon h_2. \end{cases} \tag{2.14}$$

By Lemmas 2.4 and 2.6, $\|F\|_{L^3(Q)} \le c$. Now, the theory of linear parabolic equations with smooth coefficients (see [6, Chapter IV, section 9]) can be applied to derive the estimate

$$\int_Q |u_{2xx}|^3 \, dx \, dt \le c, \tag{2.15}$$

which is the second estimate in (2.13) for u_2. The third estimate in (2.13) for u_2 follows because of the inequality

$$\int_Q |uv|^2 \, dx \, dt \le \left(\int_Q |u|^6 \, dx \, dt \right)^{1/3} \left(\int_Q |v|^3 \, dx \, dt \right)^{2/3}. \tag{2.16}$$

The first equation in (2.1) writes

(2.17)

$$\begin{aligned} u_{1t} &= B_{11} u_{1xx} + F, \\ F = u_{1x} &= \left(\frac{\partial B_{11}}{\partial u_1} u_{1x} + \frac{\partial B_{11}}{\partial u_2} u_{2x} \right) + u_{2x} \left(\frac{\partial B_{12}}{\partial u_1} u_{1x} + \frac{\partial B_{12}}{\partial u_2} u_{2x} \right) + B_{12} u_{2xx} \\ &\quad - \left(\frac{\partial f_1}{\partial u_1} u_{1x} + \frac{\partial f_1}{\partial u_2} u_{2x} \right) + \varepsilon h_1, \end{aligned}$$

where, as above, $\|F\|_{L^3(Q)} \le c$. Hence, the estimate (2.15) is also valid for u_{1xx}, which concludes the verification of the second estimate in (2.13). Now, the third estimate in (2.13) for u_1 follows also due to (2.16). ∎

LEMMA 2.8. *There are constants c and $\alpha \in (0,1)$ such that $|u_{2x}|_Q^{(\alpha)} \le c$.*

Proof. Let us differentiate the second equation in (2.1) with respect to x. The function $v := u_{2x}$ solves the linear equation

$$\begin{aligned} v_t &= (B_{22} v_x)_x + F_2 + g_x, \\ F_2 &= \frac{\partial^2 B_{22}}{\partial u_2^2} (u_{2x})^3 + 2 \frac{\partial B_{22}}{\partial u_2} u_{2x} u_{2xx} + \frac{\partial B_{22}}{\partial u_1} (u_{1xx} u_{2x} + u_{1x} u_{2xx}) \\ &\quad + \frac{\partial^2 B_{22}}{\partial u_1^2} u_{1x}^2 u_{2x} + 2 \frac{\partial^2 B_{22}}{\partial u_1 \partial u_2} u_{1x} u_{2x}^2, \\ g &= -\frac{\partial f_2}{\partial u_1} u_{1x} - \frac{\partial f_2}{\partial u_2} u_{2x} + \varepsilon h_2. \end{aligned}$$

By the above lemmas,

$$\|F_2\|_{q,r,Q} \equiv \left(\int \left(\int_\Omega F_2^q \right)^{r/q} dt \right)^{1/r} \le c, \quad \|g^2\|_{q,r,Q} \le c,$$

when $q = 2$, $r = 2$. Clearly, the constants q and r satisfy the conditions

$$\frac{1}{r} + \frac{1}{2q} = 1 - \kappa, \quad 0 < \kappa < \frac{1}{2}, \quad q \in [1, \infty], \quad r \in \left[\frac{1}{1-\kappa}, \frac{2}{1-2\kappa} \right],$$

with $\kappa = 1/4$. Moreover, it follows from the boundary conditions (2.2) and Lemma 2.3, that

$$\|v(t, \pm 1)\|_{H^{\alpha/2}([0,T])} \le c.$$

Thus, by the theory of linear parabolic equations (see [6, Chapter III, section 10])

$$|v|_Q^{(\alpha')} \le c$$

for some $\alpha' \le \alpha$. ∎

LEMMA 2.9. *There are constants c and $\alpha \in (0,1)$ such that $|u_{1x}|_Q^{(\alpha)} \le c$.*

Proof. Denoting $u_{1x} = v$ and differentiating the first equation in (2.1) with respect to x, we have

(2.18)

$$v_t = (B_{11}(u_1,u_2)v_x)_x + F_1 + g_{1x},$$

$$F_1 = u_{1xx}\left(\frac{\partial B_{11}}{\partial u_1}u_{1x} + \frac{\partial B_{11}}{\partial u_2}u_{2x}\right) + u_{1x}\left\{\frac{\partial B_{11}}{\partial u_1}u_{1xx} + u_{1x}\left(\frac{\partial^2 B_{11}}{\partial u_1^2}u_{1x} + \frac{\partial^2 B_{11}}{\partial u_1 \partial u_2}u_{2x}\right)\right.$$
$$\left. + \frac{\partial B_{11}}{\partial u_2}u_{2xx} + u_{2x}\left(\frac{\partial^2 B_{11}}{\partial u_1 \partial u_2}u_{1x} + \frac{\partial^2 B_{11}}{\partial u_2^2}u_{2x}\right)\right\},$$

$$g_1 = u_{2x}\left(\frac{\partial B_{12}}{\partial u_1}u_{1x} + \frac{\partial B_{12}}{\partial u_2}u_{2x}\right) + B_{12}u_{2xx} - \left(\frac{\partial f_1}{\partial u_1}u_{1x} + \frac{\partial f_1}{\partial u_2}u_{2x}\right) + \varepsilon h_1.$$

By Lemma 2.7, $\|F_1\|_{L^2(Q)} \le c$. With the estimate of Lemma 2.8 for u_{2x} at hand, the function F in (2.14) meets the estimate $\|F\|_{L^4(Q)} \le c$. It implies that

$$\|u_{2xx}\|_{L^4(Q)} \le c, \quad \|g_1\|_{L^4(Q)} \le c.$$

Now, one may treat (2.18) as a linear parabolic equation for v, with

$$\|F_1, g_1^2\|_{2,2,Q} \le c, \quad \|v(t,\pm 1)\|_{H^{\alpha/2}([0,T])} \le c.$$

By the same argument as in Lemma 2.8, we conclude that

$$|u_{1x}|_Q^{(\alpha')} \le c$$

for some $\alpha' < \alpha$. ∎

LEMMA 2.10. *Let*

$$u_0 \in H^{2+\beta}(\overline{\Omega}), \quad u_\pm \in H^{1+\beta/2}([0,T]), \quad 0 < \beta < 1,$$

and the compatibility conditions

$$\pm\delta u_0'(\pm 1) + u_0(\pm 1) = u_\pm(0) \tag{2.19}$$

be satisfied. Then there is a constant c such that solutions to problem (2.1), (2.2) *satisfy the estimate $|u|_Q^{(2+\beta)} \le c$. The constant c does not depend on ε but depends on T, $\|u_0|H^{2+\beta}(\overline{\Omega})\|$, $\|u_\pm|H^{1+\beta/2}([0,T])\|$, and the L^∞-norms of $f(u)$, $\nabla_u f$, $B_{ij}(u)$, $\nabla_u B_{ij}$, and $\nabla_u^2 B_{ij}$.*

Proof. First, we observe that the data $u_{0\varepsilon}$ and $u_{b\varepsilon}$ also satisfy the compatibility conditions (2.19). We know from the above lemmas that there are constants c and $\alpha \in (0,1)$ such that $|u_i, u_{ix}|_Q^{(\alpha)} \le c$. If $\gamma = \min\{\alpha, \beta\}$, it follows from the linear equation (2.14) that $|u_2|_Q^{(2+\gamma)} \le c$ (see [6, Chapter IV, section 5]). By the same argument, we conclude from (2.17) that $|u_1|_Q^{(2+\gamma)} \le c$. To increase γ up to β, one should return to problem (2.14), which now ensures that $|u_2|_Q^{(2+\beta)} \le c$. Next, one should pass to (2.17) to make sure that $|u_1|_Q^{(2+\beta)} \le c$. ∎

3. Existence and uniqueness. To prove the solvability of problem (2.1), (2.2), we apply a fixed-point argument in the form of the Leray–Schauder principle as in [6]. Let $\mathcal{B}$ be a Banach space of vector functions $u(t,x) \in \mathbb{R}^2$, having the bounded norm

$$\|u\|_{\mathcal{B}} = |u|_Q^{(\beta)} + |u_x|_Q^{(\beta)}.$$

Given $v \equiv (v_1, v_2) \in \mathcal{B}$ and $\lambda \in [0,1]$, we define $u = (u_1, u_2)$ as a solution to the linear problem

$$\begin{aligned} &u_{1t} + \lambda\left[\frac{\partial f_1(v)}{\partial x} - (B_{11}(v)u_{1x})_x - (B_{12}(v)u_{2x})_x - \varepsilon h_1(\mathbf{a})\right] = (1-\lambda)u_{1xx}, \\ &u_{2t} + \lambda\left[\frac{\partial f_2(v)}{\partial x} - (B_{22}(v)u_{2x})_x - \varepsilon h_2(v)\right] = (1-\lambda)u_{2xx}, \end{aligned} \tag{3.1}$$

$$\delta u_\eta + u = u_{b\varepsilon}, \quad u|_{t=0} = u_{0\varepsilon} = (1-\varepsilon)\left(\frac{\varepsilon}{2} + u_{01}, \frac{\varepsilon}{2} + u_{02}\right). \tag{3.2}$$

The second equation does not involve the function u_1. So, by the theory of linear parabolic equations, given $\lambda \in [0,1]$, the operator A_λ, which associates with each $v \in \mathcal{B}$ the solution $A_\lambda(v)$ of (3.1), (3.2), is well defined, and its fixed points are solutions to problem (2.1), (2.2) when $\lambda = 1$.

In order to apply Leray–Schauder theorem, as in [6], one must choose appropriately a domain $\mathcal{U}$ for the operators A_λ, $\lambda \in [0,1]$, and verify each of the conditions of the theorem. By repeating the arguments of the lemmas in section 2, one arrives at the a priori estimates for the fixed points u_λ of the operator A_λ:

$$u_\lambda \in \Delta, \quad |u_\lambda, u_{\lambda x}|_Q^{(\beta)} \le M,$$

where the constants M, M_1 are independent of λ. We choose as the domain of the operators A_λ the set

$$\mathcal{U} = \{u \in \mathcal{B} : u(x,t) \in \Delta', \quad |u_\lambda, u_{\lambda x}|_Q^{(\beta)} \le M'\},$$

where $\Delta' \subseteq \mathbb{R}^2$ is a closed set satisfying $\mathrm{int}(\Delta') \supset \Delta$ and $M' > M$. Clearly, $\mathcal{U}$ is the closure of a bounded connected open set in $\mathcal{B}$, and all the fixed points u_λ of A_λ are in the interior $\mathcal{U}$. This means that we have verified the first and most difficult of the following conditions for the application of Leray–Schauder theorem:

1. the boundary of $\mathcal{U}$ does not contain solutions of $A_\lambda(u) = u$;
2. the set $\cup_{\lambda\in[0,1]} A_\lambda(\mathcal{U})$ is compact in $\mathcal{B}$;
3. the mapping $(\lambda, v) \mapsto A_\lambda(v)$ is continuous from $\mathcal{U} \times [0,1]$ to $\mathcal{B}$;
4. the family of maps $\{v \mapsto A_\lambda(v)\}_{\lambda\in[0,1]}$ is equicontinuous on $\mathcal{U}$;
5. the operator A_0 has a unique fixed point in the interior of $\mathcal{U}$, and the mapping $v \mapsto v - A_0(v)$ has an inverse near this fixed point.

The verification of the conditions (2) to (4) above is immediate from all the preceding discussion in this section and section 2, while (5) follows immediately from the fact that $A_0(v)$ is in fact independent of v, that is, it is one and the same element of $\mathcal{B}$ for all v. Hence, problem (2.1), (2.2) has at least one solution in the Hölder space $H^{2+\beta,1+\beta/2}(\bar{Q})$. Uniqueness obtained in a standard way, by using the theory of linear equations with smooth coefficients (see [6, Chapter IV]). Thus, we have proved the following.

THEOREM 3.1. *Let the functions $f(u)$, $\nabla_u f$, $B_{ij}(u)$, $\nabla_u B_{ij}$, $\nabla_u^2 B_{ij}$, and the function $h(u)$ be Hölder continuous with the Hölder exponent $\beta \in (0,1)$. Let the*

conditions of Lemma 2.10 *be satisfied. Then problem* (2.1), (2.2) *has a unique solution* $u(t,x) \in H^{2+\beta,1+\beta/2}(\bar{Q})$ *such that* $u(t,x) \in \Delta$ *for each* $(t,x) \in Q$.

Proof of Theorem 1.1. Since the estimate of Lemma 2.10 does not depend on ε, there is a sequence $\varepsilon_k \downarrow 0$, such that the corresponding sequence u_k of solutions of problem (2.1), (2.2) converges to a function $u \in H^{2+\beta,1+\beta/2}(\overline{Q})$ in the norm $|\cdot|_Q^{(2+\gamma)}$ for any $\gamma < \beta$. Clearly, u solves the problem (1.1), (1.2). Thus, Theorem 1.1 is proved. □

4. Dirichlet boundary conditions.

Proof of Theorem 1.2. Let us consider the Dirichlet problem (1.1), (1.3). The estimates (2.6) and (2.7) are uniform with respect to $\delta \downarrow 0$. By the Aubin–Lions compactness theorem [8], they imply that there is a sequence u_k, $\delta_k \downarrow 0$, of solutions to problem (1.1), (1.2) and a function u such that

$$u \in L^\infty(Q), \quad u_x \in L^2(Q), \quad u_t \in L^2(0,T;W^{-1,2}(\Omega)), \tag{4.1}$$

$$u_k \to u \quad \text{in} \quad L^2(Q), \quad u(t,x) \in \Delta \quad \text{for each} \quad (t,x) \in Q.$$

Clearly, the function u solves (1.1) weakly. Now, given any open set Q' with $\overline{Q'} \subseteq Q$, we have that u_k is uniformly bounded in $H^{2+\beta,1+\beta/2}(\overline{Q'})$, so that $u \in H^{2+\beta,1+\beta/2}(\overline{Q'})$. In particular, u is a classical solution of (1.1) in Q. Due to estimate (2.6), the boundary condition $u|_{x=\pm 1} = u_\pm$ holds in $L^2(0,T)$. The inclusions (4.1) imply that $u \in C(0,T;L^2(\Omega))$, so the function u satisfies the initial condition $u|_{t=0} = u_0$ weakly in $L^2(\Omega)$. This proves the first part of Theorem 1.2. The last part is a consequence of the fact that, when $u_b = 0$, the estimates of Lemmas 2.2 to 2.5 are uniform in δ. Therefore, in this case we have $u \in H^{\beta,\beta/2}(\overline{Q})$ and, so, the last assertion follows. □

5. Basic equations of three-phase flow. For the reader's convenience we recall the underlying laws of multiphase flows in a porous medium [1]. We consider one-dimensional horizontal flows of three incompressible immiscible fluids formed in phases. The balance of masses is governed by the mass conservation equations

$$\frac{\partial}{\partial t}(m u_i \rho_i) + \frac{\partial}{\partial x}(\rho_i v_i) = 0, \quad \rho_i = \text{const}, \tag{5.1}$$

where m denotes porosity of the porous medium, u_i, ρ_i, and v_i are the saturation, density, and seepage velocity of the ith phase. The functions u_i satisfy the volume-balance equation

$$u_1 + u_2 + u_3 = 1. \tag{5.2}$$

The theory of multiphase flows in porous media is based on the following form of Darcy's law:

$$v_i = -k\lambda_i p_{ix}, \tag{5.3}$$

where k stands for the absolute permeability, p_i is the pressure of the ith phase, and λ_i is the mobility of the ith phase and it is assumed to be a function of $u = (u_1, u_2)$, that is, $\lambda_i = \lambda_i(u_1, u_2)$, $i = 1, 2, 3$. Experimentally, only the functions $\lambda_1(u_1, 0)$, $\lambda_2(0, u_2)$, $\lambda_3(u_1, 0)$, and $\lambda_3(0, u_2)$ are known, which correspond to mobilities in two-phase flows. But there are widely accepted models prescribing the mobility functions everywhere in Δ, based on the knowledge of them for two-phase flows, such as the one of Stone [13].

The capillary pressures are defined as pressure differences (cf., e.g., [10, 1, 3, 14]), and we assume here that they are functions of the saturations u_1, u_2, that is,

$$P_1(u_1,u_2)=p_1-p_3,\quad P_2(u_1,u_2)=p_2-p_3. \tag{5.4}$$

Similarly to what happens with the mobility functions, the capillary pressure functions are only known in practice in the case of two-phase flows. However, as opposed to the case of the mobility functions, so far there is no widely accepted way of defining these functions everywhere in Δ, assuming them to be known for two-phase flows. One of the main points in this section is the introduction of a new method to define the capillary pressures everywhere in Δ, assuming that they are known for two-phase flows and that the mobility functions are known everywhere in Δ.

Denote

$$\lambda=\sum_1^3\lambda_i,\quad f_i=\frac{\lambda_i}{\lambda},\qquad i=1,2,3. \tag{5.5}$$

For

$$v=\sum_1^3 v_i,$$

we find from (5.1) and (5.2) that $v_x=0$, so v depends on t only. We assume for simplicity that $k=m=1$ as well.

Eliminating the pressure derivative p_{3x}, we have from (2.3)

$$v_1=f_1(v(t)+\lambda_2P_{2x}-(\lambda_2+\lambda_3)P_{1x}),\quad v_2=f_2(v(t)+\lambda_1P_{1x}-(\lambda_1+\lambda_3)P_{2x}).$$

If we substitute these velocities into the first two equations in (5.1) and if we pass to the new time variable $t:=\int_0^t v(s)ds$, we reduce the function $v(t)$ to 1 and obtain a system of the form (1.1), where the 2×2-matrix B is given by

$$\begin{aligned}
B_{11}&=\frac{\lambda_1(\lambda_2+\lambda_3)}{\lambda}\frac{\partial P_1}{\partial u_1}-\frac{\lambda_1\lambda_2}{\lambda}\frac{\partial P_2}{\partial u_1},\quad B_{12}=-\frac{\lambda_1\lambda_2}{\lambda}\frac{\partial P_2}{\partial u_2}+\frac{\lambda_1(\lambda_2+\lambda_3)}{\lambda}\frac{\partial P_1}{\partial u_2},\\
B_{21}&=\frac{\lambda_2(\lambda_1+\lambda_3)}{\lambda}\frac{\partial P_2}{\partial u_1}-\frac{\lambda_1\lambda_2}{\lambda}\frac{\partial P_1}{\partial u_1},\quad B_{22}=-\frac{\lambda_1\lambda_2}{\lambda}\frac{\partial P_1}{\partial u_2}+\frac{\lambda_2(\lambda_1+\lambda_3)}{\lambda}\frac{\partial P_2}{\partial u_2}.
\end{aligned} \tag{5.6}$$

Since u is a saturation vector, we must have $u\in\Delta$. We assume that the mobilities λ_i satisfy the natural conditions (see, e.g., [1])

$$\lambda_i\ge 0,\quad \lambda_i|_{u_i=0}=0,\quad i\in\{1,2,3\}. \tag{5.7}$$

Let us describe our method for defining the capillary pressures everywhere in Δ, assuming that they are known for two-phase flows. It is based on the assumption of a special structure for the matrix B given by (5.6).

From the preceding discussion, in the system of the form (1.1) for the saturations derived above, the functions $P_i(u_1,u_2)$, $i=1,2$, appear only in the matrix B. We now impose that

$$B_{21}=0,\quad B_{22}=B_{22}(u_2)\ge 0,\quad B_{11}\ge 0\quad\text{in}\quad\Delta. \tag{5.8}$$

The hypothesis (5.8) means that the first and the third phases are not responsible for the amount of diffusion in the equation for the second phase. The first two conditions in (5.8) read

$$A\frac{\partial P_1}{\partial u_1} = \frac{\partial P_2}{\partial u_1}, \quad \frac{\partial P_2}{\partial u_2} = A\frac{\partial P_1}{\partial u_2} + \frac{\lambda B_{22}}{\lambda_2(\lambda_1+\lambda_3)}, \quad A = \frac{\lambda_1}{\lambda_1+\lambda_3}. \tag{5.9}$$

We study these equations for $P_i(u_1, u_2)$ in the case when the mobilities λ_i are linear functions:

$$\lambda_i = k_i u_i, \quad k_i = \text{const}. \tag{5.10}$$

A symmetry group analysis (see [9]), performed for system (5.9), suggests to look for solutions of the form

$$P_i = q_i(\xi) + Q_i(u_2), \quad \xi = \frac{u_1}{1-u_2} \equiv \frac{u_1}{u_1+u_3}. \tag{5.11}$$

It follows from (5.9) that the functions q_i and Q_i solve the system

$$\begin{aligned} &q_2'(\xi) = q_1'(\xi)A(\xi), \quad A = \frac{k_1\xi}{(k_1-k_3)\xi+k_3}, \quad Q_1'(u_2) = -\frac{k_0 B_{22}(u_2)}{1-u_2}, \\ &Q_2'(u_2) = B_{22}(u_2)\left(\frac{1}{k_3(1-u_2)} + \frac{1}{k_2 u_2}\right), \quad k_0 = \frac{k_3-k_1}{k_1 k_3}. \end{aligned} \tag{5.12}$$

Assume that the capillary pressure $P_1(u)$ is a given function of u_1 at the part of the boundary of the triangle Δ where $u_2 = 0$:

$$P_1|_{u_2=0} = \pi_1(u_1).$$

Assume also that the capillary pressure $P_2(u)$ is a given function of u_2 at the edge where $u_1 = 0$ of the triangle Δ:

$$P_2|_{u_1=0} = \pi_2(u_2).$$

It follows from (5.11) that

$$\pi_1(u_1) = q_1(u_1) + Q_1(0), \quad \pi_2(u_2) = q_2(0) + Q_2(u_2).$$

It is naturally to set

$$q_1(\xi) = \pi_1(\xi), \quad Q_2(u_2) = \pi_2(u_2).$$

Then the other functions $Q_1(u_2)$ and $q_2(\xi)$ are defined from (5.12) as follows:

$$q_2(\xi) = \int_0^\xi A(\xi)\pi_1'(\xi)d\xi, \quad B_{22}(u_2) = \frac{k_2 k_3 u_2(1-u_2)\pi_2'(u_2)}{k_2 u_2 + k_3(1-u_2)},$$

$$Q_1'(u_2) = -\frac{k_0}{1-u_2} B_{22}(u_2).$$

Thus, we arrive at the formulas for the capillary pressures

$$\begin{cases} P_1(u_1,u_2) = \pi_1(\xi) - \int_0^{u_2} \frac{k_0 k_2 k_3 u_2 \pi_2'(u_2)}{k_2 u_2 + k_3(1-u_2)} du_2 + \text{const}, \\ P_2(u_1,u_2) = \int_0^\xi A(\xi)\pi_1'(\xi)\, d\xi + \pi_2(u_2) + \text{const}. \end{cases} \tag{5.13}$$

We call the procedure yielding formulas (5.13) the *method of physical interpolation* since these formulas define the capillary pressures P_1 and P_2 in Δ from their values when $u_2 = 0$ and $u_1 = 0$, respectively.

Substituting (5.10) and (5.13) in (5.6), we get

$$B_{11} = k_1(1 - A(\xi))\pi_1'(\xi), \quad B_{22} = \frac{k_2 k_3 u_2(1-u_2)\pi_2'(u_2)}{k_2 u_2 + k_3(1-u_2)},$$

$$B_{21} = 0, \quad B_{12} = \xi(B_{11} - B_{22}). \tag{5.14}$$

When

$$\pi_1'(\xi) \geq 0, \quad \pi_2'(u_2) \geq 0,$$

system (1.1) is parabolic in Δ but degenerate at $\partial\Delta$.

Finally, we remark that the matrix B given by (5.14) satisfies

$$B_{12}|_{u_1=0} = 0, \quad B_{21} \equiv 0, \quad (B_{11} - B_{12} - B_{22})|_{u_1+u_2=1} = 0.$$

More generally, it is easy to verify that (5.7) implies that the functions B_{ij} of the general capillarity matrix given by (5.6) satisfy

$$B_{12}|_{u_1=0} = 0, \quad B_{21}|_{u_2=0} = 0, \quad (B_{11} + B_{21} - B_{21} - B_{22})|_{u_1+u_2=1} = 0. \tag{5.15}$$

It is also immediate to verify that (5.7) implies that the functions f_i defined in (5.5), for the general three-phase flow model, satisfy

$$f_1|_{u_1=0} = 0, \quad f_2|_{u_2=0} = 0, \quad (f_1 + f_2)|_{u_1+u_2=1} = 1. \tag{5.16}$$

Both (5.15) and (5.16) imply sufficient conditions in order to have $u(x,t) \in \Delta$ for any smooth solution of (1.1), which behaves nicely for $x \in \partial\Omega$.

6. A degenerate problem. Here, we study a particular system arising in petroleum reservoir fluid flows. We assume that mobilities are linear functions

$$\lambda_i = k_i u_i, \quad i \in \{1,2,3\}, \quad k_1 = k_3,$$

and the capillary pressures are given by the formulas (5.13). In this case, the flow is governed by the degenerate parabolic system

$$u_{1t} + f_1(u_1,u_2)_x = (B_{11}(u_1,u_2)u_{1x})_x + (B_{12}(u_1,u_2)u_{2x})_x, \tag{6.1}$$

$$u_{2t} + f_2(u_2)_x = (B_{22}(u_2)u_{2x})_x,$$

with

$$f_1 = \frac{u_1}{ku_2 + 1}, \quad f_2 = (1+k)\frac{u_2}{ku_2+1}, \quad k = \frac{k_2}{k_1} - 1,$$

and B_{11}, B_{22}, B_{12} are given by (5.14), with $k_1 = k_3$ (see (1.16)). The first and second equations in (6.1) are coupled through the condition $u(t,x) \in \Delta$, which can be written as

$$0 \leq u_i(t,x) \leq 1, \quad u_2(t,x) \leq 1 - u_1(t,x). \tag{6.2}$$

We consider the Dirichlet initial boundary value problem

$$u|_{x=\pm 1} = u_\pm(t), \quad u|_{t=0} = u_0(x). \tag{6.3}$$

Proof of Theorem 1.3. Consider the approximate nondegenerate problem

$$\begin{cases} u_t + f(u)_x = (B^\nu(u)u_x)_x, \\ \nu u_\eta + u = u_b^\nu \quad \text{at} \quad |x| = 1, \quad u|_{t=0} = u_0^\nu(x), \end{cases} \tag{6.4}$$

with

$$B_{11}^\nu = \nu + \chi_\nu(u_2)B_{11}, \quad B_{22}^\nu = \nu + \chi_\nu(u_2)B_{22}, \quad B_{12}^\nu = \chi_\nu(u_2)\xi(B_{11}^\nu - B_{22}^\nu),$$

$$u_0^\nu \in H^{2+\beta}(\overline{\Omega}), \quad u_0^\nu(x) \in \Delta, \quad u_\pm^\nu \in H^{1+\beta/2}([0,T]), \quad u_\pm^\nu(t) \in \Delta,$$

$$\pm\nu u_0^\nu(\pm 1) + u_0^\nu(\pm 1) = u_\pm^\nu(0),$$

$$\|u_\pm^\nu - u_\pm\|_{W^{1,1}(0,T)} \to 0, \quad \|u_0^\nu - u_0\|_{L^2(\Omega)} \to 0, \quad \text{as} \quad \nu \downarrow 0.$$

Here, $\chi_\nu(u_2)$ is a smooth function such that

$$\chi_\nu(u_2) = 1 \quad \text{if} \quad 0 \le u_2 \le 1 - \nu, \quad \chi_\nu(u_2) = 0 \quad \text{if} \quad 1 - \frac{\nu}{2} \le u_2 \le 1.$$

Clearly, the matrix B^ν satisfies the hypotheses of Theorem 1.1, and so we have the unique solvability of problem (6.4). We also observe that, under the conditions of Theorem 1.3 on the data u_0^ν and $u_\pm^\nu$, any smooth solution of problem (6.4) satisfies the a priori estimate

$$\delta \le u_2(t,x) \le 1 - \delta. \tag{6.5}$$

With this estimate at hand, the matrix B^ν, for small ν, reads

$$B_{11}^\nu = \nu + B_{11}, \quad B_{22}^\nu = \nu + B_{22}, \quad B_{12}^\nu = \xi(B_{11}^\nu - B_{22}^\nu). \tag{6.6}$$

We then have $B_{22}^\nu(u^\nu) \ge \delta^2$ uniformly for $\nu \downarrow 0$. Thus, the constant c in Lemma 2.2 does not depend on ν, and

$$\|u_{2x}^\nu\|_{L^2(Q)} \le c, \quad \|u_{2t}^\nu\|_{L^2(0,T;W^{-1,2}(\Omega))} \le c \tag{6.7}$$

uniformly in ν.

Taking into account the last equality in (6.6) for the entry B_{12}^ν, one can calculate from (6.4) that the function $\xi = u_1^\nu/(1 - u_2^\nu)$ solves the problem

$$\xi_t + \frac{\xi_x}{ku_2^\nu + 1} = (B_{11}^\nu \xi_x)_x - \frac{\xi_x u_{2x}^\nu (B_{11}^\nu + B_{22}^\nu)}{1 - u_2^\nu},$$

$$\frac{\nu(1 - u_2^\nu)}{1 - u_\pm^\nu}\xi_\eta + \xi = \xi_\pm \quad \text{at} \quad x = \pm 1, \quad \xi|_{t=0} = \xi_0(x).$$

As we saw in section 5, the variable ξ appears naturally when one is looking for invariant solutions of the homogeneous system corresponding to (5.9). It enables us to apply the maximum principle. By the latter,

$$\delta \le \xi(t,x) \le 1 - \delta\ , \tag{6.8}$$

uniformly in ν. Now, it is a consequence of (6.5) and (6.8) that

$$\delta^2 \le u_1^\nu(t,x) \le (1-\delta)^2, \quad B_{11}^\nu \ge \delta^2.$$

By the same argument as in Lemma 2.2, we have

$$\|u_{1x}^\nu\|_{L^2(Q)} \le c, \quad \|u_{1t}^\nu\|_{L^2(0,T;W^{-1,2}(\Omega))} \le c \tag{6.9}$$

uniformly in ν.

Estimates (6.7) and (6.9) imply by the Aubin–Lions compactness theorem that there are a sequence $u^n \equiv u^{\nu_n}$ and a function u such as described in Theorem 1.3 and such that

$$u^n(t,x) \to u(t,x) \quad \text{a.e. in} \quad Q, \quad u_x^n \to u_x \quad \text{weakly in} \quad L^2(Q).$$

Clearly, u is a weak solution of problem (6.1)–(6.3). Theorem 1.3 is proved. ∎

REFERENCES

[1] M. B. Allen, J. B. Behie, and J. A. Trangenstein, *Multiphase Flows in Porous Media: Mechanics, Mathematics and Numerics*, Lecture Notes in Engrg. 34, Springer-Verlag, Berlin, 1988.

[2] H. Amann, *Dynamic theory of quasi-linear parabolic systems*, III. *Global existence*, Math. Z., 202 (1989), pp. 219–250.

[3] Z. Chen and R. E. Ewing, *Comparison of various formulations of three-phase flow in porous media*, J. Comput. Phys., 132 (1997), pp. 362–373.

[4] H. Frid and V. Shelukhin, *A quasi-linear parabolic system for three-phase capillary flow in porous media*, SIAM J. Math. Anal., 35 (2003), pp. 1029–1041.

[5] S. M. Hassanizadeh and W. G. Gray, *Thermodynamic basis of capillary pressure in porous media*, Water Resources Research, 29 (1993), pp. 3389–3405.

[6] O. A. Ladyzhenskaya, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasi-linear Equations of Parabolic Type*, AMS, Providence, RI, 1968.

[7] O. A. Ladyzhenskaya and N. N. Ural'ceva, *Linear and Quasilinear Equations of Elliptic Type*, Academic Press, New York, 1968.

[8] J. L. Lions, *Quelques Méthodes de Résolution des Problèmes aux Limites Non Linéaires*, Dunod, Paris, 1969.

[9] L. V. Ovsiannikov, *Group Analysis of Differential Equations*, Academic Press, New York, London, 1982.

[10] D. W. Peaceman, *Fundamentals of Numerical Reservoir Simulation*, Elsevier, Amsterdam, Oxford, New York, 1977.

[11] D. Serre, *Systèmes de Lois de Conservation*, Vol. II, Diderot Editeur, Arts et Sciences, Paris, New York, Amsterdam, 1996.

[12] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, Heidelberg, Berlin, 1983.

[13] H. L. Stone, *Probability model for estimating three-phase relative permeability*, J. Petroleum Technology, 249 (1970), pp. 214–218.

[14] A. N. Varchenko and A. F. Zazovskii, *Three-phase filtration of immiscible fluids*, Itogi Nauki Tek. Ser. Kompleksnie i spetsial'nie Razdely Mekhaniki, 4 (1991), pp. 98–154 (in Russian).

SIAM J. MATH. ANAL.
Vol. 36, No. 5, pp. 1426–1433

FINITE-TIME BLOWUP FOR WAVE EQUATIONS WITH A POTENTIAL*

BORISLAV YORDANOV† AND QI S. ZHANG†

Abstract. First we give a truly short proof of the major blowup result [T. C. Sideris, *J. Differential Equations*, 52 (1984), pp. 378–406] on higher-dimensional semilinear wave equations. Using this new method, we also establish blowup phenomenon for wave equations with a potential. This complements the recent interesting existence result by [V. Georgiev, C. Heiming, and H. Kubo, *Comm. Partial Differential Equations*, 26 (2001), pp. 2267–2303], where the blowup problem was left open.

Key words. semilinear wave equation, finite-time blowup

AMS subject classification. 35L70

DOI. 10.1137/S0036141004440198

1. Introduction.

We study the blowup of the solutions to the following semilinear wave equation:

$$\begin{cases} \Delta u - Vu - u_{tt} + |u|^p = 0 \quad \text{in} \quad \mathbf{R}^n \times (0,\infty), \\ u(x,0) = u_0(x), \quad u_t(x,0) = u_1(x) \quad \text{in} \quad \mathbf{R}^n, \end{cases} \tag{1.1}$$

where $\Delta = \Sigma_{i=1}^n \partial^2/\partial x_i^2$ is the Laplace operator and $V = V(x)$ is a potential. We consider dimensions $n \geq 3$ and exponents $p \in (1,\ p_c(n))$, where $p_c(n)$ is the positive root of the quadratic equation

$$(n-1)p^2 - (n+1)p - 2 = 0.$$

The number $p_c(n)$ is known as the critical exponent of the semilinear wave equation with $V = 0$ (see, e.g., [St]). The study of this equation has an interesting and exciting history. We will give only a brief summary here and refer the reader to [St], [L], [DL], and a recent paper [JZ] for details. Let the initial values be compactly supported and nonnegative. John [J] proved that for $n = 3$ and $1 < p < p_c(3)$, nontrivial solutions must blow up in finite time. If $p > p_c(3)$, global solutions exist for small initial values. Glassey [Gl1], [Gl2] established the same result in the case $n = 2$. Schaeffer [Sc] proved that the critical power $p = p_c(n)$ also belongs to the blowup case when $n = 2$, 3. In [GLS] the authors showed that when $p > p_c(n)$ and $n \geq 3$, (1.1) has global solutions for small initial values (see also [LS] and [T]). When $n \geq 4$ and $1 < p < p_c(n)$, the blowup result was proven by Sideris in [Si]. The proof is quite delicate, using sophisticated computation involving spherical harmonics. His proof was simplified in [R] and [JZ], where spherical harmonics still play an important role. In this paper we discover a truly short proof of the blowup result using only a simple test function. More importantly the proof carries over to the case when the potential V is positive. It is a well-known fact that the presence of potentials greatly increases the complexity of wave motion. In fact there is not much progress in either the existence or blowup

*Received by the editors January 26, 2004; accepted for publication (in revised form) August 2, 2004; published electronically March 25, 2005.
http://www.siam.org/journals/sima/36-5/44019.html
†Department of Mathematics, University of California Riverside, Riverside, CA 92521 (yordanov@math.ucr.edu, qizhang@math.ucr.edu).

problems in higher-dimensional cases of (1.1). In the three-dimensional case, it is known that there exist global small solutions when $p > p_c(3)$ and $V \in C_0^\infty(\mathbf{R}^3)$ is nonnegative; see [GHK]. In the same case, [ST] establishes, among other things, a blowup result for some $V \leq 0$. The current paper complements the result of [GHK] in dimension $n = 3$ and shows the blowup of solutions in all dimensions $n \geq 3$ when $1 < p < p_c(n)$ and V is a nonnegative potential satisfying the following condition: There exist two functions ϕ_0, $\phi_1 \in C^2(\mathbf{R}^n)$ such that

$$\begin{cases} \Delta\phi_0 - V\phi_0 = 0, & C_0^{-1} \leq \phi_0(x) \leq C_0, \\ \Delta\phi_1 - V\phi_1 = \phi_1, & 0 < \phi_1(x) \leq C_1(1+|x|)^{-(n-1)/2}e^{|x|}, \end{cases} \tag{1.2}$$

with positive constants C_0 and C_1. We show (see Lemma 3.1) that this condition is satisfied by nonnegative potentials under very mild additional assumptions about regularity and behavior at infinity.

We consider compactly supported nonnegative data $(u_0, u_1) \in H^1(\mathbf{R}^n) \times L^2(\mathbf{R}^n)$:

$$u_0(x) \geq 0, \quad u_1(x) \geq 0 \quad \text{a.e.}, \quad u_0(x) = u_1(x) = 0 \quad \text{for} \quad |x| > R. \tag{1.3}$$

Our main result is the following theorem.

THEOREM 1.1. *Let (u_0, u_1) satisfy (1.3) and V satisfy (1.2). Suppose that problem (1.1) has a solution $(u, u_t) \in C([0,T),\ H^1(\mathbf{R}^n) \times L^2(\mathbf{R}^n))$ such that*

$$\operatorname{supp}(u, u_t) \subset \{(x,t) : |x| \leq t + R\}.$$

If $1 < p < p_c(n)$, then $T < \infty$.

In particular, the conclusion holds if V is locally Hölder continuous and $0 \leq V(x) \leq \frac{C}{1+|x|^{2+\delta}}$ for some C, $\delta > 0$ and all $x \in \mathbf{R}^n$.

When $V = 0$, we choose the functions

$$\begin{cases} \phi_0(x) = 1, \\ \phi_1(x) = \int_{S^{n-1}} e^{x\cdot\omega} d\omega, \quad \phi_1(x) \sim C_n|x|^{-(n-1)/2}e^{|x|} \quad \text{as} \quad |x| \to \infty. \end{cases}$$

Since condition (1.2) holds, we can apply Theorem 1.1 and deduce the well-known results of John [J] and Sideris [Si].

The proof of Theorem 1.1 is given in sections 2 and 3. To outline the method, we introduce

$$\begin{aligned} F_0(t) &= \int u(x,t)\phi_0(x)dx, \\ F_1(t) &= \int u(x,t)\psi_1(x,t)dx, \qquad \psi_1(x,t) = \phi_1(x)e^{-t}. \end{aligned} \tag{1.4}$$

The assumptions on u imply that $F_0(t)$ and $F_1(t)$ are well-defined C^2-functions for all t. By a standard procedure, we derive a nonlinear differential inequality for $F_0(t)$. We also derive a linear differential inequality for $F_1(t)$ and combine these to obtain a polynomial lower bound on $F_0(t)$ as $t \to \infty$. Theorem 1.1 is a consequence of the lower bound and the blowup result about nonlinear differential inequalities in Lemma 2.1.

In section 3 we prove the existence of ϕ_0 and ϕ_1 in (1.2) when V is locally Hölder continuous and $0 \leq V(x) \leq \frac{C}{1+|x|^{2+\delta}}$ for some $C, \delta > 0$ and all $x \in \mathbf{R}^n$. This relies on a latest sharp estimate of heat kernels with a potential.

2. Proof of Theorem 1.1. We will use the following well-known ODE result from, e.g., [Si, p. 386] to show that $F_0(t)$ in (1.4) blows up in finite time.

LEMMA 2.1. *Let $p > 1$, $a \geq 1$, and $(p-1)a > q-2$. If $F \in C^2([0,T))$ satisfies*
(a) $F(t) \geq K_0(t+R)^a$,
(b) $\frac{d^2F(t)}{dt^2} \geq K_1(t+R)^{-q}[F(t)]^p$,
with some positive constants K_0, K_1, and R, then $T < \infty$.

To show that F_0 satisfies the above differential inequalities for suitable a, q, we multiply (1.1) by ϕ_0 and integrate over $\mathbf{R}^n$. Condition (1.2) on ϕ_0 yields

$$\frac{d^2F_0(t)}{dt^2} = \int |u(x,t)|^p \phi_0(x)dx.$$

Note that for a fixed t, $u(\cdot,t) \in H_0^1(D_t)$ where D_t is the support of $u(\cdot,t)$. Hence the above equality is justified using integration by parts.

Estimating the right side by the Hölder inequality, we have

$$\int |u(x,t)|^p \phi_0(x)dx \geq \frac{\left|\int u(x,t)\phi_0(x)dx\right|^p}{\left(\int_{|x|\leq t+R} \phi_0(x)dx\right)^{p-1}}.$$

By condition (1.2),

$$\int_{|x|\leq t+R} \phi_0(x)dx \leq C_0 \mathrm{vol}\{x : |x| < t+R\} = C_0 \mathrm{vol}(\mathbf{B}^n)(t+R)^n.$$

Thus, we obtain the differential inequality

$$\frac{d^2F_0(t)}{dt^2} \geq L_1(t+R)^{-n(p-1)}|F_0(t)|^p \tag{2.1}$$

with some $L_1 > 0$.

To show that F_0 admits the lower bound in Lemma 2.1(a), we relate d^2F_0/dt^2 to F_1 using again (1.1) and the Hölder inequality:

$$\frac{d^2F_0(t)}{dt^2} = \int |u(x,t)|^p \phi_0(x)dx \geq \frac{\left|\int u(x,t)\psi_1(x,t)dx\right|^p}{\left(\int_{|x|\leq t+R} [\phi_0(x)]^{-1/(p-1)}[\psi_1(x,t)]^{p/(p-1)}dx\right)^{p-1}}.$$

By (1.2), the last inequality becomes

$$\frac{d^2F_0(t)}{dt^2} \geq \frac{C_0\,|F_1(t)|^p}{\left(\int_{|x|\leq t+R} [\psi_1(x,t)]^{p/(p-1)}dx\right)^{p-1}}. \tag{2.1$'$}$$

The following lemmas estimate the numerator and denominator, respectively, and provide a lower bound on d^2F_0/dt^2.

LEMMA 2.2. *Let V satisfy* (1.2) *and (u_0,u_1) satisfy* (1.3). *Assume that u meets the conditions of Theorem* 1.1. *Then for all $t \geq 0$,*

$$F_1(t) \geq \frac{1}{2}(1-e^{-2t})\int [u_0(x)+u_1(x)]\phi_1(x)dx + e^{-2t}\int u_0(x)\phi_1(x)dx \geq c > 0.$$

LEMMA 2.3. *Let $p > 1$. Assume that ϕ_0 and ϕ_1 satisfy condition* (1.2). *Then for all $t \geq 0$,*

$$\int_{|x|\leq t+R} [\psi_1(x,t)]^{p/(p-1)}dx \leq C(t+R)^{n-1-(n-1)p'/2},$$

where $p' = p/(p-1)$.

Taking the two lemmas for granted, we combine them with (2.1′) to obtain

$$\frac{d^2F_0(t)}{dt^2} \geq L_2(t+R)^{n-1-(n-1)p/2}, \quad t \geq 0,$$

where $L_2 > 0$. Integrating twice, we have the final estimate

$$F_0(t) \geq L_0(t+R)^{n+1-(n-1)p/2} + \frac{dF_0(0)}{dt}t + F_0(0)$$

with some $L_0 > 0$. When $1 < p < p_c(n)$, it is easy to check that $n+1-(n-1)p/2 > 1$. Hence the following estimate is valid when t is large:

$$F_0(t) \geq L_0(t+R)^{n+1-(n-1)p/2}. \tag{2.2}$$

Estimates (2.1) and (2.2) and Lemma 2.1 with parameters

$$a \equiv n+1-(n-1)p/2 \quad \text{and} \quad q \equiv n(p-1)$$

imply Theorem 1.1 for all exponents p such that

$$(p-1)(n+1-(n-1)p/2) > n(p-1)-2 \quad \text{and} \quad p > 1.$$

It is easy to see that the solution set is $p \in (1,\, p_c(n))$. The proof of Theorem 1.1 is complete, assuming Lemmas 2.2 and 2.3 and the validity of (1.2).

Proof of Lemma 2.2. We multiply (1.1) by a test function $\psi \in C^2(\mathbf{R}^{n+1})$ and integrate over $\mathbf{R}^n \times [0,\, t]$:

$$\begin{aligned} &\int_0^t \int u(\Delta\psi - V\psi - \psi_{ss})dxds + \int_0^t \int |u|^p\psi\, dxds \\ &\qquad = \int (u_s\psi - u\psi_s)dx|_{s=t} - \int (u_s\psi - u\psi_s)dx|_{s=0}. \end{aligned} \tag{2.3}$$

We will apply this identity to $\psi = \psi_1$. Notice that for a fixed t, $u(\cdot,t) \in H_0^1(D_t)$ where D_t is the support of $u(\cdot,t)$. Hence all terms involving lateral boundary vanish during integration by parts. Notice also that

$$(\psi_1)_t = -\psi_1, \quad \Delta\psi_1 - V\psi_1 - (\psi_1)_{tt} = 0,$$

and

$$\begin{aligned} \int (u_s\psi_1 - u(\psi_1)_s)dx|_{s=t} &= \int (u_t\psi_1 + u(\psi_1)_t)dx - 2\int u(\psi_1)_t dx \\ &= \frac{d}{dt}\int u\psi_1 dx + 2\int u\psi_1 dx. \end{aligned}$$

Hence, (2.3) becomes

$$\frac{dF_1(t)}{dt} + 2F_1(t) = \int [u_0(x) + u_1(x)]\phi_1(x)dx + \int_0^t \int |u(x,s)|^p\psi_1(x,s)dxds.$$

Since $\psi_1 > 0$, we conclude that

$$\frac{dF_1(t)}{dt} + 2F_1(t) \geq \int [u_0(x) + u_1(x)]\phi_1(x)dx.$$

We multiply by e^{2t} and integrate on $[0,\, t]$. Then

$$e^{2t}F_1(t) - F_1(0) \geq \frac{1}{2}(e^{2t}-1)\int [u_0(x)+u_1(x)]\phi_1(x)dx.$$

Dividing through by e^{2t}, we obtain the lower bound in Lemma 2.2. □

Proof of Lemma 2.3. Let $I(t)$ be the integral in Lemma 2.3. Condition (1.2) shows that

$$I(t) \leq \text{area}(\mathbf{S}^{n-1})C_1^{p/(p-1)}e^{-p't}\int_0^{t+R}(1+r)^{-(n-1)p'/2}e^{p'r}r^{n-1}dr,$$

where $p' = p/(p-1)$. Since $r < r+1$, it is sufficient to show that

$$I(t) \leq Ce^{-p't}\int_0^{t+R}(1+r)^{n-1-(n-1)p'/2}e^{p'r}dr \leq C(t+R)^{n-1-(n-1)p'/2}.$$

This estimate is evident after splitting the last integral into two parts:

$$\begin{aligned}\int_0^{(t+R)/2}(1+r)^{n-1-(n-1)p'/2}e^{rp'}dr &\leq (1+t+R)^{q_1}\int_0^{(t+R)/2}e^{p'r}dr\\ &\leq \frac{e^{p'R/2}}{p'}(1+t+R)^{q_1}e^{p't/2},\end{aligned}$$

where $q_1 = \max(0,\, n-1-(n-1)p'/2)$, and

$$\begin{aligned}\int_{(t+R)/2}^{t+R}(1+r)^{n-1-(n-1)p'/2}e^{p'r}dr &\leq 2^{-q_2}(1+t+R)^{n-1-(n-1)p'/2}\int_{(t+R)/2}^{t+R}e^{p'r}dr\\ &\leq \frac{2^{-q_2}e^{p'R}}{p'}(1+t+R)^{n-1-(n-1)p'/2}e^{p't},\end{aligned}$$

where $q_2 = \min(0,\, n-1-(n-1)p'/2)$. This proves Lemma 2.3. □

To complete the proof of Theorem 1.1, it remains to prove Lemma 3.1. In the special case $V = 0$, the next section is redundant.

3. Existence of the two functions in (1.2). In this section we prove the following lemma.

LEMMA 3.1. *Suppose V is locally Hölder continuous and $0 \leq V(x) \leq \frac{C}{1+|x|^{2+\delta}}$ for some $C, \delta > 0$ and all $x \in \mathbf{R}^n$. Then there exist two functions ϕ_0 and ϕ_1 satisfying* (1.2)*, i.e.,*

$$\begin{cases}\Delta\phi_0 - V\phi_0 = 0, & C_0^{-1} \leq \phi(x) \leq C_0,\\ \Delta\phi_1 - V\phi_1 = \phi_1, & 0 < \phi_1(x) \leq C_1(1+|x|)^{-(n-1)/2}e^{|x|}.\end{cases}$$

Proof. Let H_0 and H be the fundamental solutions of

$$\Delta u - u - u_t = 0, \qquad \Delta u - u - Vu - u_t = 0$$

in $\mathbf{R}^n \times (0,\infty)$, respectively. Then $H_0 = e^{-t}G_0$ and $H = e^{-t}G$, where G_0 and G are the fundamental solution of

$$\Delta u - u_t = 0, \qquad \Delta u - Vu - u_t = 0.$$

By Theorem 1.1(a) and Remark 1.1 in [Z1], there exists a positive constant c such that

$$cG_0(x,t;y,0) \le G(x,t;y,0) \le G_0(x,t;y,0) = \frac{c_n}{t^{n/2}} e^{-\frac{|x-y|^2}{4t}}$$

for all $x, y \in \mathbf{R}^n$ and $t > 0$. We should mention that the global lower bound is nontrivial since one needs to keep the exact coefficient $1/4$ in each exponential term.

Hence we have the following global bounds:

$$cH_0(x,t;y,0) \le H(x,t;y,0) \le H_0(x,t;y,0). \tag{3.1}$$

Let μ_0 be a positive solution of $\Delta\mu_0 - \mu_0 = 0$ such that

$$\mu_0(x) \sim e^{|x|}/(1+|x|)^{(n-1)/2}.$$

The existence of such μ_0 is well known and is explained in the introduction. Consider the function

$$u(x,t) \equiv \int_{\mathbf{R}^n} H(x,t;y,0)\mu_0(y)dy. \tag{3.2}$$

Since for fixed (x,t), $H(x,t;y,0)$ decays super-exponentially near infinity, the above integral is well defined. Moreover, u is a solution to

$$\Delta u - u - Vu - u_t = 0. \tag{3.3}$$

By (3.1) and (3.2) we have

$$c\int_{\mathbf{R}^n} H_0(x,t;y,0)\mu_0(y)dy \le u(x,t) \le \int_{\mathbf{R}^n} H_0(x,t;y,0)\mu_0(y)dy.$$

Since $\Delta\mu_0 - \mu_0 = 0$, it is clear from differentiation that

$$\mu_0(x) = \int_{\mathbf{R}^n} H_0(x,t;y,0)\mu_0(y)dy,$$

even though the right-hand side apparently depends on time. Indeed,

$$\begin{aligned}\frac{\partial}{\partial t}\int_{\mathbf{R}^n} H_0(x,t;y,0)\mu_0(y)dy &= -\int_{\mathbf{R}^n} (\Delta_y - 1)H_0(x,t;y,0)\mu_0(y)dy \\ &= -\int_{\mathbf{R}^n} H_0(x,t;y,0)(\Delta_y - 1)\mu_0(y)dy = 0.\end{aligned}$$

Here we observe that integration by parts is legitimate since, for fixed $t > 0$ and x, $H_0(x,t;y,0)$ has superexponential decay near infinity while μ_0 only grows exponentially.

Hence

$$c\mu_0(x) \le u(x,t) \le \mu_0(x). \tag{3.4}$$

Differentiating (3.3) with respect to t, we obtain

$$\begin{cases}\Delta u_t - u_t - Vu_t - (u_t)_t = 0, \quad (x,t) \in \mathbf{R}^n \times (0,\infty), \\ u_t|_{t=0} = \Delta\mu_0 - \mu_0 - V\mu_0 \le 0.\end{cases}$$

Here we remark that under our assumption that V is locally Hölder continuous, it is not clear whether u_{tt} exists. However, we can work on the finite difference of u_t and use a standard approximation argument to achieve the same result.

By the maximum principle, we know that $u_t(x,t) \le 0$ everywhere. This and (3.4) show that $u(x,t)$ converges to a function $\phi_1 = \phi_1(x)$ as $t \to \infty$. Moreover,

$$c\mu_0(x) \le \phi_1(x) \le \mu_0(x). \tag{3.5}$$

We are going to show that

$$\Delta\phi_1 - \phi_1 - V\phi_1 = 0. \tag{3.6}$$

To this end, let us consider the function $w = w(x,t) = \int_t^{t+1} u(x,s)ds$. Direct computation shows that

$$\Delta w(x,t) - w(x,t) - V(x)w(x,t) = u(x,t+1) - u(x,t).$$

It is also clear that $w(x,t) \to \phi_1(x)$ when $t \to \infty$. Let $\eta = \eta(x)$ be any function in $C_0^\infty(\mathbf{R}^n)$. Then we obtain

$$\begin{aligned}&\int_{\mathbf{R}^n} [w(x,t)\Delta\eta(x) - w(x,t)\eta(x) - V(x)w(x,t)\eta(x)]dx \\ &\quad = \int_{\mathbf{R}^n} [u(x,t+1) - u(x,t)]\eta(x)dx.\end{aligned}$$

Letting $t \to \infty$, we have

$$\int_{\mathbf{R}^n} [\phi_1(x)\Delta\eta(x) - \phi_1(x)\eta(x) - V(x)\phi_1(x)\eta(x)]dx = 0.$$

Since η is arbitrary and ϕ_1 is locally bounded, we know that ϕ_1 is a classical solution to (3.6), which also satisfies (3.5). This proves the existence of ϕ_1 in (1.2). The existence of ϕ_0 under our assumption is well known (see, e.g., [Z2, Theorem B]). In fact it can be proven by exactly the same method except that we drop the term $-u$ everywhere. □

Remark 3.1. The decay condition for V in Lemma 3.1 can be generalized. In [Z1], a necessary and sufficient condition for the validity of the sharp comparison result right before (3.1) was found for all nonnegative V. This class of V resembles the Kato class in mathematical physics. It overlaps with $L^{n/2}(\mathbf{R}^n)$. But they are not the same.

REFERENCES

[DL] K. Deng and H. A. Levine, *The role of critical exponents in blow-up theorems: The sequel*, J. Math. Anal. Appl., 243 (2000), pp. 85–126.

[GHK] V. Georgiev, C. Heiming, and H. Kubo, *Supercritical semilinear wave equation with non-negative potential*, Comm. Partial Differential Equations, 26 (2001), pp. 2267–2303.

[Gl1] R. T. Glassey, *Finite-time blow-up for solutions of nonlinear wave equations*, Math. Z., 177 (1981), pp. 323–340.

[Gl2] R. T. Glassey, *Existence in the large for* $\Box u = F(u)$ *in two space dimensions*, Math. Z., 178 (1981), pp. 233–261.

[GLS] V. Georgiev, H. Lindblad, and C. D. Sogge, *Weighted Strichartz estimates and global existence for semilinear wave equations*, Amer. J. Math., 119 (1997), pp. 1291–1319.

[J] F. John, *Blow-up of solutions of nonlinear wave equations in three space dimensions*, Manuscripta Math., 28 (1979), pp. 235–268.

[JZ] H. Jiao and Z. Zhou, *An elementary proof of the blow-up for semilinear wave equation in high space dimensions*, J. Differential Equations, 189 (2003), pp. 355–365.

[L] H. A. Levine, *The role of critical exponents in blowup theorems*, SIAM Rev., 32 (1990), pp. 262–288.

[LS] H. Lindblad and C. Sogge, *Long-time existence for small amplitude semilinear wave equations*, Amer. J. Math., 118 (1996) pp. 1047–1135.

[R] M. A. Rammaha, *Finite-time blow-up for nonlinear wave equations in high dimensions*, Comm. Partial Differential Equations, 12 (1987), pp. 677–700.

[Sc] J. Schaeffer, *The equation* $\Box u = |u|^p$ *for the critical value of* p, Proc. Roy. Soc. Edinburgh Sect. A, 101 (1985), pp. 31–44.

[Si] T. C. Sideris, *Nonexistence of global solutions to semilinear wave equations in high dimensions*, J. Differential Equations, 52 (1984), pp. 378–406.

[St] W. A. Strauss, *Nonlinear Wave Equations*, CBMS Reg. Conf. Ser. Math. 73, AMS, Providence, RI, 1989.

[ST] W. A. Strauss and K. Tsutaya, *Existence and blow up of small amplitude nonlinear waves with a negative potential*, Discrete Contin. Dynam. Systems, 3 (1997), pp. 175–188.

[T] D. Tataru, *Strichartz estimates in the hyperbolic space and global existence for the semilinear wave equation*, Trans. Amer. Math. Soc., 353 (2001), pp. 795–807.

[Z1] Q. S. Zhang, *A sharp comparison result concerning Schrödinger heat kernels*, Bull. London Math. Soc., 35 (2003), pp. 461–472.

[Z2] Q. S. Zhang, *An optimal parabolic estimate and its applications in prescribing scalar curvature on some open manifolds with* Ricci ≥ 0, Math. Ann., 316 (2000), pp. 703–731.

SIAM J. MATH. ANAL.
Vol. 36, No. 5, pp. 1434–1443

EXISTENCE OF TRAVELLING WAVES IN DISCRETE SINE-GORDON RINGS*

GUY KATRIEL†

Abstract. We prove existence results for travelling waves in discrete, damped, dc-driven sine-Gordon equations with periodic boundary conditions. Methods of nonlinear functional analysis are employed. Some unresolved questions are raised.

Key words. discrete sine-Gordon equation, Frenkel–Kontorova model, travelling waves

AMS subject classifications. 34C15, 34C60, 37N20, 78A55

DOI. 10.1137/S0036141004440174

1. Introduction. The damped, dc-driven discrete sine-Gordon equation, known also as the driven Frenkel–Kontorova model, with periodic boundary conditions, arises as a model of many physical systems, including circular arrays of Josephson junctions, the motions of dislocations in a crystal, the adsorbate layer on the surface of a crystal, ionic conductors, glassy materials, charge-density wave transport, sliding friction, as well as the mechanical interpretation as a model for a ring of pendula coupled by torsional springs (we refer to the reader [8, 10, 11] and the references therein). This model has thus become a fundamental one for nonlinear physics, and has been the subject of many theoretical, numerical, and experimental studies. The system of equations is

$$\phi_j'' + \Gamma\phi_j' + \sin(\phi_j) = F + K[\phi_{j+1} - 2\phi_j + \phi_{j-1}] \quad \forall j \in \mathbb{Z} \tag{1}$$

with the parameters $\Gamma > 0, K > 0, F > 0$, with the periodic boundary condition

$$\phi_{j+n}(t) = \phi_j(t) + 2\pi m \quad \forall j \in \mathbb{Z}, \tag{2}$$

where $m \geq 1$ (we note that in view of the boundary conditions we are really dealing with an n-dimensional system of ODE's rather than an infinite-dimensional one). In numerical simulations, as well as in experimental work on systems modelled by (1), (2), it is observed that solutions often converge to a travelling wave: a solution satisfying

$$\phi_j(t) = f\Big(t + j\frac{m}{n}T\Big), \tag{3}$$

where the waveform $f : \mathbb{R} \to \mathbb{R}$ is a function satisfying

$$f(t+T) = f(t) + 2\pi \quad \forall t \in \mathbb{R}. \tag{4}$$

The velocity of the travelling wave is given by

$$v = \frac{2\pi}{T}. \tag{5}$$

*Received by the editors January 26, 2004; accepted for publication (in revised form) June 11, 2004; published electronically March 25, 2005. This research was partially supported by the Edmund Landau Center for Research in Mathematical Analysis and Related Areas, sponsored by the Minerva Foundation (Germany).

http://www.siam.org/journals/sima/36-5/44017.html

†Einstein Institute of Mathematics, The Hebrew University of Jerusalem, Jerusalem, 91904, Israel (haggaik@wowmail.com).

However, as has been pointed out in [10], even the *existence* of such a solution has not been proven, except for the case of small K in which existence of a travelling wave for some values of F had been proven in [4].

In the "superdamped" case, in which the second-derivative term in (1) is removed, there are very satisfactory results about existence and also global stability of travelling waves (see [1, Theorem 2],). Such results rely strongly on monotonicity arguments. Recently Baesens and MacKay [2] have managed to extend these arguments to the "overdamped" case of (1): their results apply when

$$\Gamma > 2\sqrt{2K+1}, \tag{6}$$

and say that there exists a travelling-wave solution which is globally stable if and only if (1), (2) does not have stationary solutions. We do not know whether in general the nonexistence of stationary solutions implies the existence of a travelling wave.

We note that a function f is a waveform if and only if it satisfies (4) and

$$f''(t) + \Gamma f'(t) + \sin(f(t)) = F + K\Big[f\Big(t + \frac{m}{n}T\Big) - 2f(t) + f\Big(t - \frac{m}{n}T\Big)\Big]. \tag{7}$$

Here we obtain several existence results for travelling waves under conditions not covered by the existing work, described above.

THEOREM 1. *Fixing any $\Gamma > 0$ and $K > 0$, and given any velocity $v > 0$, there exists a travelling-wave solution of* (1), (2) *with velocity v for an appropriate $F > 0$.*

THEOREM 2. *For any $F > 1$ there exists a travelling-wave solution of* (1), (2).

THEOREM 3. *Assume that n does not divide m. Fixing any $\tilde{F} > 0$ and $\tilde{\Gamma} > 0$, for all K sufficiently large there exists a travelling-wave solution of* (1), (2) *for any $F \geq \tilde{F}$ and $\Gamma \geq \tilde{\Gamma}$.*

THEOREM 4. *Fixing any $\tilde{F} > 0$ and $\tilde{K} > 0$, for all $\Gamma > 0$ sufficiently small there exists a travelling-wave solution of* (1), (2) *for any $F \geq \tilde{F}$ and $0 < K \leq \tilde{K}$.*

We remark that the assumption that n does not divide m cannot be removed from Theorem 3, since if n divides m the coupling term vanishes and (7), (4) reduce to the equation of a running solution of a dc-forced pendulum, which, fixing $\Gamma > 0$, is known to have a solution only when F exceeds a positive critical value [5].

It is interesting to note that Theorem 4 demonstrates that for some parameter ranges, there is coexistence of stationary solutions and travelling waves of (1), (2). Indeed, it is well known [3] that, fixing K, for F sufficiently small, there exist stationary solutions of (1), (2), and these obviously do not depend on Γ. Hence we can take $\Gamma > 0$ sufficiently small so that Theorem 4 ensures also the existence of travelling waves. This phenomenon cannot happen in the superdamped case, nor in the overdamped case in which (6) holds, since in these cases existence of stationary solutions implies that the ω-limit set of every orbit is contained in the set of stationary solutions [1, 2].

Along the way we will prove the following proposition.

PROPOSITION 5. *An upper bound for the velocity v of any travelling wave is given by*

$$v < \frac{F}{\Gamma}, \tag{8}$$

and a lower bound, in the case $F > 1$, is given by

$$v > \frac{F-1}{\Gamma}. \tag{9}$$

In the next section we prove the results stated above. In section 3 we discuss the meaning of our results in connection with existing numerical studies of the discrete sine-Gordon equation, and point out some further mathematical questions which arise from our results and remain open.

2. Proofs of the results. Our method of proof involves reformulating the problem as a fixed-point problem in a Banach space, and applying results of nonlinear functional analysis. Our approach is thus close in spirit to [6], which deals with travelling waves in globally coupled Josephson junctions.

We transform the problem (4), (7) by setting

$$f(t) = u(vt) + vt,$$

where the wave velocity v is defined by (5) and u satisfies

$$u(z + 2\pi) = u(z) \quad \forall z \in \mathbb{R}. \tag{10}$$

Equation (7) can then be written as

$$\begin{aligned} &v^2 u''(z) + \Gamma v u'(z) + \sin(z + u(z)) \\ &\qquad = F - \Gamma v + K\Big[u\Big(z + 2\pi\frac{m}{n}\Big) - 2u(z) + u\Big(z - 2\pi\frac{m}{n}\Big)\Big]. \end{aligned} \tag{11}$$

Dividing by v^2 and setting

$$\lambda = \frac{1}{v},$$

we rewrite (11) in the form

$$\begin{aligned} &u''(z) + \lambda\Gamma u'(z) + \lambda^2 \sin(z + u(z)) \\ &\qquad = \lambda^2 F - \lambda\Gamma + \lambda^2 K\Big[u\Big(z + 2\pi\frac{m}{n}\Big) - 2u(z) + u\Big(z - 2\pi\frac{m}{n}\Big)\Big]. \end{aligned} \tag{12}$$

We note that if $u(z)$ satisfies (10), (12), then so does $\tilde{u}(z) = u(z + c) + c$, for any $c \in \mathbb{R}$. Thus by adjusting c we may assume that u satisfies

$$\int_0^{2\pi} u(s)ds = 0. \tag{13}$$

We note now that if u satisfies (10),(12), then by integrating both sides of (12) over $[0, 2\pi]$ we obtain

$$F = \frac{\Gamma}{\lambda} + \frac{1}{2\pi}\int_0^{2\pi} \sin(s + u(s))ds. \tag{14}$$

We can thus rewrite (12) as

$$\begin{aligned} u''(z) + \lambda\Gamma u'(z) + \lambda^2 \sin(z + u(z)) &= \lambda^2 \frac{1}{2\pi}\int_0^{2\pi} \sin(s + u(s))ds \\ &+ \lambda^2 K\Big[u\Big(z + 2\pi\frac{m}{n}\Big) - 2u(z) + u\Big(z - 2\pi\frac{m}{n}\Big)\Big]. \end{aligned} \tag{15}$$

Conversely, if u satisfies (14) and (15), then it satisfies (12). We have thus reformulated our problem as: find solutions (λ, u) of (10), (13), (14), (15). The idea now is to

consider λ as a *parameter* in (15) and try to find solutions u satisfying (10), (13), (15), and then substitute λ and u into (14) to obtain the corresponding value of F. This is the same idea as used in the numerical method presented in [8], but here it is used as part of existence proofs. We have the following claim.

PROPOSITION 6. *For any value λ, there exists a solution u of* (15) *satisfying* (10), (13).

We note that this proposition immediately implies Theorem 1, since given any $v > 0$ it shows that we can solve (15) with $\lambda = \frac{1}{v}$, hence obtain a travelling wave with velocity v, for the value of F given by (14).

To prove Proposition 6 we will use the Schauder fixed-point theorem. We denote by X and Y the Banach spaces of real-valued functions:

$$X = \left\{ u \in H^2[0, 2\pi] \mid u(0) = u(2\pi),\ \ u'(0) = u'(2\pi),\ \ \int_0^{2\pi} u(s)ds = 0 \right\},$$

$$Y = \left\{ u \in L^2[0, 2\pi] \mid \int_0^{2\pi} u(s)ds = 0 \right\},$$

with the norm

$$\|u\|_Y = \Big(\frac{1}{2\pi} \int_0^{2\pi} (u(s))^2 ds\Big)^{\frac{1}{2}},$$

and by $L_\lambda : X \to Y$ the linear mapping

$$L_\lambda(u)(z) = u''(z) + \lambda\Gamma u'(z) - \lambda^2 K\Big[u\Big(z + 2\pi\frac{m}{n}\Big) - 2u(z) + u\Big(z - 2\pi\frac{m}{n}\Big)\Big].$$

We want to show that this mapping is invertible and to derive an upper bound for the norm of its inverse. Noting that any $u \in X$ can be decomposed in a Fourier series $u(z) = \sum_{l\neq 0} a_l e^{ilz}$ (with $a_{-l} = \overline{a_l}$), we apply L_λ to the Fourier elements, obtaining

$$L_\lambda(e^{ilz}) = \mu_l e^{ilz},$$

where

$$\mu_l = -l^2 - 2K\lambda^2\Big(\cos\Big(\frac{2\pi ml}{n}\Big) - 1\Big) + \lambda l\Gamma i,$$

so that

$$|\mu_l| = \Big[\Big(l^2 + 2K\lambda^2\Big(\cos\Big(\frac{2\pi ml}{n}\Big) - 1\Big)\Big)^2 + \lambda^2 l^2\Gamma^2\Big]^{\frac{1}{2}}, \tag{16}$$

which does not vanish if $\Gamma > 0$. Thus the mapping L_λ has an inverse satisfying $L_\lambda^{-1}(e^{ilz}) = \frac{1}{\mu_l} e^{ilz}$. Since L_λ^{-1} takes Y onto X, and since X is compactly embedded in Y, we may consider L_λ^{-1} as a mapping from Y to itself, in which case it is a *compact* mapping. We also note, using (16), that

$$\|L_\lambda^{-1}\|_{Y,Y} \le \max_{l\ge 1} \frac{1}{|\mu_l|} \le \frac{1}{\lambda\Gamma}. \tag{17}$$

We also define the nonlinear operator $N : Y \to Y$ by

$$N(u)(z) = -\sin(z + u(z)) + \frac{1}{2\pi}\int_0^{2\pi} \sin(s + u(s))ds.$$

It is easy to see that N is continuous, and that the range of N is contained in a bounded ball in Y, indeed we have

$$\|\sin(z+u(z))\|_{L_2} = \Big(\frac{1}{2\pi}\int_0^{2\pi} (\sin(s+u(s)))^2 ds\Big)^{\frac{1}{2}} \leq 1,$$

and since $N(u)$ is the orthogonal projection of $-\sin(z+u(z))$ into Y, we have

$$\|N(u)\|_Y \leq 1 \quad \forall u \in Y. \tag{18}$$

We can now rewrite the problem (10), (13), (15) as the fixed-point problem

$$u = \lambda^2 L_\lambda^{-1} \circ N(u). \tag{19}$$

The operator on the right-hand side is compact by the compactness of L_λ^{-1}, and has a bounded range by (17), (18), so that Schauder's fixed-point theorem implies that (19) has a solution, proving Proposition 6 (we note that by a simple bootstrap argument a solution in Y is in fact smooth). Moreover, defining

$$\Sigma = \{(\lambda, u) \in [0,\infty) \times Y \mid u = \lambda^2 L_\lambda^{-1} \circ N(u)\},$$

Rabinowitz's continuation theorem [7] implies that the connected component of Σ containing $(\lambda, u) = (0,0)$, which we denote by C, is unbounded in $[0,\infty) \times Y$. Since for any $\lambda_0 > 0$ we have, from (18), (19), the bound $\|u\|_Y \leq \frac{\lambda_0}{\Gamma}$ for solutions (λ, u) of (19) with $\lambda \in [0, \lambda_0]$, the unboundedness of the set C must be in the λ-direction, that is, there exists $(\lambda, u) \in C$ with arbitrarily large values of λ.

We can now consider the right-hand side of (14) as a functional on $(0,\infty) \times Y$:

$$\Phi(\lambda, u) = \frac{\Gamma}{\lambda} + \frac{1}{2\pi}\int_0^{2\pi} \sin(s+u(s))ds, \tag{20}$$

and our strategy in proving Theorems 2–4 is to prove solvability of the equation

$$\Phi(\lambda, u) = F, \quad (\lambda, u) \in \Sigma \tag{21}$$

(in fact we shall prove solvability of (21) with Σ replaced by $C \subset \Sigma$). We note that by the boundedness of the sine function we have

$$\lim_{\lambda \to 0+,\ (\lambda,u)\in C} \Phi(\lambda, u) = +\infty, \tag{22}$$

$$\limsup_{\lambda \to +\infty,\ (\lambda,u)\in C} \Phi(\lambda, u) \leq 1. \tag{23}$$

Since C is a connected set and Φ is continuous, (22) implies the following proposition.

PROPOSITION 7. *For any F satisfying*

$$F > \underline{F} \equiv \inf_{(\lambda,u)\in C} \Phi(\lambda, u), \tag{24}$$

there exists a travelling wave.

Since (23) implies that $\underline{F} \leq 1$, this proves Theorem 2.

We now prove the lower and upper bounds for the velocities of travelling waves given in Proposition 5. These follow from (5), (14) and the following proposition.

PROPOSITION 8. *For any* $(\lambda, u) \in \Sigma$ *with* $\lambda > 0$ *we have*

$$0 < \frac{\Gamma}{\lambda} < \Phi(\lambda, u) < \frac{\Gamma}{\lambda} + 1.$$

The upper bound follows immediately from the definition (20) of $\Phi(\lambda, u)$ since $\frac{1}{2\pi}\int_0^{2\pi} \sin(s + u(s))ds < 1$. The lower bound follows from the claim that

$$(\lambda, u) \in \Sigma \;\Rightarrow\; \int_0^{2\pi} \sin(s + u(s))ds > 0. \tag{25}$$

To prove this claim we multiply (15) by $1 + u'(z)$ and integrate over $[0, 2\pi]$, noting that

$$\begin{aligned}\int_0^{2\pi} u\Big(s + 2\pi\frac{m}{n}\Big)u'(s)ds &= \int_0^{2\pi} u(s)u'\Big(s - 2\pi\frac{m}{n}\Big)ds \\ &= -\int_0^{2\pi} u'(s)u\Big(s - 2\pi\frac{m}{n}\Big)ds,\end{aligned}$$

so that we obtain

$$(\lambda, u) \in \Sigma \Rightarrow \Gamma\frac{1}{2\pi}\int_0^{2\pi} (u'(s))^2 ds = \lambda\frac{1}{2\pi}\int_0^{2\pi} \sin(s + u(s))ds.$$

This proves (25) since the left-hand side is nonnegative and cannot vanish unless $u \equiv 0$, but $(\lambda, 0) \notin \Sigma$ for $\lambda > 0$.

We now turn to the proof of Theorem 3.

PROPOSITION 9. *Assume* n *does not divide* m *and* $\tilde{\Gamma} > 0$. *Given any* $\lambda_0 > 0$ *and* $\epsilon > 0$, *there exists* K_0 *such that for* $K \geq K_0$ *and* $\Gamma \geq \tilde{\Gamma}$ *we have that*

$$\Big|\frac{1}{2\pi}\int_0^{2\pi} \sin(s + u(s))ds\Big| < \epsilon \text{ if } (\lambda_0, u) \in \Sigma. \tag{26}$$

To see that Proposition 9 implies Theorem 3, we fix some $\tilde{F} > 0$, $\tilde{\Gamma} > 0$, and assume $\Gamma \geq \tilde{\Gamma}$. We choose $\lambda_0 > \frac{\Gamma}{\tilde{F}}$ and set $\epsilon = \tilde{F} - \frac{\Gamma}{\lambda_0}$. We then choose K_0 according to Proposition 9, so that (26) holds, which implies that when $K \geq K_0$ we have $\Phi(\lambda_0, u) < \tilde{F}$ for any u with $(\lambda_0, u) \in C$. Thus $\underline{F} < \tilde{F}$, where $\underline{F}$ is defined by (24), so Proposition 7 implies the existence of a travelling wave for any $F \geq \tilde{F}$.

We now prove Proposition 9. Let $\lambda_0 > 0$ and $\epsilon > 0$ be given. Assume $(\lambda_0, u) \in \Sigma$, so that (19) holds with $\lambda = \lambda_0$. Let (m, n) denote the greatest common divisor of m, n, and let

$$p = \frac{m}{(m, n)}, \qquad q = \frac{n}{(m, n)}.$$

Since we assume n does not divide m we have $q \geq 2$. Let Y_0 be the subspace of Y consisting of $\frac{2\pi}{q}$-periodic functions, and let Y_1 be its orthogonal complement in Y. We denote using P the orthogonal projection of Y to Y_0. Setting

$$u_0 = P(u), \qquad u_1 = (I - P)(u),$$

we have $u = u_0 + u_1$ with $u_0 \in Y_0$ and $u_1 \in Y_1$. Applying P and $I - P$ to (19), and noting that L_λ commutes with P, we have

$$u_0 = \lambda_0^2 L_{\lambda_0}^{-1} \circ P \circ N(u_0 + u_1), \tag{27}$$

$$u_1 = \lambda_0^2 L_{\lambda_0}^{-1} \circ (I-P) \circ N(u_0+u_1). \tag{28}$$

We will now use (16) to derive a bound for $\|L_{\lambda_0}^{-1}|_{Y_1}\|_{Y_1,Y_1}$ which goes to 0 as $K \to \infty$. We note that

$$\|L_{\lambda_0}^{-1}|_{Y_1}\|_{Y_1,Y_1} \leq \max_{l\geq 1,\ q\not| l} \frac{1}{|\mu_l|}, \tag{29}$$

so we need to find lower bounds for the $|\mu_l|$'s for which q does not divide l. We define

$$\rho = \max_{l\geq 1,\ q\not| l} \cos\Big(\frac{2\pi pl}{q}\Big)$$

and note that since p and q are coprime we have $\rho < 1$.

We define

$$\alpha = 2K\lambda_0^2(1-\rho) - \sqrt{K},$$

and we shall henceforth assume that K is sufficiently large so that $\alpha > 0$. For each $l \geq 1$ we have either $l^2 < \alpha$ or $l^2 \geq \alpha$, and we treat each of these cases separately.

(1) In case $l^2 < \alpha$, we have

$$l^2 + 2K\lambda_0^2(\rho - 1) < -\sqrt{K},$$

and by the definition of ρ,

$$\cos\Big(\frac{2\pi ml}{n}\Big) = \cos\Big(\frac{2\pi pl}{q}\Big) \leq \rho,$$

so that

$$l^2 + 2K\lambda_0^2\Big(\cos\Big(\frac{2\pi ml}{n}\Big) - 1\Big) < -\sqrt{K},$$

which by (16) implies

$$|\mu_l| > \sqrt{K}. \tag{30}$$

(2) In case $l^2 \geq \alpha$, we have, since (16) implies $|\mu_l| > \lambda_0 \Gamma l$,

$$|\mu_l| \geq \lambda_0\Gamma\sqrt{\alpha} \geq \lambda_0\tilde{\Gamma}\sqrt{\alpha} = \lambda_0\tilde{\Gamma}\Big[2K\lambda_0^2(1-\rho) - \sqrt{K}\Big]^{\frac{1}{2}}. \tag{31}$$

From (30), (31) we obtain that $\lim_{K\to\infty} |\mu_l| = +\infty$ uniformly with respect to $l \geq 1$ which are not multiples of q, hence by (29)

$$\lim_{K\to\infty} \|L_{\lambda_0}^{-1}|_{Y_1}\|_{Y_1,Y_1} = 0.$$

In particular, we may choose K_0 such that for $K \geq K_0$ we will have

$$\|L_{\lambda_0}^{-1}|_{Y_1}\|_{Y_1,Y_1} < \frac{\epsilon}{\lambda_0^2}.$$

By (28) and (18) this implies

$$\|u_1\|_Y \leq \epsilon. \tag{32}$$

Thus

$$\begin{aligned}\Big|\frac{1}{2\pi}\int_0^{2\pi}\sin(s+u(s))ds\Big| &\le \Big|\frac{1}{2\pi}\int_0^{2\pi}\sin(s+u_0(s))ds\Big| \\ &\quad + \Big|\frac{1}{2\pi}\int_0^{2\pi}[\sin(s+u(s))-\sin(s+u_0(s))]ds\Big| \\ &\le \Big|\frac{1}{2\pi}\int_0^{2\pi}\sin(s+u_0(s))ds\Big| + \frac{1}{2\pi}\int_0^{2\pi}|u_1(s)|ds. \qquad (33)\end{aligned}$$

From (32) and the Cauchy–Schwarz inequality we have

$$\frac{1}{2\pi}\int_0^{2\pi}|u_1(s)|ds \le \frac{1}{\sqrt{2\pi}}\Big(\int_0^{2\pi}(u_1(s))^2ds\Big)^{\frac{1}{2}} \le \epsilon. \qquad (34)$$

From trigonometry we have

$$\int_0^{2\pi}\sin(s+u_0(s))ds = \int_0^{2\pi}\sin(s)\cos(u_0(s))ds + \int_0^{2\pi}\cos(s)\sin(u_0(s))ds,$$

but the functions $\cos(u_0(s))$ and $\sin(u_0(s))$ are $\frac{2\pi}{q}$-periodic with $q \ge 2$, which implies that they are orthogonal to $\cos(s)$ and $\sin(s)$, so that we have

$$\int_0^{2\pi}\sin(s+u_0(s))ds = 0,$$

which together with (33) and (34) implies (26), concluding the proof of Proposition 9.

We now turn to the proof of Theorem 4. We first note that from (16) we have

$$|\mu_l| \ge \Big|l^2 + 2K\lambda^2\Big(\cos\Big(\frac{2\pi ml}{n}\Big) - 1\Big)\Big|,$$

so that if we assume

$$0 < \lambda < \lambda_0 = \frac{1}{\sqrt{8\tilde{K}}} \le \frac{1}{\sqrt{8K}},$$

then we have $|\mu_l| > \frac{1}{2}$ for all $l \ge 1$, hence $\|L_\lambda^{-1}\|_{Y,Y} < 2$, independently of Γ. From (19) we thus have

$$(\lambda, u) \in \Sigma,\ 0 < \lambda < \lambda_0 \ \Rightarrow\ \|u\|_Y < 2\lambda^2.$$

We now choose $\lambda_1 \le \lambda_0$ so that $2\lambda_1^2 \le \frac{1}{2}\tilde{F}$. Thus $(\lambda_1, u) \in \Sigma$ implies that

$$\begin{aligned}\Big|\frac{1}{2\pi}\int_0^{2\pi}\sin(s+u(s))ds\Big| &= \Big|\frac{1}{2\pi}\int_0^{2\pi}[\sin(s+u(s))-\sin(s)]ds\Big| \\ &\le \frac{1}{2\pi}\int_0^{2\pi}|u(s)|ds \le \|u\|_Y < 2\lambda_1^2 \le \frac{1}{2}\tilde{F}. \qquad (35)\end{aligned}$$

Finally, we choose Γ_0 so that

$$\frac{\Gamma_0}{\lambda_1} < \frac{1}{2}\tilde{F}. \qquad (36)$$

Relations (35) and (36) thus imply that when $0 < \Gamma < \Gamma_0$,

$$(\lambda_1, u) \in \Sigma \ \Rightarrow\ \Phi(\lambda_1, u) < \tilde{F},$$

so that we have $\underline{F} < \tilde{F}$, where $\underline{F}$ is defined by (24), hence Proposition 7 implies the existence of a travelling wave for any $F \ge \tilde{F}$.

3. Discussion and further questions. In the numerical and experimental explorations of the dynamics of sine-Gordon rings [8, 10, 11], a useful method of representation consists in displaying the velocity-force characteristic. In the case of the travelling waves studied here, since the velocity of the waves is given by $v = \frac{1}{\lambda}$, the velocity-force characteristic is the subset of the (F, v)-plane given by

$$\Big\{\Big(\Phi\Big(\frac{1}{v}, u\Big), v\Big) \;\Big|\; \Big(\frac{1}{v}, u\Big) \in \Sigma\Big\},$$

where the set Σ and the functional Φ are as defined in the previous section. Examining the velocity-force characteristic as numerically computed in [8, Figure 2], we see that F is a nonmonotone function of v. This means that for some values of F equation (21) has more than one solution, or in other words that there exist multiple travelling waves with different velocities for the same value of F. On the other hand, the fact that, according to the available numerical evidence F is a function of v, leads us to conjecture that, fixing $\Gamma > 0$ and $K > 0$ for each given velocity $v > 0$, there is a unique travelling wave with velocity v, for an appropriate F. Thus, our conjecture is that uniqueness holds in Theorem 1, or in other words that the fixed-point problem (19) always has a unique solution. Let us note that for $0 < \lambda < \Gamma$ it is easy to show, using (17), that the right-hand side of (19) is a contraction from Y to itself, hence we may replace the use of the Schauder fixed-point theorem by the Banach contraction-mapping principle, which implies uniqueness. Thus, at least for velocities $v > \frac{1}{\Gamma}$, we have uniqueness in Theorem 1. However, for lower velocities this argument does not work so a proof of the above conjecture will require some new idea.

The nonuniqueness of travelling waves for some values of the parameters Γ, K, F, mentioned above, implies important consequences for the dynamics of the discrete sine-Gordon ring, such as instability of some of the travelling waves, and bistability of travelling waves leading to hysteresis as the force F is varied. It would be interesting to determine whether these phenomena can occur in the large K and the small Γ regimes for which existence of a travelling wave has been proved in Theorems 3 and 4. Moreover, stationary solutions and travelling waves do not exhaust the dynamical repertoire of the discrete sine-Gordon equations in the underdamped case: quasi-periodic and chaotic behavior is reported in [8, 9, 11]. An interesting question is to determine conditions on the parameters which ensure that there exists at least one *locally asymptotically stable* travelling wave.

Returning to the issue of existence of travelling waves, which has been the focus of our investigation, we note an intriguing question which arises from our results, and remains unanswered.

Let us define, for fixed $\Gamma > 0$ and $K > 0$

$$F_0(\Gamma, K) = \inf\{F \geq 0 \mid \text{a travelling wave of (1), (2) exists}\}.$$

Theorem 2 implies that $F_0(\Gamma, K) \leq 1$ for all $\Gamma > 0$ and $K > 0$. Theorem 3 implies that, when n does not divide m, $\lim_{K\to\infty} F_0(\Gamma, K) = 0$. Theorem 4 implies that $\lim_{\Gamma\to 0} F_0(\Gamma, K) = 0$. Is it true, though, that for each $\Gamma > 0$ and $K > 0$ we have $F_0(\Gamma, K) > 0$? In other words, is it always true that (fixing $\Gamma > 0$ and $K > 0$) for sufficiently small $F > 0$ a travelling wave does not exist? We have not been able to prove or disprove this conjecture, and can only offer the following remarks.

(i) If Γ and K satisfy (6), then indeed $F_0(\Gamma, K) > 0$, since for sufficiently small F there exists a stationary solution of (1), (2), so the results of [2] imply that no travelling wave exists for small F. However, as we have remarked, Theorem 4 shows that in general travelling waves and stationary solutions may coexist.

(ii) If n divides m, then, as was noted in the introduction, the existence of travelling waves reduces to that of running solutions of the forced pendulum, hence it is well known that $F_0(\Gamma, K) > 0$ for all Γ and K. However, in the case that n divides m we also have noted that the result of Theorem 3 does not hold, so that the case that n divides m is rather special and may not be indicative of the general case.

(iii) The conjecture that $F_0(\Gamma, K) > 0$ is supported by the notion of "pinning"—the phenomenon whereby travelling waves are unable to propagate in discrete systems when the applied force is small. However, whether this effect indeed holds in general in underdamped systems (as opposed to the overdamped case—see (i) above) is unclear to the best of our knowledge. Moreover, it is conceivable that $F_0(\Gamma, K) = 0$ but pinning still occurs—if for small F a travelling wave exists but is unstable.

(iv) Since, by Proposition 8, we have $\Phi(\lambda, u) > 0$ for all $(\lambda, u) \in \Sigma$, $\lambda > 0$, we have that $F_0(\Gamma, K) = 0$ if and only if

$$\liminf_{\lambda \to +\infty,\ (\lambda,u)\in\Sigma} \Phi(\lambda, u) = 0.$$

Thus, determining whether the above equality can hold could be a route to resolving our question, but we have not been able to do so.

We conclude with one more question: clarify the connection, if any, between the travelling waves obtained in [4] for small values of $K > 0$ and those obtained by us for large values of K in Theorem 3.

REFERENCES

[1] C. Baesens and R. S. MacKay, *Gradient dynamics of tilted Frenkel-Kontorova models*, Nonlinearity, 11 (1998), pp. 949–964.

[2] C. Baesens and R. S. MacKay, *A novel preserved partial order for cooperative networks of units with overdamped second-order dynamics and application to tilted Frenkel Kontorova chains*, Nonlinearity, 17 (2004), pp. 567–580.

[3] L. M. Floría and J. J. Mazo, *Dissipative dynamics of the Frenkel-Kontorova model*, Adv. Phys., 45 (1996), pp. 505–598.

[4] M. Levi, *Dynamics of discrete Frenkel-Kontorova models*, in Analysis, et cetera, P. Rabinowitz and E. Zehnder, eds., Academic Press, Boston, 1990.

[5] M. Levi, F. C. Hoppensteadt, and W. L Miranker, *Dynamics of the Josephson junction*, Quart., Appl. Math., 36 (1978/79), pp. 167–198.

[6] R. Mirollo and N. Rosen, *Existence, uniqueness, and nonuniqueness of single-wave-form solutions to Josephson junction systems*, SIAM J. Appl. Math, 60 (2000), pp. 1471–1501.

[7] P. Rabinowitz, *Some global results for nonlinear eigenvalue problems*, J. Funct. Anal., 7 (1971), pp. 487–513.

[8] T. Strunz and F. J. Elmer, *Driven Frenkel-Kontorova model:* I. *Uniform sliding states and dynamical domains of different particle densities*, Phys. Rev. E, 58 (1998), pp. 1601–1611.

[9] T. Strunz and F. J. Elmer, *Driven Frenkel-Kontorova model:* II. *Chaotic sliding and nonequilibrium melting and freezing*, Phys. Rev. E, 58 (1998), pp. 1612–1620.

[10] S. Watanabe, H. S. J. van der Zant, S. Strogatz, and T. P. Orlando, *Dynamics of circular arrays of Josephson junctions and the discrete sine-Gordon equation*, Phys. D, 97 (1996), pp. 429–470.

[11] Z. Zheng, B. Hu, and G. Hu, *Resonant steps and spatiotemporal dynamics in the damped dc-driven Frenkel-Kontorova chain*, Phys. Rev. B, 58 (1998), pp. 5453–5461.

SIAM J. MATH. ANAL.
Vol. 36, No. 5, pp. 1444–1460

AN ELLIPTIC PROBLEM RELATED TO PLANAR VORTEX PAIRS*

GONGBAO LI†, SHUSEN YAN‡, AND JIANFU YANG§

Abstract. In this paper, we study the existence and limiting behavior of the mountain pass solutions of the elliptic problem $-\Delta u = \lambda f(u - q(x))$ in $\Omega \subset R^2; u = 0$ on $\partial\Omega$, where q is a positive harmonic function. We show that the "vortex core" $A_\lambda = \{x \in \Omega : u_\lambda(x) > q(x)\}$ of the solution u_λ shrinks to a global minimum point of q on the boundary $\partial\Omega$ as $\lambda \to +\infty$. Furthermore, we show that for each strict local minimum x_0 point of $q(x)$ on the boundary $\partial\Omega$, there exists a solution u_λ whose vortex core shrinks to this strict local minimum point x_0 as $\lambda \to +\infty$.

Key words. vortex ring, free boundary problem

AMS subject classifications. 35R35, 35J25

DOI. 10.1137/S003614100343055X

1. Introduction.

In cylindrical coordinates in R^3, a steady vortex ring corresponds mathematically to a Stokes stream function Ψ defined on a domain $\bar{\Omega} \subset R^2$, and an open set $A \subset \Omega$, called the cross-section of a steady vortex ring and unknown a priori, such that $\Psi \in C^1(\bar{\Omega}) \cap C^2(\Omega \backslash A)$ and satisfies the equations

$$-L\Psi = \begin{cases} \lambda r^2 f(\Psi) & \text{in } A, \\ 0 & \text{in } \Omega \backslash \bar{A}, \end{cases} \tag{1.1}$$

$$\Psi|_{\partial A} = 0, \quad \Psi|_{\partial\Omega} = -\frac{1}{2} W r^2 - k, \tag{1.2}$$

where $L = r(\partial/\partial r)(1/r\partial/\partial r) + \partial^2/\partial z^2$. The vorticity function f is supposed to be positive if $t > 0$ and equal to zero if $t \le 0$, while $W > 0$ and k are prescribed constants; λ is a positive parameter which is also regarded as prescribed.

When we study the steady plane flow of an inviscid fluid of uniform density, we are led to the following free-boundary problem:

$$-\Delta\Psi = \begin{cases} \lambda f(\Psi) & \text{in } A, \\ 0 & \text{in } \Omega \backslash \bar{A}, \end{cases} \tag{1.3}$$

$$\Psi|_{\partial A} = 0, \quad \Psi|_{\partial\Omega} = -W x_1 - k, \tag{1.4}$$

*Received by the editors June 23, 2003; accepted for publication (in revised form) March 6, 2004; published electronically March 25, 2005. The work of the first and third authors was supported by National Natural Sciences Foundation of China grant 10271118. The work of the first author was partially supported by the Academy of Finland.

http://www.siam.org/journals/sima/36-5/43055.html

†Department of Mathematics, Huazhong Normal University, Wuhan 430079, People's Republic of China, and Wuhan Institute of Physics and Mathematics, Chinese Academy of Sciences, P.O. Box 71010, Wuhan 430071, People's Republic of China (ligb@wipm.ac.cn).

‡School of Mathematics, Statistics and Computer Sciences, the University of New England, Armidale, NSW 2351, Australia.

§Wuhan Institute of Physics and Mathematics, Chinese Academy of Sciences, P.O. Box 71010, Wuhan 430071, People's Republic of China (jfyang@wipm.ac.cn).

where Ω is a bounded domain in $\{(x_1, x_2) : x_1 > 0\} \subset R^2$, both Ψ and $A \subset \Omega$ are unknown, and $\Psi \in C^1(\bar{\Omega}) \cap C^2(\Omega \setminus A)$. See, for example, [11]. Let $A = \{x \in \Omega : \Psi(x) > Wx_1 + k\}$. Since $f(t) = 0$ if $t \leq 0$, problem (1.3)–(1.4) can be rewritten as

$$\begin{cases} -\Delta\Psi = \lambda f(\Psi) & \text{in } \Omega, \\ \Psi = -Wx_1 - k < 0 & \text{on } \partial\Omega. \end{cases} \tag{1.5}$$

In general, we shall consider the free boundary problem

$$-\Delta\Psi = \begin{cases} \lambda f(\Psi) & \text{in } A, \\ 0 & \text{in } \Omega \backslash \bar{A}, \end{cases} \tag{1.6}$$

$$\Psi|_{\partial A} = 0, \quad \Psi|_{\partial\Omega} = -q_0(x) < 0, \tag{1.7}$$

where $q_0(x)$ is a C^1 function defined on $\partial\Omega$. Problem (1.6), (1.7) can be reduced to the following problem:

$$\begin{cases} -\Delta w = \lambda f(w) & \text{in } \Omega \subset R^2, \\ w = -q_0(x) < 0 & \text{on } \partial\Omega. \end{cases} \tag{1.8}$$

Let $q(x)$ be the solution of

$$\begin{cases} -\Delta v = 0 & \text{in } \Omega, \\ v = q_0(x) & \text{on } \partial\Omega. \end{cases} \tag{1.9}$$

Then $q(x) > 0$ and $q(x)$ achieves its maximum and minimum on $\partial\Omega$.

Let $u = w + q(x)$. Then (1.8) becomes

$$\begin{cases} -\Delta u = \lambda f(u - q(x)) & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases} \tag{1.10}$$

Our main objective is to obtain the existence result and investigate the asymptotic behavior of the solution pair (u_λ, A_λ) of problem (1.10) as $\lambda \to \infty$, where $A_\lambda = \{x \in \Omega : u_\lambda(x) > q(x)\}$.

The existence problems were considered in [1, 3, 5, 6, 8, 10, 11, 12]. Precisely, the work of [5, 6, 8, 11, 12] is related to constrained variation. Particularly, in Turkington's setting [12], the vorticity function f is unknown a priori, while in Badiani's work [5], the stream function Ψ for the flow satisfies the equation in (1.8) in a region bounded by the line of symmetry, where f is an increasing function that is unknown a priori either. Moreover, in [6, 8, 11], the parameter λ is a Lagrange multiplier and hence is not known a priori. From the existence point of view, the solutions obtained in [5, 6, 8, 11, 12] are related to the eigenvalue problem, so the corresponding eigenvalue λ is not arbitrary. In the works of [1, 3, 4, 10, 13, 14], the mountain pass lemma was used to get the existence results for a large class of nonlinearities $f(x, u)$ and any $\lambda > 0$. These results are related to "free variation" which provides more information for the existence.

The asymptotic behavior of the solution pair (u_λ, A_λ) of problem (1.10) as $\lambda \to \infty$ was initiated by the work of [6], where the desingularization problem was studied for a confined steady vortex ring. (The reader should refer to Appendix A in [6] for a

detailed explanation of the physical background of the problem.) The asymptotic behavior of solution pair (u_λ, A_λ) of problem (1.10) was studied by [4, 6, 13, 14] for vortex rings and vortex pairs. Particularly, it is proved that the cross-section A_λ of a steady vortex ring shrinks to a point, and a vortex ring degenerates into a singular vortex circle as $\lambda \to +\infty$. Moreover, the Stokes stream function Ψ_λ of a vortex ring converges to the Stokes stream function of the filament, which is the Green's function of the operator $-\Delta$ in Ω. In [6], the asymptotic behavior of the solutions obtained by the constrained variational method was discussed. There, the parameter λ is unknown a priori. As pointed out in [3], the existence result for given λ and $f(t)$, and the asymptotic behavior of the core as $\lambda \to +\infty$ are desirable in application. In [4], Ambrosetti and Yang freed the parameter λ. Using the mountain pass lemma, they discussed the existence of solutions and then studied the limiting behavior of the solution pair. The study of the limiting behavior of the solution pair relies on upper estimates of critical level and connectedness of the A_λ. These facts allow one to show that, under certain conditions as proposed in [4, 6], the solution u_λ satisfies

$$\frac{u_\lambda(\cdot)}{\lambda \int_\Omega f(u_\lambda - q(x))\,dx} - G(\cdot, x_\lambda) \to 0 \ \text{ in } \ W^{1,p}(\Omega) \tag{1.11}$$

as $\lambda \to \infty$, where $x_\lambda \in A_\lambda, 1 \le p < 2$, G is the Green's function of Ω.

It would be interesting to describe precisely the location of the core of A_λ as $\lambda \to \infty$. Our first result deals with the problem.

Suppose that $f(t) \in C^1(R^1)$ and satisfies the following conditions:

(f_1) If $t \le 0$, $f(t) = 0$.

(f_2) There is $\theta > 1$, such that $f'(t)t \ge \theta f(t) > 0 \ \forall\ t > 0$.

(f_3) There are $p_1 > 2$ and $C > 0$, such that $\forall t \ge 0$, $f(t) \le C(1 + t^{p_1 - 1})$.

(f_4) There are $p > 2$, $t_0 > 0$ and $a_0 > 0$, such that $f(t) \ge a_0 t^{p-1} \ \forall\ t \in [0, t_0]$.

We mention that (f_2) implies $tf(t) \ge (\theta + 1)F(t) > 0$ if $t > 0$, where $F(t) = \int_0^t f(s)\,ds$.

Define

$$I_\lambda(u) = \frac{1}{2}\int_\Omega |Du|^2\,dx - \lambda \int_\Omega F(u - q(x))\,dx, \quad u \in H_0^1(\Omega).$$

Using the mountain pass theorem, we know that I_λ has a critical point u_λ satisfying

$$I_\lambda(u_\lambda) = c_\lambda, \quad I'_\lambda(u_\lambda) = 0$$

with

$$c_\lambda = \inf_{w \in H_0^1(\Omega)} \max_{t \ge 0} I_\lambda(tw).$$

Our first result is as follows.

THEOREM 1.1. *Assume that $q_0(x)$ is not a constant. Then we have the following.*

(i) *A_λ is connected and $\mathrm{diam} A_\lambda \to 0$ as $\lambda \to +\infty$.*

(ii) *For any $x_\lambda \in A_\lambda$, suppose $x_\lambda \to x_0$. Then $q(x_0) = q_m := \min_{x \in \partial\Omega} q(x)$.*

(iii) *As $\lambda \to +\infty$,*

$$c_\lambda = \frac{2\pi q_m^2}{\ln \lambda}(1 + o(1)),$$

where $o(1) \to 0$ as $\lambda \to 0$.

(iv) *As* $\lambda \to +\infty$, *we have*

$$\frac{u_\lambda(\cdot)}{\lambda \int_\Omega f(u_\lambda - q(x))\,dx} - G(\cdot, x_\lambda) \to 0 \ \text{ in } \ W^{1,p}(\Omega), \tag{1.12}$$

where $x_\lambda \in A_\lambda, 1 \le p < 2$, G *is the Green's function of* Ω, *and*

$$\frac{u_\lambda}{\lambda \int_\Omega f(u_\lambda - q(x))\,dx} - G(\cdot, x_\lambda) \to 0 \ \text{ in } \ C^{1,\alpha}_{loc}(\Omega \setminus \{x_0\}), \tag{1.13}$$

where α *is any constant in* $(0,1)$.

From Theorem 1.1 we know that there is a solution of (1.10) whose cross-section A_λ shrinks to a global minimum point of q_0 on the boundary $\partial\Omega$. If q_0 has, for instance, k different local minimum points on $\partial\Omega$, does (1.10) possess at least k solution? It seems that there is not previous work in this direction. Our next result concerns the effect of the function $q_0(x)$ on the number of the solution pairs. It shows that each strictly local minimum point x_0 of q_0 on the boundary $\partial\Omega$ corresponds to a solution u_λ of (1.10) with A_λ shrinking to x_0 as $\lambda \to \infty$.

THEOREM 1.2. *Let* x_0 *be a strictly local minimum point of* $q(x)$ *on the boundary* $\partial\Omega$. *Then there exists a solution* u_λ *of* (1.10) *such that*

(i) A_λ *is connected and* $diam A_\lambda \to 0$ *as* $\lambda \to +\infty$;

(ii) *for any* $x_\lambda \in A_\lambda$, *we have* $x_\lambda \to x_0$ *as* $\lambda \to +\infty$.

For the case $q_0(x) = C$, let us point out that by using the estimates in sections 2 and 3, together with the technique in [6], it might prove that if $q_0(x) = C$, the core would shrink to a point x_0, which is a critical point of the Robin function $H(x,x)$, where $H(y,x)$ is the regular part of the Green's function of $-\Delta$ with Dirichlet boundary condition. In particular, if q_0 is a positive constant, the core lies in the interior of the domain. It is known that the Robin function $H(x,x)$ is related to the domain geometry and the domain topology. In the case $q_0 \neq C$, our results show that the vortex core shrinks to a point on the boundary no matter how close q_0 is to a constant. Then the effect from the boundary data is so strong that the effect from the domain is negligible. In other words, if $q_0(x) \neq C$, the location of the limit of the core as $\lambda \to \infty$ is not related to the Robin function $H(x,x)$.

This paper is arranged as follows. We first obtain an upper bound for c_λ in section 2. Then the connectedness of A_λ is shown in section 3. Theorems 1.1 and 1.2 are proven in sections 4 and 5, respectively.

2. Upper bound. We will analyze the limit behavior of u_λ as $\lambda \to +\infty$. First we will obtain an upper bound for c_λ. To this end, let $\bar{q} > 0$ be a constant. We consider a related problem in a ball B_δ:

$$\begin{cases} -\Delta u = \lambda f(u - \bar{q}) & \text{in } B_\delta, \\ u = 0 & \text{on } \partial B_\delta, \end{cases} \tag{2.1}$$

where $\delta > 0$ is a fixed constant.

Since $f(0) = 0$ and f is a C^1 function, it is superlinear at zero; i.e., $\lim_{t\to 0} \frac{f(t)}{t} = 0$. By (f_1)–(f_3), we may verify that for each $\lambda > 0$, (2.1) has a mountain pass solution with critical value defined as

$$c_{\lambda,\bar{q}} = \inf_{v \in H_0^1(B_\delta), v \not\equiv 0} \max_{t \ge 0} \Big(\frac{1}{2} t^2 \int_{B_\delta} |Dv|^2\,dx - \lambda \int_{B_\delta} F(tv - \bar{q})\,dx \Big)$$

by the mountain pass lemma [2].

To estimate $c_{\lambda,\bar{q}}$, we consider the following problem:

$$\begin{cases} -\Delta u = \lambda (u-\bar{q})_+^{p-1} & \text{in } B_\delta, \\ u = 0 & \text{on } \partial B_\delta, \end{cases} \tag{2.2}$$

where $p > 2$ is a fixed constant.

The existence of a solution for (2.2) can be proven in the following way.

First, we know from [2] that the problem

$$-\Delta \phi = \phi_+^{p-1} \quad \text{in } B_1,\ \phi \in H_0^1(B_1), \tag{2.3}$$

has a unique positive solution $\phi \in C^2(B_1(0))$, and by [9], ϕ is radially symmetric, i.e., $\phi(x) = \phi(r)$ with $r = |x|$.

Second, let ϕ be the solution of (2.3), and for any fixed $\delta > s > 0$, we know that

$$w = \lambda^{-\frac{1}{p-2}} s^{-\frac{2}{p-2}} \phi\left(\frac{r}{s}\right)$$

is the solution of

$$-\Delta w = \lambda w_+^{p-1} \quad \text{in } B_s,\ w \in H_0^1(B_s).$$

It is known that $\bar{q}\ln\frac{r}{\delta}/\ln\frac{s}{\delta}$ is the solution of

$$\begin{cases} -\Delta u = 0 & \text{for } s \le r \le \delta, \\ u = 0 & \text{on } r = \delta, \\ u = \bar{q} & \text{on } r = s. \end{cases} \tag{2.4}$$

Define

$$v_\lambda = \begin{cases} \bar{q} + \lambda^{-\frac{1}{p-2}} s^{-\frac{2}{p-2}} \phi(\frac{r}{s}) & \text{if } 0 \le r \le s, \\ \bar{q}\ln\frac{r}{\delta}/\ln\frac{s}{\delta} & \text{if } s \le r \le \delta. \end{cases}$$

Choose $s \in (0,\delta)$ such that $v_\lambda \in C^1(B_\delta(0))$; that is, s satisfies

$$\lambda^{-\frac{1}{p-2}} s^{-\frac{2}{p-2}} \frac{1}{s}\phi'(1) = \bar{q}\frac{1}{s}/\ln\frac{s}{\delta}.$$

Then we have

$$\frac{s/\delta}{(\ln\frac{\delta}{s})^{\frac{p-2}{2}}} = \lambda^{-\frac{1}{2}}\delta^{-1}\left(\frac{-\phi'(1)}{\bar{q}}\right)^{\frac{p-2}{2}}, \tag{2.5}$$

and v_λ is a solution of (2.2). We know from Lemma C.2 of [6] that for $\lambda > 0$ large, (2.5) has a unique solution s, which satisfies

$$s \sim \lambda^{-\frac{1}{2}} \ln^{\frac{p-2}{2}} \lambda. \tag{2.6}$$

We have the following estimate for the energy of v_λ.

Lemma 2.1.

$$\frac{1}{2}\int_{B_\delta(0)} |Dv_\lambda|^2\,dx - \frac{\lambda}{p}\int_{B_\delta(0)} (v_\lambda - \bar{q})_+^p\,dx = \frac{2\pi\bar{q}^2}{\ln\lambda}\left(1 + O\left(\frac{\ln(\ln\lambda)}{\ln\lambda}\right)\right) \tag{2.7}$$

as $\lambda \to \infty$.

Proof. A direct calculation by using (2.6) yields

$$\begin{aligned}\int_{B_\delta(0)} |Dv_\lambda|^2\,dx &= \int_{B_s(0)} |Dv_\lambda|^2\,dx + \int_{B_\delta(0)\setminus B_s(0)} |Dv_\lambda|^2\,dx\\ &= 2\pi\lambda^{-\frac{2}{p-2}} s^{-\frac{4}{p-2}} \frac{1}{s^2}\int_0^s r\left|\phi'\left(\frac{r}{s}\right)\right|^2 dr + 2\pi\bar q^2 \int_s^\delta \frac{1}{r}\Big/\left(\ln\frac{s}{\delta}\right)^2 dr\\ &= 2\pi\bar q^2\Big/\left(\ln\frac{\delta}{s}\right) + 2\pi\lambda^{-\frac{2}{p-2}} s^{-\frac{4}{p-2}} \int_0^1 r|\phi'(r)|^2\,dr \qquad (2.8)\\ &= 2\pi\bar q^2/(\ln(\lambda^{\frac{1}{2}}\ln^{-\frac{p-2}{2}}\lambda) + O(1)) + O\left(\frac{1}{\ln^2\lambda}\right)\\ &= \frac{4\pi\bar q^2}{\ln\lambda}\left(1 + O\left(\frac{\ln\ln\lambda}{\ln\lambda}\right)\right)\end{aligned}$$

and

$$\begin{aligned}\int_{B_\delta(0)} (v_\lambda - \bar q)_+^p\,dx &= \int_{B_s(0)} (v_\lambda - \bar q)^p\,dx = 2\pi\lambda^{-\frac{p}{p-2}} s^{-\frac{2p}{p-2}} \int_0^s r\phi^p\left(\frac{r}{s}\right) dr \qquad (2.9)\\ &= O(\lambda^{-\frac{p}{p-2}} s^{-\frac{2p}{p-2}} s^2) = O\left(\frac{1}{\lambda\ln^2\lambda}\right)\end{aligned}$$

as $\lambda \to \infty$. ∎

We are now ready to estimate $c_{\lambda,\bar q}$.

PROPOSITION 2.2.

$$\frac{2\pi\bar q^2}{\ln\lambda}\left(1 - O\left(\frac{1}{\ln\lambda}\right)\right) \le c_{\lambda,\bar q} \le \frac{2\pi\bar q^2}{\ln\lambda}\left(1 + O\left(\frac{\ln\ln\lambda}{\ln\lambda}\right)\right) \qquad (2.10)$$

as $\lambda \to +\infty$.

Proof. By (f_4), we know that there is a constant $b > 0$, such that

$$f(t) \ge bt^{p-1} \quad \forall\, t \in [0,1).$$

Let $v_{b\lambda}$ be the solution of (2.2), with λ replaced by $b\lambda$. Then we know that $|v_{b\lambda}| \le \bar q + 1$ if $\lambda > 0$ is large enough. See the formula for v_λ.

Consider

$$g(t) =: t^2 \int_{B_\delta(0)} |Dv_{b\lambda}|^2\,dx - \lambda \int_{B_\delta(0)} f(tv_{b\lambda} - \bar q)v_{b\lambda}\,dx.$$

Since f is superlinear at zero, we see that for $t > 0$ small, $g(t) > 0$. On the other hand, since $v_{b\lambda}$ is a solution of (2.2) and $|v_{b\lambda}| \le \bar q + 1$, we obtain

$$\begin{aligned}g(1) &= \int_{B_\delta(0)} |Dv_{b\lambda}|^2 - \lambda \int_{B_\delta(0)} f(v_{b\lambda} - \bar q)v_{b\lambda}\\ &\le \int_{B_\delta(0)} |Dv_{b\lambda}|^2 - \lambda b \int_{B_\delta(0)} (v_{b\lambda} - \bar q)_+^{p-1} v_{b\lambda} = 0.\end{aligned}$$

As a result, we know that there is a $\bar t \in (0,1]$, such that $g(\bar t) = 0$. Thus, by Lemma 2.1 and the fact that $\bar q > 0, v_{b\lambda} > 0$ and

$$\begin{aligned}&\max_{t\ge 0}\Big(\frac{1}{2}t^2 \int_{B_\delta(0)} |Dv_{b\lambda}|^2\,dx - \frac{1}{p}\int_{B_\delta(0)} \lambda b(tv_{b\lambda} - \bar q)_+^p\,dx\Big)\\ &= \frac{1}{2}\int_{B_\delta(0)} |Dv_{b\lambda}|^2\,dx - \frac{1}{p}\int_{B_\delta(0)} \lambda b(v_{b\lambda} - \bar q)_+^p\,dx,\end{aligned}$$

we obtain

$$\begin{aligned}
c_{\lambda,\bar q} \le & \max_{t\ge 0} I(tv_{b\lambda}) = I(\bar t v_{b\lambda}) = \frac{1}{2}\bar t^2 \int_{B_\delta(0)} |Dv_{b\lambda}|^2\,dx - \lambda \int_{B_\delta(0)} F(\bar t v_{b\lambda} - \bar q)\,dx \\
\le & \frac{1}{2}\bar t^2 \int_{B_\delta(0)} |Dv_{b\lambda}|^2\,dx - \frac{1}{p}\lambda \int_{B_\delta(0)} b(\bar t v_{b\lambda} - \bar q)_+^p\,dx \\
\le & \max_{t\ge 0}\Big(\frac{1}{2}t^2 \int_{B_\delta(0)} |Dv_{b\lambda}|^2\,dx - \frac{1}{p}\int_{B_\delta(0)} \lambda b(tv_{b\lambda} - \bar q)_+^p\,dx\Big) \\
= & \frac{2\pi\bar q^2}{\ln(b\lambda)}\left(1 + O\left(\frac{\ln\ln(b\lambda)}{\ln(b\lambda)}\right)\right) = \frac{2\pi\bar q^2}{\ln\lambda}\left(1 + O\left(\frac{\ln\ln\lambda}{\ln\lambda}\right)\right).
\end{aligned}$$

The upper bound thus follows.

It remains to prove the lower bound.

Let w_λ be the mountain pass solution of (2.1) with critical value $c_{\lambda,\bar q}$. By [9], we know that w_λ is radially symmetric and $w_\lambda'(r) < 0$. Since $c_{\lambda,\bar q} > 0$, we see that the set $\{x \in B_\delta : w_\lambda(x) > \bar q\}$ is not empty. In fact, otherwise, we would have $w_\lambda \equiv 0$ by (2.1) and the maximum principle. So there is unique $\bar s_\lambda \in (0,\delta)$, such that $w_\lambda(\bar s_\lambda) = \bar q$. Thus $-\Delta w_\lambda = 0$ if $x \in B_\delta(0) \setminus B_{\bar s_\lambda}(0)$, $-\Delta w_\lambda = \lambda f(w_\lambda - \bar q)$ if $x \in B_{\bar s_\lambda}(0)$. We may solve the boundary value problem $-\Delta w_\lambda = 0,\ x \in B_\delta(0) \setminus B_{\bar s_\lambda}(0)$ with $w_\lambda = 0$ if $|x| = \delta$ and $w_\lambda = \bar q$ if $|x| = \bar s_\lambda$ as (2.4). Therefore, letting $\tilde w_\lambda(r) = w_\lambda(r/\bar s_\lambda)$, we obtain

(2.11)

$$\begin{aligned}
c_{\lambda,\bar q} = & I(w_\lambda) = \frac{1}{2}\int_{B_\delta(0)\setminus B_{\bar s_\lambda}(0)} |Dw_\lambda|^2\,dx + \frac{1}{2}\int_{B_{\bar s_\lambda}(0)} |Dw_\lambda|^2\,dx - \lambda\int_{B_{\bar s_\lambda}(0)} F(w_\lambda - \bar q)\,dx \\
= & \frac{1}{2}\int_{B_\delta(0)\setminus B_{\bar s_\lambda}(0)} |Dw_\lambda|^2\,dx \\
& + \max_{t\ge 0}\Big(\frac{1}{2}t^2\int_{B_{\bar s_\lambda}(0)} |D(w_\lambda - \bar q)|^2\,dx - \lambda\int_{B_{\bar s_\lambda}(0)} F(t(w_\lambda - \bar q))\,dx\Big) \\
= & \frac{1}{2}\int_{B_\delta(0)\setminus B_{\bar s_\lambda}(0)} |Dw_\lambda|^2\,dx \\
& + \max_{t\ge 0}\Big(\frac{1}{2}t^2\int_{B_1(0)} |D(\tilde w_\lambda - \bar q)|^2\,dx - \lambda\bar s_\lambda^2\int_{B_1(0)} F(t(\tilde w_\lambda - \bar q))\,dx\Big) \\
= & \pi\bar q^2 \Big/ \left(\ln\frac{\delta}{\bar s_\lambda}\right) + \max_{t\ge 0}\Big(\frac{1}{2}t^2\int_{B_1(0)} |D(\tilde w_\lambda - \bar q)|^2\,dx - \lambda\bar s_\lambda^2\int_{B_1(0)} F(t(\tilde w_\lambda - \bar q))\,dx\Big) \\
\ge & \pi\bar q^2 \Big/ \left(\ln\frac{\delta}{\bar s_\lambda}\right) + \inf_{v\in H_0^1(B_1(0)),v\not\equiv 0}\max_{t\ge 0}\Big(\frac{1}{2}t^2\int_{B_1(0)} |Dv|^2\,dx - \lambda\bar s_\lambda^2\int_{B_1(0)} F(tv)\,dx\Big).
\end{aligned}$$

We claim that

$$\lambda\bar s_\lambda^2 > 1.$$

In fact, suppose that there are $\lambda_j \to +\infty$, such that

$$\lambda_j \bar s_{\lambda_j}^2 \le 1.$$

Then

$$\inf_{v\in H_0^1(B_1(0)),v\not\equiv 0}\max_{t\geq 0}\left(\frac{1}{2}t^2\int_{B_1(0)}|Dv|^2\,dx-\lambda_j\bar{s}_{\lambda_j}^2\int_{B_1(0)}F(tv)\,dx\right)$$
$$\geq \inf_{v\in H_0^1(B_1(0)),v\not\equiv 0}\max_{t\geq 0}\left(\frac{1}{2}t^2\int_{B_1(0)}|Dv|^2-\int_{B_1(0)}F(tv)\right)=:\bar{c}>0.$$

Thus, (2.11) yields $c_{\lambda_j,\bar{q}}\geq\bar{c}$. This is a contradiction because of $c_{\lambda_j,\bar{q}}\leq C/\ln\lambda_j$.

From $\lambda\bar{s}_\lambda^2>1$, we see $\frac{1}{\bar{s}_\lambda}<\lambda^{\frac{1}{2}}$. Using (2.11), we obtain

$$c_{\lambda,\bar{q}}\geq\pi\bar{q}^2/\left(\ln\frac{\delta}{\bar{s}_\lambda}\right)\geq\pi\bar{q}^2/(\ln(\delta\lambda^{\frac{1}{2}}))=\frac{2\pi\bar{q}^2}{\ln\lambda}\left(1-O\left(\frac{1}{\ln\lambda}\right)\right).\qquad\square$$

Let $x_0\in\partial\Omega$ be a point such that $q(x_0)=q_m=\min_{x\in\bar{\Omega}}q(x)$. We turn to the estimate of c_λ.

PROPOSITION 2.3. *For any small $\tau>0$, we have*

$$c_\lambda\leq\frac{2\pi(q_m+\tau)^2}{\ln\lambda}\left(1+O\left(\frac{\ln\ln\lambda}{\ln\lambda}\right)\right)\tag{2.12}$$

as $\lambda\to+\infty$.

Proof. For any $\tau>0$, take $\bar{x}_0\in\Omega$ close to x_0 and $\delta>0$ small so that

$$q(x)\leq q_m+\tau\quad\forall x\in B_\delta(\bar{x}_0).$$

Let $\bar{u}_\lambda\in H_0^1(B_\delta(\bar{x}_0))$ be the mountain pass solution of

$$-\Delta\bar{u}=\lambda(\bar{u}-(q_m+\tau))_+^{p-1}\quad\text{in } B_\delta(\bar{x}_0).$$

Then we deduce by Proposition 2.2 that

$$\begin{aligned}c_\lambda&\leq\max_{t\geq 0}I_\lambda(t\bar{u}_\lambda)\\&=\max_{t\geq 0}\left(\frac{1}{2}\int_{B_\delta(\bar{x}_0)}|Dt\bar{u}_\lambda|^2\,dx-\lambda\int_{B_\delta(\bar{x}_0)}F(t\bar{u}_\lambda-q(x))\,dx\right)\\&\leq\max_{t\geq 0}\left(\frac{1}{2}\int_{B_\delta(\bar{x}_0)}|Dt\bar{u}_\lambda|^2\,dx-\lambda\int_{B_\delta(\bar{x}_0)}F(t\bar{u}_\lambda-(q_m+\tau))\,dx\right)\\&=\frac{2\pi(q_m+\tau)^2}{\ln\lambda}\left(1+O\left(\frac{\ln\ln\lambda}{\ln\lambda}\right)\right)\end{aligned}\tag{2.13}$$

$\forall\tau>0.\qquad\square$

3. Connectedness of A_λ. Let u_λ be the mountain pass solution of (1.10) obtained in section 1 and $A_\lambda=\{x:u_\lambda(x)>q(x)\}$.

PROPOSITION 3.1. *A_λ is connected.*

Proof. We argue by contradiction. Suppose that A_λ has two components, A_1 and A_2. Let $\psi_i=u_\lambda-q(x)$ in A_i, $\psi_i=0$ for $x\in\Omega\setminus A_i, i=1,2$.

Let η_0 be a constant, which will be chosen later. Let $\widetilde{w}_\lambda=u_\lambda+s\psi_1-s\eta_0\psi_2$. It is easy to see that $\widetilde{w}_\lambda\in H_0^1(\Omega)$. Now we calculate $\max_{t\geq 0}I_\lambda(t\widetilde{w}_\lambda)$. We know that t_0 is a maximum point of $I_\lambda(t\widetilde{w}_\lambda)$ if and only if

$$t_0\int_\Omega|D\widetilde{w}_\lambda|^2\,dx=\lambda\int_\Omega f(t_0\widetilde{w}_\lambda-q(x))\widetilde{w}_\lambda\,dx.\tag{3.1}$$

In fact, if t_0 is a maximum point of $I_\lambda(t\tilde{w}_\lambda)$, it is clear that t_0 satisfies (3.1). On the other hand, from (f_2) we know that $\frac{d}{dt}\frac{f(t-q(x))}{t} > 0$ if $t > 0$; therefore

$$\begin{aligned}\frac{d}{dt}I(t\tilde{w}_\lambda) &= \frac{t\lambda}{t_0}\int_\Omega \tilde{w}_\lambda f(t_0\tilde{w}_\lambda - q(x))\,dx - \lambda\int_\Omega \tilde{w}_\lambda f(t\tilde{w}_\lambda - q(x))\,dx\\ &= t\lambda\int_\Omega\Big[\frac{f(t_0\tilde{w}_\lambda - q(x))}{t_0\tilde{w}_\lambda} - \frac{f(t\tilde{w}_\lambda - q(x))}{t\tilde{w}_\lambda}\Big]\tilde{w}_\lambda^2\,dx,\end{aligned}$$

which is positive if $t < t_0$ and negative if $t > t_0$. Thus, we have (3.1). Let

$$K(s,t) = t\int_\Omega |D\widetilde{w}_\lambda|^2\,dx - \lambda\int_\Omega f(t\widetilde{w}_\lambda - q(x))\widetilde{w}_\lambda\,dx.$$

Then, because u_λ is a solution, we obtain $K(0,1) = 0$ and

$$\begin{aligned}\frac{\partial K(0,t)}{\partial t}|_{t=1} = \frac{\partial K(0,1)}{\partial t} &= \int_\Omega |Du_\lambda|^2\,dx - \lambda\int_\Omega f'(u_\lambda - q(x))u_\lambda^2\,dx\\ &= \lambda\int_\Omega f(u_\lambda - q(x))u_\lambda\,dx - \lambda\int_\Omega f'(u_\lambda - q(x))u_\lambda^2\,dx < 0.\end{aligned} \tag{3.2}$$

Thus, by the implicit function theorem we know that for s small, there exists $t(s)$ which is differentiable such that $t(s) \to 1$ as $s \to 0$ and $K(s,t(s)) = 0$. On the other hand,

$$K_s(0,1) + K_t(0,1)t'(0) = 0.$$

Thus,

$$t'(0) = -\frac{K_s(0,1)}{K_t(0,1)}.$$

But by (f_2) we see that

(3.3)

$$\begin{aligned}K_s(0,1) &= 2\int_\Omega Du_\lambda D(\psi_1 - \eta_0\psi_2)\,dx - \lambda\int_\Omega f(u_\lambda - q(x))(\psi_1 - \eta_0\psi_2)\,dx\\ &\quad - \lambda\int_\Omega f'(u_\lambda - q(x))u_\lambda(\psi_1 - \eta_0\psi_2)\,dx\\ &= \lambda\int_\Omega f(u_\lambda - q(x))(\psi_1 - \eta_0\psi_2)\,dx - \lambda\int_\Omega f'(u_\lambda - q(x))u_\lambda(\psi_1 - \eta_0\psi_2)\,dx\\ &= \lambda\Big[\Big(\int_\Omega f(u_\lambda - q(x))\psi_1\,dx - \int_\Omega f'(u_\lambda - q(x))u_\lambda\psi_1\,dx\Big)\\ &\qquad - \eta_0\Big(\int_\Omega f(u_\lambda - q(x))\psi_2\,dx - \int_\Omega f'(u_\lambda - q(x))u_\lambda\psi_2\,dx\Big)\Big] = 0\end{aligned}$$

if

$$\eta_0 = \frac{\int_\Omega f(u_\lambda - q(x))\psi_1\,dx - \int_\Omega f'(u_\lambda - q(x))u_\lambda\psi_1\,dx}{\int_\Omega f(u_\lambda - q(x))\psi_2\,dx - \int_\Omega f'(u_\lambda - q(x))u_\lambda\psi_2\,dx} > 0.$$

With such $\eta_0 > 0$, we have $t'(0) = 0$. As a result,

$$t - 1 = t'(0)s + O(s^2) = O(s^2). \tag{3.4}$$

Now fix $s > 0$ small. Let $t(s)$ be the maximum point of $I_\lambda(t\widetilde{w}_\lambda)$. Then $K(s, t(s)) = 0$. Since $\mathrm{spt}\psi_1 \cap \mathrm{spt}\psi_2 = \emptyset$,

$$\begin{aligned}\int_\Omega |D\widetilde{w}_\lambda|^2\,dx &= \int_\Omega |Du_\lambda|^2\,dx + s^2\int_\Omega |D\psi_1|^2\,dx + \eta_0^2 s^2\int_\Omega |D\psi_2|^2\,dx\\ &\quad + 2s\int_\Omega Du_\lambda D\psi_1\,dx - 2\eta_0 s\int_\Omega Du_\lambda D\psi_2\,dx\\ &= \int_\Omega |Du_\lambda|^2\,dx + s^2\int_\Omega |D\psi_1|^2\,dx + \eta_0^2 s^2\int_\Omega |D\psi_2|^2\,dx\\ &\quad + 2s\lambda\int_\Omega f(u_\lambda - q(x))\psi_1\,dx - 2\eta_0 s\lambda\int_\Omega f(u_\lambda - q(x))\psi_2\,dx.\end{aligned} \tag{3.5}$$

Noting that $t(s) = 1 + O(s^2)$, we have

$$\begin{aligned}&\int_\Omega F(t(s)\widetilde{w}_\lambda - q(x))\,dx\\ &= \int_\Omega F(t(s)u_\lambda - q(x))\,dx + \int_\Omega f(t(s)u_\lambda - q(x))(s\psi_1 - \eta_0 s\psi_2)\,dx\\ &\quad + \frac{1}{2}\int_\Omega f'(t(s)u_\lambda - q(x))(s\psi_1 - \eta_0 s\psi_2)^2\,dx + O(s^{2+\sigma})\\ &= \int_\Omega F(t(s)u_\lambda - q(x))\,dx + \int_\Omega f(u_\lambda - q(x))(s\psi_1 - \eta_0 s\psi_2)\,dx\\ &\quad + \frac{1}{2}\int_\Omega f'(u_\lambda - q(x))(s^2\psi_1^2 + \eta_0^2 s^2\psi_2^2)\,dx + O(s^{2+\sigma}),\end{aligned} \tag{3.6}$$

where $\sigma > 0$. As a result,

$$\begin{aligned}&\max_{t\ge 0} I_\lambda(t\widetilde{w}_\lambda) = I_\lambda(t(s)\widetilde{w}_\lambda)\\ &\quad = \frac{1}{2}t^2(s)\int_\Omega |Du_\lambda|^2\,dx - \lambda\int_\Omega F(t(s)u_\lambda - q(x))\,dx\\ &\quad + \frac{1}{2}s^2\Big[\int_\Omega |D\psi_1|^2\,dx - \lambda\int_\Omega f'(u_\lambda - q(x))\psi_1^2\,dx\\ &\qquad + \eta_0^2\Big(\int_\Omega |D\psi_2|^2\,dx - \lambda\int_\Omega f'(u_\lambda - q(x))\psi_2^2\,dx\Big)\Big] + O(s^{2+\sigma}).\end{aligned} \tag{3.7}$$

Since

$$\begin{aligned}&\frac{1}{2}t^2(s)\int_\Omega |Du_\lambda|^2\,dx - \lambda\int_\Omega F(t(s)u_\lambda - q(x))\,dx\\ &\le \max_{t\ge 0} I_\lambda(tu_\lambda) = I_\lambda(u_\lambda) = c_\lambda\end{aligned}$$

and

$$\int_\Omega |D\psi_i|^2\,dx - \lambda\int_\Omega f'(u_\lambda - q(x))\psi_i^2\,dx = \lambda\int_\Omega (f(\psi_i) - f'(\psi_i)\psi_i)\psi_i\,dx < 0,$$

we obtain for s small enough that

$$\max_{t\ge 0} I_\lambda(t\widetilde{w}_\lambda) < c_\lambda = \inf_{u\in H_0^1(\Omega)}\max_{t\ge 0} I_\lambda(tu).$$

This is a contradiction. □

PROPOSITION 3.2. *diam$A_\lambda \to 0$ as $\lambda \to +\infty$.*

Proof. It is known from (2.12) that

$$I_\lambda(u_\lambda) \le \frac{C}{\ln\lambda}.$$

Since u_λ is a solution of (1.10), we deduce by (f_2) that

$$\begin{aligned} I_\lambda(u_\lambda) &= \frac{1}{2}\lambda \int_\Omega f(u_\lambda - q(x))u_\lambda \, dx - \lambda \int_\Omega F(u_\lambda - q(x)) \, dx \\ &= \lambda \int_\Omega \left(\frac{1}{2} f(u_\lambda - q(x))(u_\lambda - q(x)) - F(u_\lambda - q(x)) \right) dx \\ &+ \frac{\lambda}{2} \int_\Omega q(x) f(u_\lambda - q(x)) \, dx \\ &\ge \lambda \int_\Omega \left(\frac{1}{2} - \frac{1}{\theta + 1} \right) f(u_\lambda - q(x))(u_\lambda - q(x)) \, dx \\ &+ \frac{\lambda}{2} q_m \int_\Omega f(u_\lambda - q(x)) \, dx. \end{aligned}$$

As a result,

$$\lambda \int_\Omega f(u_\lambda - q(x)) \, dx \le \frac{C}{\ln\lambda},$$

$$\lambda \int_\Omega f(u_\lambda - q(x))(u_\lambda - q(x)) \, dx \le \frac{C}{\ln\lambda},$$

$$\|u_\lambda\|^2 = \int_\Omega |Du_\lambda|^2 \, dx = \lambda \int_\Omega f(u_\lambda - q(x))u_\lambda \, dx \le \frac{C}{\ln\lambda}.$$

Since A_λ is connected, thus we may prove $\big($as Theorem 5 of [4]; see also [7]$\big)$ that

$$\text{diam}A_\lambda \to 0 \ \text{ as } \ \lambda \to 0. \tag{3.8}$$

For the reader's convenience, we sketch the proof as follows.

Let $P, Q \in \bar{A}_\lambda$ be such that $|P - Q| = \text{diam}A_\lambda$, and consider a family of straight lines l_X passing through the point $X \in [P, Q]$. Denote by $L_X = [Y_X, Z_X]$ a segment in l_X such that $Y_X \in \partial\Omega, Z_X \in \partial A_\lambda$ and $\text{int}L_X \subset \Omega \backslash \bar{A}_\lambda$. Then one has

$$u_\lambda(Y_X) - u_\lambda(Z_X) = \int_{L_X} \frac{\partial u_\lambda}{\partial L_X} \, dL_X.$$

Note that $u_\lambda(Y_X) = 0$ while $u_\lambda(Z_X) = q(Z_X) \ge q_m > 0$; hence we obtain

$$q_m \le \left| \int_{L_X} \frac{\partial u_\lambda}{\partial L_X} \, dL_X \right| \le C \int_{L_X} |\nabla u_\lambda| \, dL_X.$$

Integrating with respect to X in $[P, Q]$ and using the Hölder inequality, we get

$$q_m |P - Q| \le c \int_{\overline{PQ}} dX \int_{L_X} |\nabla u_\lambda| \, dL_X \le C|P - Q|^{\frac{1}{2}} \|u_\lambda\|$$

which gives (3.8). □

Let $r_\lambda = \mathrm{diam} A_\lambda$. Then by Proposition 3.2, $r_\lambda \to 0$ as $\lambda \to +\infty$. Let $x_\lambda \in A_\lambda$ be such that $A_\lambda \subset B_{r_\lambda}(x_\lambda)$.

PROPOSITION 3.3. *There holds*

$$\frac{u_\lambda}{\lambda \int_\Omega f(u_\lambda - q(x))\,dx} - G(\cdot, x_\lambda) \to 0 \ \ in \ \ W^{1,p}(\Omega) \tag{3.9}$$

as $\lambda \to +\infty$, *where* $1 \le p < 2$.

Proof. We follow the arguments of Theorem 5.2 of [6]. Denote $h(\lambda) = \lambda \int_\Omega f(u_\lambda - q(x))\,dx$ and $G(x,y)$ the Green's function of $-\Delta$ in Ω, subject to the Dirichlet condition. Since

$$u_\lambda(y) = \lambda \int_{A_\lambda} G(x,y) f(u_\lambda(x) - q(x))\,dx$$

and

$$\frac{\lambda}{h(\lambda)} \int_{A_\lambda} f(u_\lambda(x) - q(x))\,dx = 1,$$

for $z(\lambda) \in A_\lambda$ we have

$$\frac{u_\lambda(y)}{h(\lambda)} - G(y, z(\lambda)) = \frac{\lambda}{h(\lambda)} \int_{A_\lambda} \{G(x,y) - G(y, z(\lambda))\} f(u_\lambda(x) - q(x))\,dx.$$

By the Minkowski inequality, there holds

$$\begin{aligned} &\left\| \frac{u_\lambda(\cdot)}{h(\lambda)} - G(\cdot, z(\lambda)) \right\|_{W^{1,p}(\Omega)} \\ \le &\frac{\lambda}{h(\lambda)} \int_{A_\lambda} \Big[\int_\Omega |\nabla_y (G(y,x) - G(y, z(\lambda))|^p\,dy \Big]^{\frac{1}{p}} f(u_\lambda(x) - q(x))\,dx. \end{aligned}$$

Lemma 5.1 of [6] yields

$$\int_\Omega |\nabla_y \{G(y,x) - G(y, z(\lambda))|^p\,dy \le C |x - z(\lambda)|^{2-p} \left(1 + \ln \left(\frac{\mathrm{diam}\Omega}{|x - z(\lambda)|} \right) \right)^2.$$

Since x and $z(\lambda)$ are both in A_λ, then $|x - z(\lambda)| \le \mathrm{diam} A_\lambda$. The conclusion follows from Proposition 3.2. □

PROPOSITION 3.4. *Let* $G(x,y)$ *be the Green's function of* $-\Delta$ *in* Ω, *subject to the Dirichlet condition. Then, for any* $\alpha \in (0,1)$,

$$\frac{u_\lambda}{\lambda \int_\Omega f(u_\lambda - q(x))\,dx} - G(\cdot, x_\lambda) \to 0 \ \ in \ \ C^{1,\alpha}(\Omega \setminus B_\delta(x_0))$$

as $\lambda \to +\infty$.

Proof. By Proposition 3.3, we know that

$$\frac{u_\lambda}{\lambda \int_\Omega f(u_\lambda - q(x))\,dx} - G(\cdot, x_\lambda) \to 0 \ \text{ in } \ W^{1,p}(\Omega) \tag{3.10}$$

for $x_\lambda \in A_\lambda$, where $1 \le p < 2$. On the other hand, for any $\delta > 0$, we have

$$A_\lambda \subset B_\delta(x_0)$$

for $\lambda > 0$ large enough, where $x_0 = \lim_{\lambda\to+\infty} x_\lambda$. Thus,

$$-\Delta u_\lambda = 0 \ \text{ in } \ \Omega \setminus B_\delta(x_0),$$

that is,

$$-\Delta v_\lambda := -\Delta\left(\frac{u_\lambda}{\lambda\int_\Omega f(u_\lambda - q(x))\,dx}\right) = 0 \ \text{ in } \ \Omega \setminus B_\delta(x_0).$$

By the L^p-estimate of the elliptic equation, we have for any $r > 0$

$$\|v_\lambda\|_{w^{2,r}(B_\tau(y_0))} \le C\|v_\lambda\|_{L^r(B_{2\tau}(y_0))}$$

for any $B_{2\tau}(y_0) \subset \Omega \setminus B_\delta(x_0)$. As

$$G(x, x_\lambda) = -\frac{1}{2\pi}\ln|x - x_\lambda| - R(x, x_\lambda) \quad \forall x \in \Omega,$$

where

$$\begin{cases} -\Delta_x R(x, x_\lambda) = 0, & x \in \Omega, \\ R(x, x_\lambda) = -\frac{1}{2\pi}\ln|x - x_\lambda|, & x \in \partial\Omega. \end{cases}$$

Clearly, $\left|-\frac{1}{2\pi}\ln|x - x_\lambda|\right| \le C$ on $\Omega \setminus B_\delta(x_0)$ and

$$\sup_{\partial B_\delta(x_0)\cap\bar\Omega} |R(x, x_\lambda)| = |R(y_\lambda, x_\lambda)|$$

for some $y_\lambda \in \partial B_\delta(x_0) \cap \bar\Omega$. If $|R(y_\lambda, x_\lambda)| \to +\infty$ as $\lambda \to +\infty$, then we may assume there exist subsequences $x_n := x_{\lambda_n}, y_n := y_{\lambda_n} \in \partial B_\delta(x_0) \cap \bar\Omega$ such that $x_n \to x_0$ and $y_n \to y_0 \in \partial B_\delta(x_0) \cap \bar\Omega$. Therefore, $R(x_0, y_0) = \infty$, which is a contradiction since $x_0 \ne y_0$. Hence

$$|G(x, x_\lambda)| \le C \quad \forall x \in \Omega \setminus B_\delta(x_0). \tag{3.11}$$

By (3.10) and (3.11), we know that

$$\|v_\lambda\|_{L^r(B_{2\tau}(y_0))} \le 1 + C\|G(\cdot, x_\lambda)\|_{L^r(B_{2\tau}(y_0))} \le C'.$$

Therefore, the conclusion follows. □

4. Proof of Theorem 1.1. We first get a lower bound for c_λ. Suppose that $x_\lambda \to x_0 \in \bar\Omega$. For any $\tau > 0$, let $\delta > 0$ be small enough such that

$$|q(x) - q(x_0)| < \tau \quad \forall x \in B_\delta(x_0).$$

Then we have the following proposition.

PROPOSITION 4.1. *The point x_0 is on the boundary $\partial\Omega$ and $q(x_0) = q_m = \min_{x\in\partial\Omega} q(x)$. Moreover,*

$$c_\lambda \ge \frac{2\pi(q(x_0) - \tau)^2}{\ln\lambda}\left(1 + O\left(\frac{\ln\ln\lambda}{\ln\lambda}\right)\right). \tag{4.1}$$

Proof. By Proposition 3.2, Proposition 3.4, and (3.11), we have

$$|u_\lambda(y)|, |Du_\lambda(y)| \le C\lambda\int_\Omega f(u_\lambda - q(x))\,dx \le \frac{1}{\ln\lambda} \quad \forall y \in \Omega \setminus B_{\frac{\delta}{2}}(x_0). \tag{4.2}$$

Let $\xi \in C_0^\infty(B_\delta(x_0)),\ 0 \le \xi \le 1$ and $\xi = 1$ in $B_{\frac{\delta}{2}}(x_0)$. Then

$$\int_\Omega |D(u_\lambda - \xi u_\lambda)|^2\,dx = \int_{\Omega\setminus B_{\frac{\delta}{2}}(x_0)} |D(u_\lambda - \xi u_\lambda)|^2\,dx \le \frac{C}{\ln^2\lambda}. \tag{4.3}$$

Choose $\tilde{t}_\lambda > 0$ such that

$$\begin{aligned}&\frac{1}{2}\int_{B_\delta(x_0)} |D(\tilde{t}_\lambda \xi u_\lambda)|^2\,dx - \lambda\int_{B_\delta(x_0)} F(\tilde{t}_\lambda \xi u_\lambda - (q(x_0)-\tau))\,dx\\ &= \max_{t\ge 0}\Big(\frac{1}{2}\int_{B_\delta(x_0)} |D(t\xi u_\lambda)|^2\,dx - \lambda\int_{B_\delta(x_0)} F(t\xi u_\lambda - (q(x_0)-\tau))\,dx\Big).\end{aligned}$$

We claim that $\tilde{t}_\lambda$ is bounded. In fact, let $\bar{t}_\lambda > 0$ be such that

$$\begin{aligned}&\frac{1}{2}\int_{B_\delta(x_0)} |D(\bar{t}_\lambda \xi u_\lambda)|^2\,dx - \lambda\int_{B_\delta(x_0)} F(\bar{t}_\lambda \xi u_\lambda - q(x))\,dx\\ &= \max_{t\ge 0}\Big(\frac{1}{2}\int_{B_\delta(x_0)} |D(t\xi u_\lambda)|^2\,dx - \lambda\int_{B_\delta(x_0)} F(t\xi u_\lambda - q(x))\,dx\Big).\end{aligned}$$

From $f(t-q(x)) \le f(t-(q(x_0)-\tau))\ \forall\ x \in B_\delta(x_0)$, we can check that $\tilde{t}_\lambda \le \bar{t}_\lambda$. Thus we just need to prove that $\bar{t}_\lambda$ is bounded.

Let $q_m = \min_{x\in\Omega} q(x)$. Take $R > 0$ large enough such that $\Omega \subset B_R(0)$. Then, we have

$$\begin{aligned}c_\lambda \ge& \inf_{v\in H_0^1(B_R(0)),v\not\equiv 0}\ \max_{t\ge 0}\Big(\frac{1}{2}t^2\int_{B_R(0)} |Dv|^2\,dx - \int_{B_R(0)} F(tv - q_m)\,dx\Big)\\ =&\frac{2\pi q_m^2}{\ln\lambda}(1+o(1)) \ge \frac{b}{\ln\lambda}\end{aligned}$$

for some $b > 0$. Thus, we obtain by (f_2) that

$$\lambda\int_\Omega f(u_\lambda - q(x))u_\lambda\,dx \ge \frac{b'}{\ln\lambda}, \tag{4.4}$$

where $b' > 0$ is a constant.

For $t \in [0,2]$, we define

$$g(t) =: t\int_{B_\delta(x_0)} |D(\xi u_\lambda)|^2\,dx - \lambda\int_{B_\delta(x_0)} f(t\xi u_\lambda - q(x))\xi u_\lambda\,dx.$$

It follows from Proposition 3.4 that for $x \in \Omega\setminus B_{\delta/2}(x_0)$, we have

$$|u_\lambda(x)| \le C\lambda\int_\Omega f(u_\lambda - q(x))\,dx \le \frac{C}{\ln\lambda}.$$

Since q is a positive harmonic function uniformly bounded below in $\bar{\Omega}$, we see that

$$t\xi u_\lambda - q(x) \le 0 \quad \forall\ t\in[0,2],\ x \in \Omega\setminus B_{\delta/2}(x_0)$$

for $\lambda > 0$ large enough. This, together with (4.3), gives

$$g(t) = t\int_\Omega |Du_\lambda|^2\,dx - \lambda\int_\Omega f(tu_\lambda - q(x))u_\lambda\,dx + O\Big(\frac{1}{\ln^2\lambda}\Big). \tag{4.5}$$

Define

$$h(\sigma) = (1+\sigma)\int_\Omega |Du_\lambda|^2\,dx - \lambda\int_\Omega f((1+\sigma)u_\lambda - q(x))u_\lambda\,dx.$$

Then $h(0)=0$, and by (f_2) and (4.4) we get

$$\begin{aligned}
h'(0) =& \int_\Omega |Du_\lambda|^2\,dx - \lambda\int_\Omega f'(u_\lambda - q(x))u_\lambda^2\,dx\\
=&\lambda\int_\Omega \Big(f(u_\lambda - q(x)) - f'(u_\lambda - q(x))u_\lambda\Big)u_\lambda\,dx\\
\le&\lambda\int_\Omega \Big(\theta^{-1}f'(u_\lambda - q(x))(u_\lambda - q(x)) - f'(u_\lambda - q(x))u_\lambda\Big)u_\lambda\,dx\\
\le& -c_1\lambda\int_\Omega f'(u_\lambda - q(x))(u_\lambda - q(x))u_\lambda\,dx\\
\le& -c_2\lambda\int_\Omega f(u_\lambda - q(x))u_\lambda\,dx \le -c_3/\ln\lambda,
\end{aligned}$$

where all the constants $c_i, i=1,2,3$, in the above relation are positive. As a result,

$$h(\sigma) \le -c_3\sigma/\ln\lambda + o(\sigma)$$

if $\sigma>0$. Thus

$$g(1+\sigma) = h(\sigma) + O\Big(\frac{1}{\ln^2\lambda}\Big) \le -c_3\sigma/\ln\lambda + o(\sigma) + O\Big(\frac{1}{\ln^2\lambda}\Big) < 0$$

if $\sigma>0$ is small enough and λ is large enough. Thus, $\bar t_\lambda \le 1+\sigma$. Therefore, we have proved that $\bar t_\lambda$, and thus $\tilde t_\lambda$ is bounded.

Using Proposition 3.4 again, we may deduce as (4.5) that for $\lambda>0$ large enough

$$\begin{aligned}
&\lambda\int_\Omega F(\tilde t_\lambda u_\lambda - q(x))\,dx - \lambda\int_\Omega F(\tilde t_\lambda \xi u_\lambda - q(x))\,dx\\
&= \lambda\int_{\Omega\setminus B_{\delta/2}(x_0)} F(\tilde t_\lambda u_\lambda - q(x))\,dx - \lambda\int_{\Omega\setminus B_{\delta/2}(x_0)} F(\tilde t_\lambda \xi u_\lambda - q(x))\,dx = 0.
\end{aligned} \tag{4.6}$$

By Proposition 2.2, (4.3), and (4.6), we have

$$\begin{aligned}
c_\lambda &= I_\lambda(u_\lambda) = \max_{t\ge0} I_\lambda(tu_\lambda)\\
&\ge I_\lambda(\tilde t_\lambda u_\lambda) = I_\lambda(\tilde t_\lambda\xi u_\lambda) + O\Big(\frac{1}{\ln^2\lambda}\Big)\\
&= \frac12\int_{B_\delta(x_0)} |D(\tilde t_\lambda\xi u_\lambda)|^2\,dx - \lambda\int_{B_\delta(x_0)} F(\tilde t_\lambda\xi u_\lambda - q(x))\,dx + O\Big(\frac{1}{\ln^2\lambda}\Big)\\
&\ge \frac12\int_{B_\delta(x_0)} |D(\tilde t_\lambda\xi u_\lambda)|^2\,dx - \lambda\int_{B_\delta(x_0)} F(\tilde t_\lambda\xi u_\lambda - (q(x_0)-\tau))\,dx + O\Big(\frac{1}{\ln^2\lambda}\Big)\\
&= \max_{t\ge0}\Big(\frac12\int_{B_\delta(x_0)} |D(t\xi u_\lambda)|^2\,dx - \lambda\int_{B_\delta(x_0)} F(t\xi u_\lambda - (q(x_0)-\tau))\,dx\Big) + O\Big(\frac{1}{\ln^2\lambda}\Big)\\
&\ge \inf_{v\in H_0^1(B_\delta(x_0))}\max_{t\ge0}\Big(\frac12\lambda\int_{B_\delta(x_0)} |D(tv)|^2\,dx - \lambda\int_{B_\delta(x_0)} F(tv - (q(x_0)-\tau))\,dx\Big)\\
&+ O\Big(\frac{1}{\ln^2\lambda}\Big)\\
&\ge \frac{2\pi(q(x_0)-\tau)^2}{\ln\lambda}\Big(1 + o\Big(\frac{\ln\ln\lambda}{\ln\lambda}\Big)\Big). \qquad \square
\end{aligned}$$

Proof of Theorem 1.1. We claim that $q(x_0) = q_m = \min_{x\in\partial\Omega} q(x)$. In fact, by Propositions 3.3 and 4.1, we have

$$\frac{2\pi(q(x_0)-\tau)^2}{\ln\lambda}(1-o(1)) \le c_\lambda \le \frac{2\pi(q_m+\tau)^2}{\ln\lambda}(1+o(1)) \quad \forall \tau > 0,$$

where $o(1) \to 0$ as $\lambda \to +\infty$. Therefore,

$$q_m \le q(x_0) \le q_m + 2\tau \quad \forall \tau > 0.$$

This completes the proof.

5. Proof of Theorem 1.2. Let $\bar{x}_0 \in \partial\Omega$ be a strict local minimum point of $q(x)$. Let $\delta > 0$ be small so that

$$q(x) > q(\bar{x}_0) \ \ \forall x \in B_\delta(\bar{x}_0) \cap (\bar{\Omega} \setminus \{\bar{x}_0\}).$$

Let

$$\bar{f}(x,t) = \chi_{B_\delta(\bar{x}_0)} f(t - q(x)),$$

where $\chi_{B_\delta(\bar{x}_0)}$ is the characteristic function of $B_\delta(\bar{x}_0)$. Consider

$$\begin{cases} -\Delta u = \lambda \bar{f}(x,u) & \text{in } \Omega, \\ u \in H_0^1(\Omega). \end{cases} \tag{5.1}$$

It is easy to check that (5.1) has a mountain pass solution $\bar{u}_\lambda$ with critical value $\bar{c}_\lambda$. Then $\bar{c}_\lambda$ has a upper bound as (2.12). Denote by $A_\lambda = \{x \in \Omega : \bar{u}(x)_\lambda > q(x)\}$ the vortex core of $\bar{u}_\lambda$. Then we have the following lemma.

LEMMA 5.1. *A_λ is connected.*

Proof. Let A_i be any component of A_λ. Then $A_i \cap B_\delta(\bar{x}_0) \neq \emptyset$. In fact, if $A_i \cap B_\delta(\bar{x}_0) = \emptyset$, then

$$-\Delta u_\lambda = 0 \ \text{ in } \ A_i; \quad u_\lambda = q(x) \ \text{ on } \ \partial A_i.$$

But $q(x)$ is harmonic; thus $u_\lambda = q(x)$ in A_i. This is a contradiction. The rest of the proof is the same as Proposition 3.1. Since $A_i \cap B_\delta(\bar{x}_0) \neq \emptyset$, η_0 as in the proof of Proposition 3.1 is well defined, and the coefficient in (3.7) in front of s^2 is negative. Then the conclusion follows in the same way. □

Let $x_\lambda \in A_\lambda$ and $x_\lambda \to x_0$.

LEMMA 5.2. $x_0 = \bar{x}_0$.

Proof. First, we have $x_0 \in \overline{B_\delta(\bar{x}_0)} \cap \bar{\Omega}$. If not, we could choose $\tilde{\delta} > 0$ small such that for λ large

$$A_\lambda \subset B_{\tilde{\delta}}(x_0) \subset \bar{\Omega} \setminus \overline{B_\delta(\bar{x}_0)}.$$

Thus, $A_\lambda \cap \overline{B_\delta(\bar{x}_0)} = \emptyset$. This is a contradiction, because as in the proof of Lemma 5.1, we know that $A_\lambda \cap \overline{B_\delta(\bar{x}_0)} \neq \emptyset$.

Since $q(x_0) = \min_{\overline{B_\delta(\bar{x}_0)} \cap \bar{\Omega}} q(x)$ and x_0 is the only point which attains the minimum in $B_\delta(x_0) \cap \bar{\Omega}$, we can prove $x_0 = \bar{x}_0$ in a similar way as in the proof of Theorem 1.1. Since $x_\lambda \to x_0$, we see $A_\lambda \subset B_\delta(x_0)$ for $\lambda > 0$ large. So $f(u_\lambda - q(x)) = 0$ for all $x \in \Omega \setminus B_\delta(x_0)$. As a result, u_λ is a solution of (1.10). □

Acknowledgments. The authors would like to thank the referees for their valuable suggestions.

REFERENCES

[1] A. Ambrosetti and G. Mancini, *On some free boundary problems*, in Recent Contributions to Nonlinear Partial Differential Equations, H. Berestycki and H. Brezis, eds., Pitman, London, 1981, pp. 24–36.
[2] A. Ambrosetti and P. H. Rabinowitz, *Dual variational methods in critical point theory and applications*, J. Funct. Anal., 14 (1973), pp. 349–381.
[3] A. Ambrosetti and M. Struwe, *Existence of steady vortex rings in an ideal fluid*, Arch. Rational Mech. Anal., 108 (1989), pp. 97–109.
[4] A. Ambrosetti and J. Yang, *Asymptotic behaviour in planar vortex theory*, Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl., 1 (1990), pp. 285–291.
[5] T. V. Badiani, *Existence of steady symmetric vortex pairs on a planar domain with an obstacle*, Math. Proc. Cambridge Philos. Soc., 123 (1998), pp. 365–384.
[6] M. S. Berger and L. E. Fraenkel, *Nonlinear desingularization in certain free-boundary problems*, Comm. Math. Phys., 77 (1980), pp. 149–172.
[7] L. A. Caffarelli and A. Friedman, *Asymptotic estimates for the plasma problem*, Duke Math. J., 47 (1980), pp. 705–742.
[8] L. E. Fraenkel and M. S. Berger, *A global theory of steady vortex rings in an ideal fluid*, Acta Math., 132 (1974), pp. 13–51.
[9] B. Gidas, W. M. Ni, and L. Nirenberg, *Symmetry and related properties via the maximum principle*, Comm. Math. Phys., 68 (1979), pp. 209–243.
[10] W. M. Ni, *On the existence of global vortex rings*, J. Anal. Math., 37 (1980), pp. 208–247.
[11] J. Norbury, *Steady planar vortex pairs in an ideal fluid*, Comm. Pure Appl. Math., 28 (1975), pp. 679–700.
[12] B. Turkington, *On steady vortex flow in two dimensions*, I *and* II, Comm. Partial Differential Equations., 8 (1983), pp. 1031–1071.
[13] J. Yang, *Existence and asymptotic behavior in planar vortex theory*, Math. Models Methods Appl. Sci., 1 (1991), pp. 461–475.
[14] J. Yang, *Global vortex rings and asymptotic behaviour*, Nonlinear Anal., 25 (1995), pp. 531–546.

SIAM J. MATH. ANAL.
Vol. 36, No. 5, pp. 1461–1488

BIFURCATIONS OF PERIODIC SOLUTIONS SATISFYING THE ZERO-HAMILTONIAN CONSTRAINT IN REVERSIBLE DIFFERENTIAL EQUATIONS*

R. E. BEARDMORE†, M. A. PELETIER‡, C. J. BUDD§, AND M. AHMER WADEE¶

Abstract. This is a study of the existence of bifurcation branches for the problem of finding even, periodic solutions in fourth-order, reversible Hamiltonian systems such that the Hamiltonian evaluates to zero along each solution on the branch. The class considered here is a generalization of both the Swift–Hohenberg and extended Fisher–Kolmogorov equations that have been studied in several recent papers. We obtain the existence of local bifurcations from a trivial solution under mild restrictions on the nonlinearity and obtain existence and disjointness results regarding the global nature of the resulting bifurcating continua for the case where the Hamiltonian has a *single-well* potential.

The local results rest on two abstract bifurcation theorems which also have applications to sixth-order problems and which show that the curves of zero-Hamiltonian solutions are contained within two-dimensional manifolds of solutions of both negative and positive Hamiltonian.

Key words. reversible Hamiltonian systems, Lyapunov–Schmidt reduction

AMS subject classifications. 37G15, 70H33, 70G70, 37C80, 37C30

DOI. 10.1137/S0036141002418637

1. Introduction.

In [25, 7, 28, 29, 36, 20, 9] the authors find periodic solutions of systems of Hamiltonian differential equations with the property of having *prescribed zero Hamiltonian*. In particular, existence theorems for even periodic orbits satisfying the zero-Hamiltonian constraint in certain fourth-order Hamiltonian systems have been derived by Peletier, Troy, and van den Berg using shooting techniques [35, 36]. While such shooting methods rely heavily on the particular form of the nonlinearities in a given problem and thus suffer from a lack of generality, the techniques do provide a great deal of quantitative information about the solutions. The problem of finding periodic solutions of Hamiltonian systems with prescribed *nonzero energy* has been studied extensively (see [31, 30] and more recently [4]).

The main contribution of this paper is to view the problem of finding zero-Hamiltonian periodic solutions of (1.1) as a one-parameter bifurcation problem from a zero solution, with either *period* or an external parameter playing the role of bifurcation parameter. To solve this bifurcation problem we formulate two abstract Hopf bifurcation theorems (Theorems 2.2 and 2.4) and deduce the existence of the desired

*Received by the editors November 26, 2002; accepted for publication (in revised form) May 21, 2004; published electronically March 25, 2005. This work was supported by TMR grant ERB 4061 PL 97-0159 on degenerate parabolic partial differential equations and by EPSRC grant GR/L17177 of the Applied Nonlinear Mathematics Initiative.

http://www.siam.org/journals/sima/36-5/41863.html

†Department of Mathematics, South Kensington Campus, Imperial College, London SW7 2AZ, United Kingdom (r.beardmore@imperial.ac.uk). The research of this author was supported by Nuffield Foundation Grant NAL/00511/G.

‡Centrum voor Wiskunde en Informatica, Kruislaan 413, NL-1098 SJ Amsterdam, The Netherlands (peletier@cwi.nl).

§Department of Mathematical Sciences, University of Bath, Claverton Down, Bath, BA2 7AY, United Kingdom (cjb@maths.bath.ac.uk).

¶Department of Civil and Environmental Engineering, South Kensington Campus, Imperial College, London SW7 2AZ, United Kingdom (a.wadee@imperial.ac.uk).

solutions as a corollary. The abstract results apply to reversible fourth-order Hamiltonian systems at 1:1 and m:n resonances, provided that the Hamiltonian is indefinite about the trivial solution. We call these *simple* and *double* bifurcations, respectively, as the theorems lead to either a single continuum or a pair of bifurcating continua of solutions. Furthermore, the proofs of the bifurcation theorems are easily modified to show that the zero-Hamiltonian solutions that we find actually lie within manifolds of solutions of positive and negative Hamiltonian.

The proofs of our abstract results are achieved using a Lyapunov–Schmidt reduction technique, as can be found in many texts [38, 2, 11], and the fact that we essentially have only one bifurcation parameter means that some of the global bifurcation results of [6] are applicable. Using arguments from the *configuration-space* formulation of fourth-order problems [34, 17, 29], we shall be able to find bifurcation invariants which demonstrate that the bifurcating continua form a countable collection of mutually disjoint sets. Subsequently, we shall be able to show that a simple bifurcation for fourth-order problems results in the existence of an unbounded (in a suitable sense) continuum of solutions, rather like the classical global Hopf bifurcation theorem described in [1]. The global aspect of the paper is peculiar to fourth-order equations and does not immediately apply to more general Hamiltonian systems (like the sixth-order problem [33, equation (2)] for which we also have local results). Consequently, we have what approaches a nonlinear Sturm–Liouville theory (which is well known in the context of elliptic two-point boundary-value problems [5]) for zero-Hamiltonian solutions of (1.1) given below.

A Lyapunov–Schmidt reduction procedure is specifically available for systems that are either reversible or Hamiltonian [37, 22]; however, we do not make use of those results in this paper. The reason for this is that it is not clear that studying the problem in a space of reduced dimension helps to elucidate the role played by the Hamiltonian constraint, and consequently we approach the problem ab initio.

So, consider the class of fourth-order differential equations

$$u'''' + pu'' + F_u(u) = 0, \tag{1.1}$$

where primes refer to differentiation with respect to x, p is a real parameter, and the function $F \in C^\omega(\mathbb{R})$ satisfies

(**F**) $F(0) = F_u(0) = 0 \quad \text{and} \quad F_{uu}(0) = 1.$

Here, $C^\omega(\mathbb{R})$ denotes the space of real-analytic functions on $\mathbb{R}$, and a subscript u denotes differentiation with respect to u. We shall assume *throughout* that F satisfies assumption (**F**) and is therefore positive in some neighborhood of $u = 0$. Note that the final condition in (**F**) is not restrictive as it can always be obtained from a suitable scaling of u and of time (denoted x), provided that $F_{uu}(0) > 0$.

Now (1.1) is reversible (see [8] for a discussion of reversible systems) and Hamiltonian, with Hamiltonian

$$H \equiv u'u''' - \frac{1}{2}u''^2 + \frac{1}{2}pu'^2 + F(u), \tag{1.2}$$

and when suitably scaled (see [36]), (1.1) provides the *extended Fisher–Kolmogorov* and *Swift–Hohenberg equations*, with $F_u(u) = \pm u(1 - u^2)$.

Fourth-order equations like (1.1) have a burgeoning literature, as can be seen from the recent studies in [3, 7, 24, 26, 27, 28, 29, 35, 36, 8, 12, 19, 18]; see, in particular, the recent monograph [25] and also [36, 20]. In these references it is shown, using a variety of variational, geometric, functional analytic, and elementary techniques, that

(1.1) may possess periodic, homoclinic, and heteroclinic solutions, with applications ranging from geology to buckling theory; in particular, zero-Hamiltonian periodic solutions play an important role in the study of cellular buckling (see [12, 19, 8] and the references therein; see also [9]).

We note at this stage that in order for a bifurcation from the trivial solution to occur as p varies, no further restrictions will be required on the nonlinearity F than those already given in assumption (**F**).

As an application of the results of the first part of the paper, we analyze the behavior of the simple bifurcating branch which connects to $(u,p) = (0,2)$ for the case

$$F(u;\epsilon) = \frac{1}{2}u^2 - \epsilon\left(\frac{1}{4}u^4 - \frac{1}{6}u^6\right), \tag{1.3}$$

where ϵ is a parameter that unfolds the degenerate problem from $\epsilon = 0$. In particular, we prove the existence of a fold bifurcation on this bifurcating branch, which was conjectured to exist in [19] and [12]. Finally, the results of some numerical calculations performed in AUTO will be presented, which indicate that similar behavior is observed for the multiple bifurcating branches which connect to the trivial solution at $p > 2$. We also compute the symmetry-breaking bifurcations on these branches and illustrate the subsequent connecting branches of solutions.

2. Bifurcation theorems. Let X, Y, and Z be Banach spaces, and $BL(X,Y)$ denote the space of continuous (bounded) linear maps from X to Y. We write X^* for the dual space of continuous linear functionals $BL(X,\mathbb{R})$. If $L \in BL(X,Y)$ and $U \subset X$ is a closed subspace of X, then $L|_U \in BL(U,Y)$ will denote the restriction of L to U. We shall use $\|\cdot\|_X$ to denote the norm on X, and $\mathrm{Iso}(X,Y)$ denotes the set of continuous linear isomorphisms from X to Y. Let $C^r_{2\pi}$ be the Banach space of 2π-periodic C^r functions from $[0,2\pi]$ to $\mathbb{R}^n$, endowed with a C^r norm.

If $f : X \to Z$ is a given smooth mapping, then $df(x)[h]$ will denote the Fréchet derivative of f. For higher derivatives, the k-form $d^k f(x)[h,\dots,h]$ will also be written as $d^k f(x)[h]^{(k)}$ for brevity. Partial derivatives of a function $f \in C^1(X \times Y, Z)$ will be written as $d_x f(x,y)[h] \in Z$ and $d_y f(x,y)[k] \in Z$, where $(h,k) \in X \times Y$, and higher derivatives will be written as in $d^2_{xy} f(x,y)[h,k]$. If $X = \mathbb{R}$, we will identify $d_x f(x,y)[h]$ with $h d_x f(x,y)[1]$, and we shall also write

$$d^k f(x)[h_1,\dots,h_k] = (\Pi_{j=1}^k h_j) d^k f(x)[1,\dots,1],$$

although we shall often omit the k-vector $[1,\dots,1]$ in this expression where no confusion results. Given $u, u_1, u_2 \in X$, we will write $\langle u\rangle \equiv \mathbb{R}\cdot u$ and $\langle u_1, u_2\rangle = \{\alpha_1 u_1 + \alpha_2 u_2 : \alpha_1, \alpha_2 \in \mathbb{R}\}$. For any continuous function v, we denote the delta functional by $\delta(v) \equiv v(0)$.

For completeness, let us recall the following. A linear mapping $L \in BL(X,Y)$ is said to be *Fredholm* if its range $\mathrm{ran}(L)$ is a closed subspace of Y with finite codimension and its null space $\ker(L)$ is a finite-dimensional subspace of X. Then

$$\mathrm{ind}(L) = \dim\ker(L) - \mathrm{codim}\ \mathrm{ran}(L)$$

is said to be the *Fredholm index* of L. We recall the following theorem, which gives a useful collections of facts that can be found in [32].

THEOREM 2.1. *If $L \in BL(X,Y)$ is Fredholm and $K \in BL(X,Y)$ is a compact linear operator, then $L + K \in BL(X,Y)$ is also Fredholm and* $\mathrm{ind}(L+K) = \mathrm{ind}(L)$.

As a consequence, if L is Fredholm of index zero and is injective, then it is an isomorphism. The set $\mathcal{F}$ of Fredholm operators is open in $BL(X,Y)$, and ind *is constant on connected components of $\mathcal{F}$.*

2.1. Statement of the abstract problem. Let $\psi \in Z^*, M \in C^\omega(X \times \mathbb{R}^2, Y)$, and $g \in C^\omega(X \times \mathbb{R}^2, Z)$; now consider the bifurcation problem

$$\mathcal{H}(u,p,\mu) \equiv \begin{pmatrix} M(u,p,\mu) \\ \psi(g(u,p,\mu)) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{2.1}$$

where $u \in X$ and $\mu, p \in \mathbb{R}$ are parameters. Throughout, we shall write $\mathcal{H} = M \times \psi(g)$ for brevity. We intend that (2.1) represent an abstract formulation of finding periodic solutions of (1.1) with the property of having zero Hamiltonian; accordingly we shall call the functional $\psi \circ g$ the *energy* of (2.1). Since the functional H in (1.2) is of *quadratic order* at the origin, we impose this degree of degeneracy into the operator g. Hence we assume that

$$M(0,p,\mu) \equiv 0 \tag{2.2}$$

and

$$g(0,p,\mu) \equiv 0, \qquad d_u g(0,p,\mu) \equiv 0. \tag{2.3}$$

By the term *bifurcation* from the trivial solution $u = 0$ of (2.1) at $(u,p,\mu) = (0,p_0,\mu_0)$ we mean that there is a sequence $(u_n,p_n,\mu_n) \subset X\backslash\{0\} \times \mathbb{R}^2$ which satisfies

$$\mathcal{H}(u_n,p_n,\mu_n) \equiv 0, \quad u_n \to 0 \quad \text{and} \quad (p_n,\mu_n) \to (p_0,\mu_0) \quad \text{as } n \to \infty.$$

2.2. Local bifurcations. Let us now seek conditions under which there is a bifurcation of (2.1) from the trivial solution. The implicit function theorem applied to (2.1) shows that $(0,p_0,\mu_0)$ can be a bifurcation point for (2.1) only if

$$d_u M(0,p_0,\mu_0) \notin \mathrm{Iso}(X,Y). \tag{2.4}$$

Furthermore, if M is a assumed to be a Fredholm mapping, then a bifurcation can occur only when $d_u M(0,p_0,\mu_0)$ is *not* injective. Motivated by this, we shall now consider two such cases:

(i) $\dim\ker(d_u M(0,p_0,\mu_0)) = 1$,
(ii) $\dim\ker(d_u M(0,p_0,\mu_0)) = 2$.

Case (i) is reminiscent of the theorem on bifurcation from a simple eigenvalue and will give rise to a *unique* bifurcating continuum. In case (ii), however, we will be able to locate exactly *two* distinct bifurcating continua. Let us now proceed with the promised results.

THEOREM 2.2 (simple abstract Hopf bifurcation). *Suppose that* (2.2)–(2.3) *hold and that $d_u M(0,p_0,\mu_0) \in BL(X,Y)$ is Fredholm of index zero, where* $\ker(d_u M(0,p_0,\mu_0)) = \langle k \rangle$. *Suppose also that $X = \langle k \rangle \oplus U$, $V = \mathrm{ran}(d_u M(0,\mu_0,p_0), Y = \langle K \rangle \oplus V, P : Y \to V$ is the projection operator along $\langle K \rangle$, and Q is the projection onto $\langle K \rangle$ which is identified with $\mathbb{R}$.*

Suppose further that $\psi(d^2_{uu} g(0,p_0,\mu_0)[k,k]) = 0$ and that the operator $D \in BL(U \times \mathbb{R}^2, V \times \mathbb{R}^2)$ given by

$$D = \begin{pmatrix} P d_u M(0,p_0,\mu_0) & P d^2_{up} M(0,p_0,\mu_0)[k,1] & P d^2_{u\mu} M(0,p_0,\mu_0)[k,1] \\ 0 & Q d^2_{up} M(0,p_0,\mu_0)[k,1] & Q d^2_{u\mu} M(0,p_0,\mu_0)[k,1] \\ \psi(d^2_{uu} g[k,\cdot]) & \psi(d^3_{uup} g[k,k,1]) & \psi(d^3_{uu\mu} g[k,k,1]) \end{pmatrix}$$

is an isomorphism.

Then, $(0, p_0, \mu_0)$ *is a bifurcation point for* (2.1). *Moreover, there is an interval* I *containing* 0 *and a unique analytic branch* $\mathcal{B}$ *of solutions of* (2.1) *on which* $(u, p, \mu) = (u(\beta), p(\beta), \mu(\beta))$ *for* $\beta \in I$ *and which satisfies* $u(\beta) \neq 0$ *for* $\beta \neq 0, (u(0), p(0), \mu(0)) = (0, p_0, \mu_0)$. *Moreover, there results* $\|u(\beta) - \beta k\|_X = O(\beta^2)$ *as* $\beta \to 0$.

Proof. Let us express u in terms of the decomposition of X as $u = \beta k + r = \beta(k + \rho) \in \langle k \rangle \oplus U$; then (2.1) is equivalent to

$$(P + Q)M(\beta(k + \rho), p, \mu) = 0, \tag{2.5}$$

$$\psi(g(\beta(k + \rho), p, \mu)) = 0. \tag{2.6}$$

Using analyticity, it follows that there are analytic mappings $\tilde{M}$ and $\tilde{g}$ such that

$$M(\beta(k + \rho), p, \mu) = \beta d_u M(0, p, \mu)[k + \rho] + \beta^2 \tilde{M}(\beta, \rho, p, \mu)$$

and

$$\psi(g(\beta(k + \rho), p, \mu)) = \psi\left(\frac{\beta^2}{2} d^2_{uu} g(0, p, \mu)[k + \rho, k + \rho] + \frac{\beta^3}{2} \tilde{g}(\beta, \rho, p, \mu)\right).$$

As we are seeking nonzero solutions to (2.1), we can divide by appropriate powers of β in (2.5)–(2.6) and solve the equivalent problems

$$(P + Q)d_u M(0, p, \mu)[k + \rho] + \beta \tilde{M}(\beta, \rho, p, \mu) = 0, \tag{2.7}$$

$$\psi\left(d^2_{uu} g(0, p, \mu)[k + \rho, k + \rho] + \beta \tilde{g}(\beta, \rho, p, \mu)\right) = 0. \tag{2.8}$$

In turn, (2.7)–(2.8) is equivalent to

$$P\big(d_u M(0, p, \mu)[k + \rho] + \beta \tilde{M}(\beta, \rho, p, \mu)\big) = 0 \in V, \tag{2.9}$$

$$Q\big(d_u M(0, p, \mu)[k + \rho] + \beta \tilde{M}(\beta, \rho, p, \mu)\big) = 0 \in \mathbb{R}, \tag{2.10}$$

$$\psi\left(d^2_{uu} g(0, p, \mu)[k + \rho, k + \rho] + \beta \tilde{g}(\beta, \rho, p, \mu)\right) = 0 \in \mathbb{R}, \tag{2.11}$$

where the one-dimensional space $\langle K \rangle$ is identified with $\mathbb{R}$.

Let us now define (2.9)–(2.11) as $\Phi_1(\beta, \rho, p, \mu) = 0$, where Φ_1 is an analytic mapping of Banach spaces

$$\Phi_1 : \mathbb{R} \times U \times \mathbb{R}^2 \to V \times \mathbb{R}^2. \tag{2.12}$$

Under the stated assumptions it is clear that $\Phi_1(0, 0, p_0, \mu_0) = 0$, and one can show that $D = d_{\rho,p,\mu}\Phi_1(0, 0, p_0, \mu_0)$, noting $Q[d_u M(0, p_0, \mu_0)] = 0$ by definition. It now follows by the implicit function theorem that we may locally solve (2.9)–(2.10) for ρ, p, and μ as a function of β. The fact that $\rho(0) = 0$ completes the proof. ∎

The following result tells us that Theorem 2.2 is a special case of a more general result which says that the branch $\mathcal{B}$ of zero energy solutions from this theorem is formed from the intersection of a manifold of solutions of $M(u, p, \mu) = 0$ with the zero-energy surface $\{(u, p, \mu) : \psi(g(u, p, \mu)) = 0\}$.

THEOREM 2.3. *The curve of zero-energy solutions* $\mathcal{B}$ *from Theorem* 2.2 *is contained within a (locally) two-dimensional analytic manifold* $\mathcal{M}$ *of solutions of*

$$M(u, p, \mu) = 0 \tag{2.13}$$

of both positive and negative energy.

Proof (sketch). Repeat the same argument as in Theorem 2.2 but for the system

$$\mathcal{H}_\epsilon(u,p,\mu) \equiv \begin{pmatrix} M(u,p,\mu) \\ \psi(g(u,p,\mu)) - \epsilon \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{2.14}$$

One obtains a mapping $\Phi_1(\beta,\rho,p,\mu,\epsilon)$, entirely analogous to (2.12), such that

$$\Phi_1(0,0,p_0,\mu_0,0) = 0 \quad \text{and} \quad D = d_{\rho,p,\mu}\Phi_1(0,0,p_0,\mu_0,0).$$

One can then solve ρ, p, and μ locally as analytic functions of (β,ϵ). ∎

The following technical lemma shows that one can parameterize the bifurcating branch from Theorem 2.2 using one of p or μ as parameters, and this result will be used at a later stage in the paper.

LEMMA 2.1. *If $(u,p,\mu) = (u(\beta),p(\beta),\mu(\beta))$ is an element of the bifurcating branch $\mathcal{B}$ obtained in Theorem* 2.2, *then at least one of*

$$d_{u,p}(M \times \psi(g)) \quad \text{or} \quad d_{u,\mu}(M \times \psi(g)) \in BL(X \times \mathbb{R}, Y \times \mathbb{R})$$

(evaluated at (u,p,μ)) is an isomorphism for sufficiently small nonzero $|\beta|$.

Proof. Note from Theorem 2.2 that

$$\left|Qd^2_{up}M(0,p_0,\mu_0)[k,1]\right| + \left|Qd^2_{u\mu}M(0,p_0,\mu_0)[k,1]\right| \neq 0,$$

and let us therefore assume for definiteness that

$$Qd^2_{up}M(0,p_0,\mu_0)[k,1] \neq 0. \tag{2.15}$$

Now define the one-parameter family of linear mappings

$$L(\beta) \equiv \begin{pmatrix} d_uM(u(\beta),p(\beta),\mu(\beta)) & d_pM(u(\beta),p(\beta),\mu(\beta)) \\ \psi(d_ug(u(\beta),p(\beta),\mu(\beta)) & \psi(d_pg(u(\beta),p(\beta),\mu(\beta)) \end{pmatrix}$$

and note that this is an (at most) rank-two perturbation of a Fredholm mapping with index zero. It follows that we need to prove only that $L(\beta)$ is injective for $\beta \neq 0$.

Using analyticity, a straightforward but lengthy calculation shows that we may write $L(\beta) = L_0(\beta) + \beta L_1(\beta) + \frac{\beta^2}{2}L_2(\beta) + O(\beta^3)$, where

$$L_0(\beta) = \begin{pmatrix} d_uM & 0 \\ 0 & 0 \end{pmatrix},$$

$$L_1(\beta) = \begin{pmatrix} d^2_{uu}M[k+O(\beta),\cdot] & d^2_{up}M[k+O(\beta),\cdot] \\ \psi(d^2_{uu}g[k+O(\beta),\cdot]) & 0 \end{pmatrix},$$

and

$$L_2(\beta) = \begin{pmatrix} d^3_{uuu}M[k+O(\beta),k+O(\beta),\cdot] & d^3_{uup}M[k+O(\beta),k+O(\beta),\cdot] \\ \psi(d^3_{uuu}g[k+O(\beta),k+O(\beta),\cdot]) & \psi(d^3_{uup}g[k+O(\beta),k+O(\beta),\cdot]) \end{pmatrix},$$

where each of the given derivatives is evaluated at $(u,p,\mu) = (0,p(\beta),\mu(\beta))$. For $\beta \neq 0$, one can see that $L(\beta)$ is injective if and only if $T(\beta)$ is, where

$$\begin{aligned} T(\beta) = &\begin{pmatrix} d_uM & 0 \\ \psi(d^2_{uu}g[k+O(\beta),\cdot]) & 0 \end{pmatrix} \\ &+ \frac{\beta}{2}\begin{pmatrix} 2d^2_{uu}M[k+O(\beta),\cdot] & 2d^2_{up}M[k+O(\beta),\cdot] \\ \psi(d^3_{uuu}g[k+O(\beta),k+O(\beta),\cdot]) & \psi(d^3_{uup}g[k+O(\beta),k+O(\beta),\cdot]) \end{pmatrix} \\ &+ O(\beta^2). \end{aligned}$$

Again, each of the derivatives is evaluated at $(u,p,\mu)=(0,p(\beta),\mu(\beta))$.

Clearly $T(0)$ is not injective, but the fact that $d^2_{uu}g(0,p_0,\mu_0)[k,k]\neq 0$ implies

$$\ker(T(0))=\langle\kappa\rangle\subset X\times\mathbb{R},$$

where $\kappa=(0_X,1)\in X\times\mathbb{R}$. To prove that $T(\beta)$ is injective for small β it suffices to prove that

$$T'(0)\kappa\notin \operatorname{ran}(T(0)). \tag{2.16}$$

However

$$T'(0)=\frac{1}{2}\begin{pmatrix} 2d^2_{uu}M[k,\cdot] & 2d^2_{up}M[k,\cdot]\\ \psi(d^3_{uuu}g[k,k,\cdot]) & \psi(d^3_{uup}g[k,k,\cdot])\end{pmatrix},$$

evaluating derivatives at $(u,p,\mu)=(0,p_0,\mu_0)$, and $\operatorname{ran}(T(0))=\operatorname{ran}(d_uM(0,p_0,\mu_0))\times\operatorname{ran}(\psi(d^2_{uu}g(0,p_0,\mu_0)))$. It follows that $T'(0)\kappa\in\operatorname{ran}(T(0))$ can be satisfied only if

$$d^2_{up}M(0,p_0,\mu_0)[k,1]\in\operatorname{ran}(d_uM(0,p_0,\mu_0)),$$

but this contradicts (2.15). Finally, one can use an analogous argument to cover the case whereby $Qd^2_{u\mu}M(0,p_0,\mu_0)[k,1]\neq 0$. □

Next we consider case (ii), where $d_uM(0,p_0,\mu_0)$ has a two-dimensional null-space.

THEOREM 2.4 (double abstract Hopf bifurcation). *Suppose that* (2.2)–(2.3) *hold and* $\ker(d_uM(0,p_0,\mu_0))=W$, *where* $\dim(W)=2$ *with* $W=\langle k_1,k_2\rangle$. *Suppose further that* $X=W\oplus U$ *and*

$$Y=Z\oplus V,\qquad V=\operatorname{ran}(d_uH(0,p_0,\mu_0)),$$

where V *is closed,* $\dim(Z)=2$, *and* $Z=\langle u_1,u_2\rangle$.

Now, let $P:Y\to V$ *be the projection along* Z *and* $Q=I-P$. *For* $i=1,2$, *let* Q_i *be the projection of* Y *onto* $\langle u_i\rangle$ *(which we identify with* $\mathbb{R}$*) such that* $Q[y]=Q_1[y]u_1+Q_2[y]u_2$. *Set*

$$A=\psi(d^2_{uu}g(0,p_0,\mu_0)[k_2,k_2]),\qquad B=\psi(d^2_{uu}g(0,p_0,\mu_0)[k_1,k_2])$$

and

$$C=\psi(d^2_{uu}g(0,p_0,\mu_0)[k_1,k_1]).$$

Suppose that $C\neq 0, B^2>AC$, *and let* $\alpha_\pm$ *be the two (real nonzero distinct) roots of the quadratic equation* $A\alpha^2+2B\alpha+C=0$. *Suppose also that*

$$\det\begin{pmatrix} Q_1d^2_{up}M[k_1+\alpha_\pm k_2,1] & Q_1d^2_{u\mu}M[k_1+\alpha_\pm k_2,1]\\ Q_2d^2_{up}M[k_1+\alpha_\pm k_2,1] & Q_2d^2_{u\mu}M[k_1+\alpha_\pm k_2,1]\end{pmatrix}\neq 0$$

when $(u,p,\mu)=(0,p_0,\mu_0)$.

Then $(0,p_0,\mu_0)$ *is a bifurcation point for* (2.1). *Moreover, there is an interval* I *containing* 0 *and exactly two analytic branches* $\mathcal{B}_\pm$ *of solutions of* (2.1) *on which* $(u,p,\mu)=(u_\pm(\beta),p_\pm(\beta),\mu_\pm(\beta))$ *for* $\beta\in I$, *with* $u_\pm(0)\neq 0$ *for* $\beta\neq 0$ *and* $(u_\pm(0),p_\pm(0),\mu_\pm(0))=(0,p_0,\mu_0)$. *Moreover, there are analytic functions* $\alpha_\pm:I\to\mathbb{R}$ *and* $\rho:I\to V$ *such that* $\alpha_\pm(0)=\alpha_\pm$, $\|\rho(\beta)\|_Y=O(\beta)$ *as* $\beta\to 0$, *and* $u_\pm(\beta)=\beta k_1+\beta\alpha_\pm(\beta)k_2+\beta\rho(\beta)$.

Proof. Using the analyticity of M, let us write

$$M(u,p,\mu) = d_u M(0,p,\,\mu)[u] + \mathcal{O}(2)$$

and

$$g(u,p,\mu) = g(0,p,\mu) + d_u g(0,p,\mu)u + \frac{1}{2}d^2_{uu}g(0,p,\mu)[u,u] + \mathcal{O}(3),$$

where $\mathcal{O}(n)$ represents any function, $\Theta(u,p,\mu)$ say, where there is a $\gamma > 0$ such that $\|\Theta(u,p,\mu)\| \le \gamma\|u\|^n$ for all (u,p,μ) in a neighborhood of $(0,p_0,\mu_0)$.

Now let $u = \beta(k_1+\alpha k_2+\alpha\rho) \in W \oplus U$, and note that there is an analytic function ϕ_1 such that the equation $M(u,p,\mu) = 0$ is locally equivalent to

$$\beta d_u M(0,p,\mu)[k_1+\alpha k_2+\alpha\rho] + \beta^2\phi_1(\alpha,\beta,\rho,p,\mu) = 0. \tag{2.17}$$

We may also use the analyticity of g to write

$$g(\beta(k_1+\alpha k_2+\alpha\rho),p,\mu) = \frac{\beta^2}{2}d^2_{uu}g(0,p,\mu)[k_1+\alpha k_2+\alpha\rho]^{(2)} + \frac{\beta^3}{2}\phi_2(\alpha,\beta,\rho,p,\mu), \tag{2.18}$$

where ϕ_2 is another suitably defined analytic function. Now the equation $M(u,p,\mu) = 0$ is equivalent to

$$(P + u_1Q_1 + u_2Q_2)M(u,p,\mu) = 0,$$

and, after dividing (2.17) and (2.18) by β and β^2, respectively, we obtain the locally equivalent problem

$$P\left[d_u M(0,p,\mu)[k_1+\alpha k_2+\alpha\rho] + \beta\phi_1(\alpha,\beta,\rho,p,\mu)\right] = 0 \in V, \tag{2.19}$$

$$Q_1\left[d_u M(0,p,\mu)[k_1+\alpha k_2+\alpha\rho] + \beta\phi_1(\alpha,\beta,\rho,p,\mu)\right] = 0 \in \mathbb{R}, \tag{2.20}$$

$$Q_2\left[d_u M(0,p,\mu)[k_1+\alpha k_2+\alpha\rho] + \beta\phi_1(\alpha,\beta,\rho,p,\mu)\right] = 0 \in \mathbb{R}, \tag{2.21}$$

$$\psi\big(d^2_{uu}g(0,p,\mu)[k_1+\alpha k_2+\alpha\rho]^{(2)} + \beta\phi_2(\alpha,\beta,\rho,p,\mu)\big) = 0 \in \mathbb{R}. \tag{2.22}$$

Setting $\beta = 0, \rho = 0, p = p_0$, and $\mu = \mu_0$ in (2.19)–(2.22), we find an equation for α:

$$\psi(d^2_{uu}g(0,p,\mu)[k_1+\alpha k_2, k_1+\alpha k_2]) = 0. \tag{2.23}$$

From the definitions of A, B, and C given in the statement of the theorem, (2.23) is simply the equation $C + 2B\alpha + A\alpha^2 = 0$ with solutions $\alpha = \alpha_\pm$.

We now write (2.19)–(2.22) as $\Phi_2(\rho,p,\mu,\alpha,\beta) = 0$, say, where Φ_2 is an analytic mapping $\Phi_2 : U \times \mathbb{R}^4 \to V \times \mathbb{R}^3$. The derivative $d_{\rho,p,\mu,\alpha}\Phi_2(0,p_0,\mu_0,\alpha_\pm,0)$ is given by the operator matrix

$$L \equiv \begin{pmatrix} A_0 & B_0 & 0 \\ 0 & D_0 & 0 \\ E_0 & F_0 & G_0 \end{pmatrix} \in BL(U \times \mathbb{R}^3, V \times \mathbb{R}^3),$$

where

$$A_0 = \alpha_\pm P d_u M^0, \qquad B_0 = \left[P d^2_{up}M^0[k_1+\alpha_\pm k_2,1]\middle|\, P d^2_{u\mu}M^0[k_1+\alpha_\pm k_2,1]\right],$$

$$E_0 = 2\alpha_\pm \psi(d^2_{uu} g^0[\cdot, k_1 + \alpha_\pm k_2]), \qquad G_0 = 2\psi(d^2_{uu} g^0[k_2, k_1 + \alpha_\pm k_2]),$$

$$F_0 = \left[\psi(d^3_{uup} g^0[1, k_1 + \alpha_\pm k_2, k_1 + \alpha_\pm k_2]) \middle| \psi(d^3_{uu\mu} g^0[1, k_1 + \alpha_\pm k_2, k_1 + \alpha_\pm k_2])\right]$$

and

$$D_0 = \begin{pmatrix} Q_1 d^2_{up} M^0[k_1 + \alpha_\pm k_2, 1] & Q_1 d^2_{u\mu} M^0[k_1 + \alpha_\pm k_2, 1] \\ Q_2 d^2_{up} M^0[k_1 + \alpha_\pm k_2, 1] & Q_2 d^2_{u\mu} M^0[k_1 + \alpha_\pm k_2, 1] \end{pmatrix},$$

and a superscript zero $(^0)$ denotes evaluation of a function at $(u, p, \mu) = (0, p_0, \mu_0)$.

Clearly, for L to be an isomorphism we require $G_0 \neq 0$, that is, $\psi(d^2_{uu} g^0[k_2, k_1 + \alpha_\pm k_2]) \neq 0$, but this is just $B + \alpha_\pm A \neq 0$, which is true by assumption. Since $A_0 \in \mathrm{Iso}(U, V)$, L is an isomorphism if $\det(D_0) \neq 0$, and this is also an assumption. Using the implicit function theorem, we can now determine all of the variables as analytic functions of β locally to the *two points* $(\rho, p, \mu, \alpha, \beta) = (0, p_0, \mu_0, \alpha_\pm, 0)$. □

As was the case for Theorems 2.2 and 2.3, we can prove that the branches of zero-energy solutions $\mathcal{B}_\pm$ from Theorem 2.4 are obtained from the intersection of solutions of $M(u, p, \mu) = 0$ with the zero-energy surface.

THEOREM 2.5. *The two curves of zero-energy solutions $\mathcal{B}_\pm$ from Theorem* 2.4 *are each contained within (locally) two-dimensional analytic manifolds $\mathcal{M}_\pm$ of solutions of $M(u, p, \mu) = 0$ of both positive and negative energy. Moreover, $\mathcal{M}_+ \cap \mathcal{M}_- = \{(0, p_0, \mu_0)\}$.*

Proof. This is almost a verbatim repetition of the proof of Theorem 2.4, but modified to deal with an energy constraint of the form $\psi(g(u, p, \mu)) = \epsilon$. □

3. The existence of bifurcations for fourth- and sixth-order systems.

3.1. Preliminaries. In this section we shall apply Theorems 2.1 and 2.2 to find bifurcating branches of periodic solutions of (1.1) which have the zero-Hamiltonian property. To do so, we shall presume that a periodic solution of (1.1) has period T, where

$$T = \frac{2\pi}{\mu}$$

and μ is a priori unknown. Upon setting

$$t = \mu x,$$

a simple rescaling of (1.1) and (1.2) leads us to consider the *two-parameter* problem

$$M(u, p, \mu) \equiv \mu^4 u'''' + p\mu^2 u'' + F_u(u) = 0, \tag{3.1}$$

$$\psi(g(u, p, \mu)) \equiv \delta \left(\mu^4 u' u''' - \frac{1}{2}\mu^4 u''^2 + \frac{1}{2} p \mu^2 u'^2 + F(u) \right) = 0, \tag{3.2}$$

where primes now denote differentiation with respect to t. With regard to (2.1), ψ corresponds to δ, and g is the Hamiltonian which appears in (3.2); again we shall write $\mathcal{H} = M \times \psi(g)$ so that (3.1)–(3.2) corresponds to the equation $\mathcal{H} = 0$.

A natural setting for the application of these Theorems 2.1 and 2.2 is in the space of even functions of period 2π. Accordingly, let $X_e = \left\{ u \in C^4_{2\pi} : u(t) = u(-t) \right\}$ and

$Y_e = \{u \in C^0_{2\pi} : u(t) = u(-t)\}$, both endowed with their usual norms. We also define the *even-odd* subspaces

$$X_{eo} = \left\{ u \in X_e : u\left(\frac{\pi}{2} - t\right) = -u\left(\frac{\pi}{2} + t\right) \right\}$$

and

$$Y_{eo} = \left\{ u \in Y_e : u\left(\frac{\pi}{2} - t\right) = -u\left(\frac{\pi}{2} + t\right) \right\}.$$

For a given subspace $S \subset Z \subset L^2(0,\pi)$ we define its orthogonal complement by $S^\perp = \{u \in Z : \int_0^{2\pi} u(t)s(t)dt = 0 \ \ \forall s \in S\}$. In this way we obtain a map $\mathcal{H} \in C^\omega(X_e \times \mathbb{R}^2, Y_e \times \mathbb{R})$, and if F is even, then $\mathcal{H}$ also provides a map $\mathcal{H} \in C^\omega(X_{eo} \times \mathbb{R}^2, Y_{eo} \times \mathbb{R})$.

3.2. Simple bifurcation from $p = 2$. The following theorem shows that the zero-Hamiltonian problem associated with (1.1) has a simple bifurcation point to a locally unique and smooth branch of solutions from the point $p = 2$.

THEOREM 3.1. *Suppose that assumption* (**F**) *holds. Then there is an interval* $I \subset \mathbb{R}$ *and a unique analytic branch* $\beta \mapsto (u(\beta), p(\beta)) \in X_e \times \mathbb{R}$ *defined on* I *of nontrivial even periodic solutions of* (1.1) *with* zero Hamiltonian *and period* $T(\beta)$. *Moreover,* $u(\beta) \neq 0$ *if* $\beta \neq 0$,

$$T(0) = 2\pi, \ u(0) = 0, \ p(0) = 2, \text{ and } \|u(\beta)(t) - \beta\cos(t)\|_{C^4} = O(\beta^2)$$

as $\beta \to 0$. *If* F *is even, then the function* $t \mapsto u(\beta)(t)$ *is an element of* X_{eo} *for all* $\beta \in I$.

Proof. To prove this result we apply Theorem 2.2 to $\mathcal{H}(u, p, \mu) = 0$ with $X = X_e$ and $Y = Y_e$. Let $L(p,\mu)[a] \equiv d_u M(0,p,\mu)[a] = \mu^4 a'''' + p\mu^2 a'' + a$, and note that the bilinear form $\psi(d^2_{uu} g(0,p,\mu)[a,b])$ from Theorem 2.2 is given by

$$B(p,\mu)[a,b] = \delta\left(\mu^4(a'b''' + a'b''' - a''b'') + p\mu^2 a'b' + ab\right).$$

In order to verify the hypotheses of Theorem 2.2 let us seek a nonzero solution $a \in X_e$ to $L(p,\mu)a = 0$, that is,

$$\mu^4 a'''' + p\mu^2 a'' + a = 0, \quad \delta\left(\mu^4(2a'a''' - (a'')^2) + p\mu^2(a')^2 + a^2\right) = 0.$$

Since a is even and of period 2π, we seek solutions of the form $a(t) = \cos(mt)$, where m is an integer. This provides the equations $\alpha^4 - p\alpha^2 + 1 = 0, -\alpha^4 + 1 = 0$, where $\alpha = \mu m$, whence $\alpha^2 = 1$, so that $p = 2$ and $\mu = 1/m$. Seeking the solution of minimal period, we may set $m = 1$ and thus define $k(t) \equiv \cos(t)$ and record the fact that $\ker(L(2,1)) = \langle k \rangle$. Let us also define $K \equiv k$ for the purposes of Theorem 2.2 and note for the moment that k is an *even-odd* function.

We now form the decompositions $X_e = \langle k \rangle \oplus \langle k \rangle^\perp$ and $Y_e = \langle k \rangle \oplus \langle k \rangle^\perp$, so that $U = \langle k \rangle^\perp \subset X_e$ and $V = \langle k \rangle^\perp \subset Y_e$ in accordance with Theorem 2.2, and define the projection $Q : L^2(0, 2\pi) \to \mathbb{R}$ by

$$(Qu)(t) = \frac{1}{\pi}\int_0^{2\pi} u(t)k(t)dx, \quad u \in L^2(0, 2\pi),$$

and let $P = I - k \cdot Q$. Now, $\psi(d^3_{uup} g(0,p,\mu)[a,b,1]) = B_p(p,\mu)[a,b] = \delta(\mu^2 a'b')$, $d^3_{uu\mu} g(0,p,\mu)[a,b] = \delta(4\mu^3(a'b''' + a'b''' - a''b'') + 2p\mu a'b')$, and $d^2_{u\mu} M(0,p,\mu)[a] =$

$L_\mu(p,\mu)[a] = 4\mu^3 a'''' + 2p\mu a''$. Finally, $d^2_{up}M(0,p,\mu)[a,1] = \mu^2 a''$. Thus the remaining hypothesis of Theorem 2.2 is satisfied if the operator matrix D is nonsingular. On inspection of the relevant derivatives we find

$$D = \begin{pmatrix} P \circ L(2,1) & 0 & 0 \\ 0 & -1 & 0 \\ * & 0 & -4 \end{pmatrix} \in BL(U \times \mathbb{R}^2, V \times \mathbb{R}^2),$$

where $P \circ L(2,1) : U \to V$ is an isomorphism and $*$ is irrelevant to the calculation at hand. It follows that D is an isomorphism, and the result follows.

The second part of the theorem is proven in exactly the same way, simply observing the change of space, using X_{eo} rather than X_e. The uniqueness of the bifurcating branch in both X_e and X_{eo}, and the fact that $X_{eo} \subset X_e$, implies that $u(\beta) \in X_{eo}$ if F is even. □

Remark 1. In order to demonstrate that the application of Theorem 2.2 is not limited to fourth-order problems, we present the following example. In [33] the authors study the problem of finding periodic solutions for sixth-order problems using a variational approach, of which

$$u^{vi} + 5u^{iv} + pu'' + u - u^3 = 0 \tag{3.3}$$

is an example (see also [10]). Equation (3.3) has Hamiltonian

$$H \equiv \frac{1}{2}(u''')^2 + u^v u' - u^{iv}u'' + 5\left(u'u''' - \frac{1}{2}(u'')^2\right) + \frac{p}{2}(u')^2 + \frac{1}{2}u^2 - \frac{1}{4}u^4. \tag{3.4}$$

THEOREM 3.2. *The points* $(u,p) = (0, 4\frac{1}{4})$ *and* $(0, 1+4\sqrt{2})$ *are simple bifurcation points to even periodic solutions of* (3.3) *with zero Hamiltonian and with period near* $2\pi\sqrt{2}$ *and* $2\pi/\sqrt{1+\sqrt{2}}$, *respectively.*

The proof of Theorem 3.2 is very similar to that of Theorem 3.1, and so we omit it; note that the existence of a locally two-dimensional manifold of positive and negative Hamiltonian solutions also follows from Theorem 2.3.

3.3. Double bifurcations from $p > 2$. The following result shows that the interval $[2,\infty)$ contains a dense set of bifurcation points for (3.1)–(3.2).

THEOREM 3.3. *Suppose that assumption* (**F**) *holds. Then to each* $n, m \in \mathbb{N}$ *such that* $n \geq m+1$ *and* $\gcd(n,m) = 1$ *there is an interval* $I \subset \mathbb{R}$ *and exactly two analytic branches* $\beta \mapsto (u_\pm(\beta), p_\pm(\beta)) \in X_e \times \mathbb{R}$ *defined on* I *of even periodic solutions of* (1.1) *with zero Hamiltonian and period* $T_\pm(\beta)$. *Moreover,* $u_\pm(\beta) \neq 0$ *for* $\beta \neq 0$,

$$T_\pm(0) = 2\pi\sqrt{nm}, \quad u_\pm(0) = 0, \quad p_\pm(0) = \frac{n}{m} + \frac{m}{n},$$

and

$$\|u_\pm(\beta)(t) - \beta\left(m\cos(nt) \pm n\cos(mt)\right)\|_{C^4} = O(\beta^2)$$

as $\beta \to 0$.

Proof. Let us apply Theorem 2.4 to $\mathcal{H}(u,p,\mu) = 0$; to identify the functions k_1 and k_2 from Theorem 2.4, we consider the linearized problem

$$L(p,\mu)[a] \equiv \mu^4 a'''' + p\mu^2 a'' + a = 0$$

of (3.1) with $a \in X_e$. This linear equation admits an even 2π-periodic solution of the form $a(t) = \lambda_1 \cos(mt) + \lambda_2 \cos(nt)$ with integer m and n, provided that $\alpha_1 = (\mu m)^2$ and $\alpha_2 = (\mu n)^2$ both satisfy the equation $\alpha^2 - p\alpha + 1 = 0$. From this we obtain $\alpha_1\alpha_2 = 1$, so that $\mu = \mu_{n,m} \equiv \frac{1}{\sqrt{nm}}$ and $p = \alpha_1 + \alpha_2$, whence $p = p_{n,m} \equiv \frac{n}{m} + \frac{m}{n}$.

Hence we define $k_1(t) \equiv \cos(nt), k_2(t) \equiv \cos(mt)$, and $u_1 \equiv k_1, u_2 \equiv k_2$. Moreover, using $\psi(d^2_{uu} g(0, p_{n,m}, \mu_{n,m})[a, b]) = \delta(-\mu^4_{n,m} a''b'' + ab)$, we have

$$A = \frac{(n^2 - m^2)}{n^2}, \quad B = 0, \quad \text{and } C = -A,$$

so that $\alpha_\pm = \pm\frac{n}{m}$ in the notation of Theorem 2.4. With $Q_i(v) = \frac{1}{\pi}\int_0^{2\pi} v(t)u_i(t)dt$ for $i = 1, 2$, we find $d^2_{up}M(0, p_{n,m}, \mu_{n,m})[a, 1] = \mu^2_{n,m}a''$ and $d^2_{u\mu}M(0, p_{n,m}, \mu_{n,m})[a, 1] = 4\mu^3_{n,m}a'''' + 2p_{n,m}\mu_{n,m}a''$. We then evaluate the determinant from Theorem 2.4, which is $-4\mu^5_{n,m}\alpha_\pm n^2m^2(m^2 - n^2)$, and the result now follows since this is nonzero. ∎

If F is even, then all the bifurcations which occur for $p \geq 2$ are pitchforks because u is then a solution of (3.1)–(3.2) if and only if $-u$ is. The uniqueness properties of Theorems 2.2 and 2.4 and symmetry then imply that the parameterization of the solution branch satisfies $-u(\beta) = u(-\beta)$. From this we infer that the bifurcation diagram of $\|u(\beta)\|$ (with any suitable norm) plotted against $p(\beta)$ has a tongue-like appearance because of the density of the union of $p_{n,m}$ in $[2, \infty)$.

Remark 2. Theorem 3.3 was essentially known some time ago and can be found in an unpublished letter by J. F. Toland (1992), as referred to in [9, equation (5.1), p. 2486] for the case $F_u(u) = u - u^2$. This letter was communicated to the present authors by A. R. Champneys, and we express our gratitude for his help in this matter. A singularly perturbed version of (1.1) for this choice of nonlinearity was studied in [15] and more recently in [16], where the authors consider both homoclinic and periodic solutions, although the latter are not of zero Hamiltonian; see also [23].

As an aside, consider the equation

$$\frac{1}{12}v^{iv} + v'' + pv + v^3 + \frac{3}{4}v(vv'' + (v')^2), \tag{3.5}$$

taken from [21], with first integral

$$H \equiv \frac{1}{12}\left(v'''v' - \frac{1}{2}(v'')^2\right) + \frac{1}{2}(v')^2 + \frac{p}{2}v^2 + \frac{v^4}{4} + \frac{3}{8}(v')^2v^2. \tag{3.6}$$

Note that a parameter λ^2 appearing in [21] has been replaced here by p. This is not in the class of Hamiltonian systems given by (1.1), but Theorems 2.2 and 2.4 are still applicable.

THEOREM 3.4. *The point $p = 3$ is a simple bifurcation point to even periodic solutions of* (3.5) *with zero first integral and with period near $\pi\sqrt{2/3}$. For each $n, m \in \mathbb{N}$ such that $n > m$ and $\gcd(n, m) = 1$, the point $p_{n,m} = 12(\frac{n}{m} + \frac{m}{n})^{-2}$ is a double bifurcation point to such solutions with period near $\pi\sqrt{(n^2 + m^2)/3}$.*

The proof of Theorem 3.4 is an application of Theorem 2.4, which is entirely analogous to the proof of Theorem 3.3, so we omit the details.

3.4. Odd solutions for even *F*. Now let us suppose that F is an even function. If we define the spaces of odd functions, $X_o = \{u \in C^4_{2\pi} : u(t) = -u(-t)\}$ and $Y_o = \{u \in C^0_{2\pi} : u(t) = -u(-t)\}$, then $\mathcal{H}$ provides a map $\mathcal{H} \in C^\omega(X_o \times \mathbb{R}^2, Y_o \times \mathbb{R})$. This means that one can obtain odd zero-Hamiltonian solutions of (1.1) in a manner entirely

analogous to the way we found the even solutions. For this reason we give the following theorem without proof.

THEOREM 3.5. *If assumption* (**F**) *holds and F is even, then to each $n, m \in \mathbb{N}$ such that $n \geq m+1$ and $\gcd(n,m) = 1$ there is an interval $I \subset \mathbb{R}$ and exactly two analytic branches $\beta \mapsto (u_\pm(\beta), p_\pm(\beta)) \in X_e \times \mathbb{R}$ defined on I of odd periodic solutions of* (1.1) *with zero Hamiltonian and period $T_\pm(\beta) = 2\pi/\mu_\pm(\beta)$. Moreover, $u_p m(\beta) \neq 0$ for $\beta \neq 0$,*

$$T_\pm(0) = 2\pi\sqrt{nm}, \quad u_\pm(0) = 0, \quad p_\pm(0) = \frac{n}{m} + \frac{m}{n},$$

and

$$\|u_\pm(\beta)(t) - \beta\left(m\sin(nt) \pm n\sin(mt)\right)\|_{C^4} = O(\beta^2)$$

as $\beta \to 0$.

Of course, there is little point in formulating a version of Theorem 3.1 in this context, since that theorem already tells us that a branch of odd solutions of (1.1) can be found by shifting time.

It is possible to formulate an extension of the results proven in this section by considering a smooth one-parameter family of reversible vector fields on $\mathbb{R}^n$ which possesses a trivial branch of equilibrium solutions and a first integral. One could use Theorems 2.2 and 2.4 to formulate sufficient conditions for the bifurcation of zero-energy symmetric periodic solutions. However, for brevity we have not done this, and we restrict our attention to the properties of fourth-order systems.

3.5. Disjointness properties of solution sets. Motivated by Theorems 3.1 and 3.3, we define the following nonempty sets, assuming (**F**) to be true. Let

$$\Sigma \equiv \{(u,p,\mu) \in X_e \times \mathbb{R} \times \mathbb{R} : \mathcal{H}(u,p,\mu) = 0, u \neq 0, \mu > 0\}, \tag{3.7}$$

and let $\overline{\Sigma}$ denote the closure of Σ in $X_e \times \mathbb{R}^2$. For any pair $(n,m) \in \mathbb{N} \times \mathbb{N}$ such that $\gcd(n,m) = 1$, let $C(n,m)$ be the maximal connected subset of $\overline{\Sigma}$ which contains the point $(u,p,\mu) = (0, p_{n,m}, \mu_{n,m})$, and define the functional $\nu : \Sigma \to (2,\infty)$ by

$$\nu(u,p,\mu) = \|u\|_{C^4} + |p| + |\mu| + \frac{1}{|\mu|}. \tag{3.8}$$

Also, let

$$\Sigma_+ = \{(u,p,\mu) \in \Sigma : p > 0\}, \tag{3.9}$$

$\overline{\Sigma}_+$ be the closure of Σ_+, and $C_+(n,m)$ be the maximal connected subset of $C(n,m) \cap \overline{\Sigma}_+$ which contains the point $(u,p,\mu) = (0, p_{n,m}, \mu_{n,m})$.

We continue with a simple lemma which is used in the subsequent analysis. Throughout this section, # is used to represent the cardinality of a set, and we introduce a *potential* $V(u,u'')$ by writing (3.2) as

$$-\mu^2 u'\left(\mu^2 u''' + \frac{p}{2}u'\right) = V(u,u'') \equiv -\frac{1}{2}\mu^4 u''^2 + F(u).$$

LEMMA 3.1. *Suppose that $n, m \geq 1$ are distinct integers; then $\#\{t \in [0,\pi] : n\cos(nt) \pm m\cos(mt) = 0\} = \max(n,m)$ and $\#\{t \in [0,\pi] : m\cos(nt) \pm n\cos(mt) = 0\} = \min(n,m)$.*

The following two theorems provide bifurcation invariants that are invaluable to the study of the global nature of Σ.

THEOREM 3.6. *If assumption* (**F**) *holds and* $F(u) > 0$ *for* $u \neq 0$, *then the mapping*

$$\iota_1 : \Sigma \to \mathbb{N}; (u, p, \mu) \mapsto \#\{t \in [0, \pi] : u(t) = 0\}$$

is continuous and satisfies $\iota_1(C(n,m)\backslash\{(0, p_{n,m}, \mu_{n,m})\}) \equiv \min(n, m)$.

We postpone the proof of this theorem until after the following preliminary lemma.

LEMMA 3.2. *Let* $(u_k, p_k, \mu_k) \subset \Sigma$ *be a sequence with* $(u_k, p_k, \mu_k) \to (u, p, \mu) \in \Sigma$, *and suppose that there is a pair of sequences* $(t_k^1), (t_k^2) \subset [0, 2\pi]$ *such that* $|t_k^1 - t_k^2| \to 0$ *and* $u_k'(t_k^{1,2}) = 0$. *Then* $u_k(t_k^1)u_k(t_k^2) > 0$ *for sufficiently large* k.

Proof. For definiteness we assume that $t_k^1 < t_k^2$ and, seeking a contradiction, we also assume that $u_k(t_k^1)u_k(t_k^2) \leq 0$; we initially also assume that $u_k(t_k^1) < 0 < u_k(t_k^2)$. By the mean-value theorem there is a $T_k \in (t_k^1, t_k^2)$ such that $u_k''(T_k) = 0$, and taking the limit $k \to \infty$ gives the existence of a T such that $u'(T) = u''(T) = 0$. Using (3.2) yields $F(u(T)) = 0$ and therefore $u(T) = 0$, and since u is not identically zero, it follows that $u'''(T) \neq 0$.

We now assume without the loss of any generality that $\|u_k\|_{C^4} \leq \|u\|_{C^4} + 1 \equiv B$ and that $\frac{1}{2}\mu \leq \mu_k \leq \frac{3}{2}\mu$. By the smoothness of F and since $F(0) = F_u(0) = 0$, we may assume that there is a C such that

$$|F(w)| \leq \frac{C^2}{2}|w|^2 \quad \text{if } |w| \leq B.$$

Since $u_k'(t_k^{1,2}) = 0$, we have $V(u_k, u_k'') = -\frac{1}{2}\mu_k^4(u_k'')^2 + F(u_k) = 0$ at $t_k^{1,2}$, and therefore

$$|u_k''| \leq \frac{C}{\mu_k^2}|u_k| \leq \frac{4C}{\mu^2}|u_k| \quad \text{at } t_k^{1,2}. \tag{3.10}$$

If we define the function $v(t) = V(u_k(t), u_k''(t))$, then the mean-value theorem gives the existence of a $\tau_k \in (t_k^1, t_k^2)$ with $v'(\tau_k) = 0$. From (3.10), the orbit of u_k is a smooth curve in the u, u''-plane that connects $(u_k, u_k'')(t_k^1)$ (in the left half-plane) to $(u_k, u_k'')(t^2)$ (in the right half-plane) with end-points in the region $\{(u, u'') : |u''| \leq \frac{4C}{\mu^2}|u|\}$. By considering the tangent vector (u', u''') of this planar curve, it follows that there is a $\hat{\tau}_k \in (t_k^1, t_k^2)$ such that

$$|u_k'''(\hat{\tau}_k)| \leq \frac{4C}{\mu^2}|u_k'(\hat{\tau}_k)|. \tag{3.11}$$

When we take the limit $k \to \infty$, it follows from the estimate

$$\sup_{(t_k^1, t_k^2)} |u_k'| \leq |t_k^2 - t_k^1| \sup_{(t_k^1, t_k^2)} |u_k''|$$

that $u_k'''(\hat{\tau}_k) \to 0$. We therefore find that the limiting solution u satisfies $u'''(T) = 0$, which is a contradiction.

If the signs of u_k at $t_k^{1,2}$ are inverted, so that $u_k(t_k^2) < 0 < u_k(t_k^1)$, then the argument given above holds unchanged. If exactly one of the two values $u_k(t_k^{1,2})$ is zero for all k, then the argument holds in a similar way: in this case the curve in the u, u''-plane connects the origin to the other point. The existence of $\hat{\tau}_k$ satisfying (3.11)

follows as before. If both of the values of u_k are zero, then the curve is closed, and again a value of $\hat{\tau}_k$ can be found satisfying (3.11). This concludes the proof of the lemma. ∎

Proof of Theorem 3.6. Suppose that $(u,p,\mu) \in \Sigma$ and that $(u_k,p_k,\mu_k) \subset \Sigma$ satisfies $(u_k,p_k,\mu_k) \to (u,p,\mu)$ in $C^4 \times \mathbb{R}^2$.

We first note that if four or more zeros of u_k collide (counted according to algebraic multiplicity), say $0 \le t_k^1 \le t_k^2 \le t_k^3 \le t_k^4 \le 2\pi$ are all zeros of u_k that converge to T, then from the mean-value theorem we have $u(T) = u'(T) = u''(T) = u'''(T) = 0$, contradicting the assumption that $u \neq 0$.

On the other hand, if two or more zeros collide, then from the mean-value theorem again there is a T such that $u'(T) = 0$ and therefore $u''(T) = 0$ by (3.2). In order to avoid the same contradiction as above, necessarily $u'''(T) \neq 0$. This implies that the zero of u at $t = T$ is topologically transverse, which rules out the possibility that two transverse zeros coalesce.

We therefore are left with two cases: either three simple zeros collide or two zeros collide of which one is a double zero. In the first case, three simple zeros, there exist $t_k^{1,2}$ such that $u_k'(t_k^{1,2}) = 0$ and $t_k^{1,2} \to T$ as $k \to \infty$, and since the zeros are transverse, we can assume that u_k has opposite signs at t_k^1 and t_k^2. The conclusions of Lemma 3.2 show that this situation leads to a contradiction. In the second case we choose t_k^1 to be the nonsimple zero, and $\tau_k \in (t_k^1, t_k^2)$ to be an intermediate point such that $u'(\tau_k) = 0$. Again an application of Lemma 3.2 leads to a contradiction, and therefore the number of zeros of u_k eventually equals that of u. This shows that ι_1 is continuous on Σ and therefore constant on connected components of Σ.

In order to evaluate $\iota_1(u,p,\mu)$ for $(u,p,\mu) \in C(n,m)$ with $u \neq 0$ we use the representation of $C(n,m)$ at bifurcation given in Theorem 3.3. From Lemma 3.1 there results

$$\|u_\pm(\beta)(t) - \beta\,(m\cos(nt) \pm n\cos(mt))\,\|_{C^4} = O(\beta^2),$$

and hence $\#\{t \in [0,\pi] : u_\pm(\beta)(t) = 0\} = \min(n,m)$ follows for sufficiently small and nonzero β, and the theorem is proven. ∎

THEOREM 3.7. *If assumption* (**F**) *holds and* $F(u) > 0$ *for* $u \neq 0$*, then the mapping*

$$\iota_2 : \Sigma_+ \to \mathbb{N}; (u,p,\mu) \mapsto \#\{t \in [0,\pi] : u''(t) = 0\}$$

is continuous.

Proof. Let $(u,p,\mu) \in \Sigma_+$, and suppose that there is a $T \in [0,2\pi]$ such that

$$u''(T) = u'''(T) = 0.$$

The zero-Hamiltonian condition (3.2) then gives

$$\frac{1}{2}p\mu^2 u'(T)^2 + F(u(T)) = 0,$$

and the hypotheses on F ensure that $u'(T) = 0$ and $u(T) = 0$. It follows that $u = 0$, which contradicts the definition of Σ_+, and this contradiction implies that the zeros of u'' are transverse. Consequently, if $(u_n,p_n,\mu_n) \subset \Sigma_+$ is a sequence such that $(u_n,p_n,\mu_n) \to (u,p,\mu)$ in Σ_+, then $u_n'' \to u''$ in the C^1 topology. Hence u_n'' has the same number of zeros as u'' for all sufficiently large n, which shows that ι_2 is continuous as claimed. ∎

Theorem 3.7 immediately implies that ι_2 is constant on connected components of Σ_+, and from this observation we deduce the following.

COROLLARY 3.1. *Suppose that* (**F**) *holds and* $uF_u(u) \geq 0$ *for all* $u \in \mathbb{R}$*; then* $\Sigma_+ = \Sigma$*, and as a consequence,* $C_+(n,m) = C(n,m)$*. Moreover,*

$$\iota_2(C(n,m)\backslash\{(0,p_{n,m},\mu_{n,m})\}) = \max(n,m),$$

so that $C(n,m) \cap C(n',m')$ *is empty unless* $(n,m) = (n',m')$*.*

Proof. Multiplying (3.1) by u and integrating gives

$$p\mu^2 \int_0^{2\pi} (u')^2 dt = \int_0^{2\pi} \mu^4 (u'')^2 + uF_u(u) dt \geq \int_0^{2\pi} \mu^4 (u'')^2 dt \geq 0.$$

Hence if there is a solution of (3.1)–(3.2) with $p = 0$ and $\mu > 0$, it follows that $u'' \equiv 0$, so $u(t) = At + B$ for constants A and B. As u is periodic, $A = 0$, and as u must have zero Hamiltonian, $F(B) = 0$ is also true. The hypotheses ensure that $F(u) = 0$ only when $u = 0$ so that $B = 0$; hence $u(t) \equiv 0$ and so $\Sigma = \Sigma_+$, from which $C(n,m) = C_+(n,m)$ by definition.

Since $\iota_2 : \Sigma \to \mathbb{N}$ is continuous by Theorem 3.7, the set $\mathcal{C}$ defined by $\mathcal{C} \equiv C(n,m)\backslash\{(0,p_{n,m},\mu_{n,m})\}$ is a connected subset of Σ, because the intersection of $C(n,m)$ with some small ball, $C(n,m) \cap B_\delta(0,p_{n,m},\mu_{n,m})$, is path-connected for all sufficiently small $\delta > 0$. Hence ι_2 is constant on $\mathcal{C}$. In order to evaluate $\iota_2(u,p,\mu)$ with $(u,p,\mu) \in \mathcal{C}$ we use the representation of $C_+(n,m)$ at bifurcation from the trivial solution described in Theorem 3.3 and then apply Lemma 3.1. From Theorem 3.3 we have

$$\| u''_\pm(\beta)(t) + \beta mn \left(n\cos(nt) \pm m\cos(mt)\right) \|_{C^2} = O(\beta^2),$$

whence $\#\{t \in [0,\pi] : u''_\pm(\beta)(t) = 0\} = \max(n,m)$ follows for sufficiently small and nonzero β.

Finally, as (**F**) ensures that F is positive in a punctured neighborhood of zero, the condition $uF_u(u) \geq 0$ then ensures that $F(u) > 0$ if $u \neq 0$. Applying Theorem 3.6, we obtain $\iota_j(C(n,m)) = \iota_j(C(n',m'))$ for $j = 1,2$, and the last part of the corollary follows. ∎

Finally, we have the following theorem, which applies to *path-connected subsets* of Σ, although it provides no information regarding the behavior of *connected subsets* of Σ which are not path-connected.

THEOREM 3.8. *Suppose that* $F(u) > 0$ *for* $u \neq 0$*, and let* (u_s, p_s) *be a continuous path of solutions of* (1.1)*, where each* u_s *is defined on a sufficiently large subset of* $\mathbb{R}$ *and the path does not contain the equilibrium solution. If two zeros of* u'_s *collide, then multiplicity is preserved in and through the collision.*

Proof. We may assume that all collisions occur at $s = 0$ and $t = 0$. Note that if all u_s are defined on a common interval $I \subset \mathbb{R}$, then by standard elliptic estimates the solution curve is bounded in $C^k(I')$ for any $k \in \mathbb{N}$ and any compact interval $I' \subset I$. Since F is smooth, we can bound F by

$$|F(w)| \leq \frac{C^2 w^2}{2} \tag{3.12}$$

for, say, $|w| \leq 1$.

Note that $u'(0) = u''(0) = 0$ at $s = 0$, and therefore by (1.2), $F(u(0)) = 0$, which implies $u(0) = 0$. The nonconstancy hypothesis on each u_s implies that $u'''(0) \neq 0$.

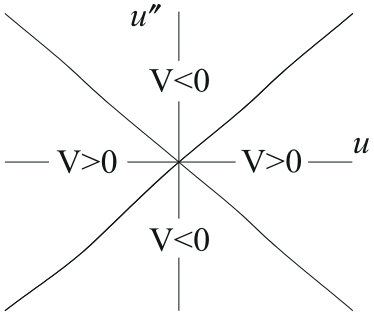

FIG. 3.1. *The configuration space u, u'' is partitioned according to the sign of V.*

This proves that multiplicity is conserved *in* the collision. Note that the zero at $(t, s) = (0, 0)$ is necessarily transverse.

To show that any subsequent perturbation preserves the multiplicity, we use some ideas from the analysis of the *configuration space* (u, u'') from [34, 17]; [29] contains a simplified description that is sufficient for our purposes. The structure of the configuration space and the set $\{V = 0\}$ is shown in Figure 3.1. Near the origin in this plane the set $\{V = 0\}$ consists of two curves that intersect in the origin. Near the origin the direction of these curves is bounded from above by $2C$, where C is the constant in (3.12).

At $s = 0$, we have $u(0) = u'(0) = u''(0)$, $u'''(0) \neq 0$, and therefore the orbit near $t = 0$ is represented in the u, u''-plane by a curve that remains inside the set $\{V \leq 0\}$ and intersects $\{V = 0\}$. We can choose appropriate translations of u_s, and small $\bar{t}, \bar{s} > 0$, such that we have the following:

1. $u_s(t)$ is defined for $(t, s) \in Q \equiv (-\bar{t}, \bar{t}) \times (-\bar{s}, \bar{s})$.
2. u_s depends smoothly on s in $C^4(-\bar{t}, \bar{t})$.
3. $u_s'''(t) \geq 0$ on Q (if not, then reverse time).
4. For each $s \in (-\bar{s}, \bar{s})$ we have $\pm u_s''(\pm\bar{t}) < 0$, and $V(u_s(\pm\bar{t}), u_s''(\pm\bar{t})) < 0$. This implies that no intersections of the solutions with $\{V = 0\}$ appear or disappear through the boundary $\pm\bar{t}$.
5. $u_s'''/u_s' \geq 4C$ on Q, where C is the constant in (3.12).

We write $\gamma_s \equiv \{(u_s(t), u_s''(t)) : -\bar{t} < t < \bar{t}\}$.

We now consider the alternatives for perturbation away from $s = 0$. First assume that $\bar{s}$ can be chosen such that $\gamma_s \cap \{V = 0\}$ has only one intersection for $0 < s < \bar{s}$; let this intersection be at $0 < t_s < \bar{s}$. The lower bound on the angle of the curve γ_s, given by condition 5 above, implies that γ_s intersects $\{V = 0\}$ only at the origin in the u, u''-plane. Since $u_0'(0) = 0$ and $u_0'''(0) \neq 0$, the smooth dependence of u_s on s implies that $u_s'''(t_s) \neq 0$ for s close to zero; therefore the requirement $u'(t_s)(u_s'''(t_s) + pu_s'(t_s)/2) = 0$ forces $u_s'(t_s) = 0$. Combining this with $u_s(t_s) = 0$ and taking the limit yields that the zero of u' at $s = 0$ is a double zero.

To cover the alternative case we assume that $\gamma_s \cap \{V = 0\}$ has at least two intersections for a sequence $0 < s_n < \bar{s}$, $s_n \downarrow 0$, at the points $0 < t_n < \tau_n < \bar{t}$. We have $\lim_{n\to\infty} t_n = \lim_{n\to\infty} \tau_n = 0$. With an argument similar to the one above, it follows that $u_{s_n}'(t_n) = u_{s_n}'(\tau_n) = 0$, and therefore the zero at $s = 0$ is of second order.

At any point where the orbit intersects $\{V = 0\}$, either $u' = 0$ or $u''' = -pu'/2$. Since at $s = 0$ we have $u' = 0$ and $u''' \neq 0$, under perturbation of s we have $u' = 0$ on the intersection of the orbit with $\{V = 0\}$. If we assume that under perturbation

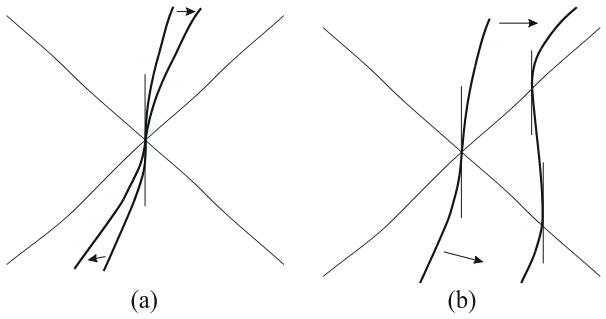

FIG. 3.2. *Two forms of perturbation:* (a) *if the curve continues to intersect* $\{V = 0\}$ *in the origin, then the tangent remains vertical;* (b) *a translation, on the other hand, creates new zeros of* u' *and therefore also conserves multiplicity.*

there is only one zero of u' (locally), then the intersection with $\{V = 0\}$ necessarily occurs at the origin in the u, u''-plane. Therefore the zero remains of multiplicity two.

Note that the only possible scenario is the reduction of multiplicity two to multiplicity zero. Multiplicity zero implies that while $u'(0) = 0$ at $s = 0$, this zero of u' disappears under perturbation. An inspection of the u, u''-plane (Figure 3.2) shows that such a perturbation is possible only if $u''' + pu'/2$, which is nonzero at $s = 0$, jumps to zero for $s \neq 0$. This contradicts the assumption of continuous dependence of the curve of solutions in C^k on the parameter s. □

4. Global bifurcations. Let us now briefly state some results from global real-analytic bifurcation theory for one-parameter problems as developed in [6]. The utility of this theory with respect to (2.1) is the fact that Lemma 2.1 identifies either p or μ which can be used as the bifurcation parameter. So, supposing that $\mathcal{U} \subset \mathbb{R} \times X$ is a given set and that $F : \mathbb{R} \times X \to Y$ is a real analytic map, define the set

$$S = \{(\lambda, x) \in \mathcal{U} : F(\lambda, x) = 0, d_x F(\lambda, x) \in \mathrm{Iso}(X, Y)\}.$$

Throughout this section the space $\mathbb{R} \times X$ is endowed with the norm $|\lambda| + \|x\|_X$, and a pair $(\lambda, x) \in S$ is said to be a regular solution of $F(\lambda, x) = 0$. While the theory developed in [6] is more powerful than we require, we shall use the following result, which is the statement of Theorem 7.4(iii) of this reference.

THEOREM 4.1. *Let* $\nu : \mathbb{R} \times X \to [0, \infty)$ *be a given function, and suppose that*

(i) S *is nonempty and* $\mathcal{U} \cap \mathcal{S}$ *is open in* $\mathcal{S}$, *where* $\mathcal{S} \equiv \{(\lambda, x) \in \mathbb{R} \times X : F(\lambda, x) = 0\}$,
(ii) $d_x F(\lambda, x)$ *is Fredholm of index zero for all* $(\lambda, x) \in \mathcal{U}$,
(iii) *subsets of* $\mathcal{S}$ *on which* ν *is bounded have compact closure,*
(iv) *there are* $\delta > 0, \lambda_0 \in \mathbb{R}$, *and an analytic function* $h : N_\delta(\lambda_0) \backslash \{\lambda_0\} \to S$, *where* $N_\delta(\lambda_0)$ *is a half-neighborhood of* λ_0, *such that* $\lim_{\lambda \to \lambda_0} h(\lambda) = 0$ *but* $(\lambda_0, 0) \notin \mathcal{U}$,
(v) *if* $\mathcal{A}_0$ *is the maximal path-connected subset of* S *which contains the graph of* h *and if* $(\xi_n) \subset \overline{S} \cap \mathcal{U}$ *is any convergent sequence with* $\xi_n \to \xi \notin \mathcal{U}$ *and* $\sup_n \nu(\xi_n) < \infty$, *then* $\xi = (\lambda_0, 0)$ *and* $\xi_n \in \mathcal{A}_0$ *for all* n *sufficiently large.*

Under conditions (i)–(v) *the maximal connected component of* $\overline{S} \cap \mathcal{U}$ *that contains* $\mathcal{A}_0$ *contains a path-connected subset* $\mathcal{P}$ *on which* ν *is unbounded.*

If we recall the definition of the sets Σ and Σ_+ in (3.7) and (3.9), respectively, then the reasoning in Corollary 3.1 ensures that $\Sigma_+ = \Sigma$ if $uF_u(u) \geq 0$ for all $u \in \mathbb{R}$.

From this observation we can obtain the following lemma.

LEMMA 4.1. *Suppose that* (**F**) *holds and* $uF_u(u) \geq 0$ *for* $u \in \mathbb{R}$*; then* $\nu(u,p,\mu)$ *(defined in* (3.8)*) is unbounded on a path-connected subset of* $C(1,1)$.

Proof. Let us verify the hypotheses of Theorem 4.1 in turn, where $\omega = \mu - 1, x = (u,\omega)$ and we write λ in place of p. To keep the notation consistent with Theorem 4.1, let $F \equiv M \times \psi(g)$ (the symbol $\mathcal{H}$ was used for this previously) and

$$\mathcal{U} \equiv \{(\lambda, x) = (p,u,\omega) \in \mathbb{R} \times X_e \times \mathbb{R} : \omega > -1, u \neq 0, \iota_1(u) = \iota_2(u) = 1\}.$$

We remark that $S \subset \Sigma \subset \mathcal{S},\ \mathcal{S} \cap \mathcal{U} \subset \Sigma$ and the space X referred to in Theorem 4.1 is $X_e \times \mathbb{R}$ and Y is $Y_e \times \mathbb{R}$.

(i) It follows from Lemma 2.1 that S is nonempty. (The argument assumes that $d_{u,p}(M \times \psi(g))$ is an isomorphism from Lemma 2.1, for if this is not the case, then we can repeat the argument of this proof using μ for λ rather than p.) One can see that the set $\mathcal{U} \cap \mathcal{S}$ is open in $\mathcal{S}$ as follows. Suppose, seeking a contradiction, that $\mathcal{U} \cap \mathcal{S}$ is not open in $\mathcal{S}$, so there is a $(\lambda_0, x_0) = (p_0, u_0, \omega_0) \in \mathcal{S} \cap \mathcal{U}$ and a sequence $(\lambda_n, x_n) = (p_n, u_n, \omega_n) \in \mathcal{S} \backslash \mathcal{U}$ such that $(\lambda_n, x_n) = (p_n, u_n, \omega_n) \to (\lambda_0, x_0) = (p_0, u_0, \omega_0)$.

Since $u_0 \neq 0$ and $u_n \overset{C^4}{\to} u_0$, it follows that $u_n \neq 0$, and as $\omega_0 > -1$, then $\omega_n > -1$, both for all sufficiently large n. Since ι_1 and ι_2 are continuous functions on Σ and Σ_+, respectively, and $\Sigma = \Sigma_+$ by the hypothesis on F, then $\iota_1(u_n) \to \iota_1(u_0) = 1$ as $n \to \infty$; but since ι_1 is integer-valued, this means that $\iota_1(u_n) \equiv 1$ for all n sufficiently large. Similar reasoning applies to ι_2. Consequently, $(\lambda_n, x_n) \in \mathcal{U}$ for all n sufficiently large, which is the required contradiction.

(ii) The operator $d_x F(\lambda, x)$ has the form

$$\begin{pmatrix} d_u M(u,p,\mu) & 0 \\ 0 & 0 \end{pmatrix} + K \in BL(X_e \times \mathbb{R}, Y_e \times \mathbb{R}),$$

where K is a continuous operator that has rank at most two and $d_u M(u,p,\mu)[h] = \mu^4 h'''' + p\mu^2 h'' + F_u(u)h \in BL(X_e, Y_e)$. However, the latter is a compact perturbation of the operator

$$E : h \mapsto \mu^4 h^{iv} + \theta h, \qquad E \in BL(X_e, Y_e).$$

Since $\mu > 0$, using a Fourier series argument, one can easily show that there is a θ such that E is an isomorphism of the given spaces. Consequently $d_x F(\lambda, x)$ is a compact perturbation of a Fredholm mapping of index zero, and therefore is itself Fredholm of index zero.

(iii) If $(u_n, p_n, \omega_n) \subset \mathcal{S}$ is a sequence such that $\omega_n = \mu_n - 1 > -1$ and

$$\nu(u_n, p_n, \mu_n) = \|u_n\|_{C^4} + |p_n| + |\omega_n + 1| + \frac{1}{|\omega_n + 1|}$$

is bounded, then there are p_0 and ω_0 such that $\omega_n \to \omega_0 \geq -1$ and $p_n \to p_0 \geq 0$. Now $\omega_0 \neq -1$ by the boundedness of ν, and therefore $u_n^v = -(\omega_n + 1)^{-4}(p_n(\omega_n + 1)^{-2} u_n''' + F_u(u_n)u_n')$ is also bounded, whence (u_n) converges to some $u_0 \in C^4$ as the embedding $C^5 \hookrightarrow C^4$ is compact, and therefore (p_n, u_n, ω_n) converges in $\mathbb{R} \times X_e \times \mathbb{R}$.

(iv) This part of the theorem follows from Theorem 2.2 and Lemma 2.1, where the bifurcating branch is represented by an analytic curve of regular solutions.

(v) If a ν-bounded sequence $(\xi_n) = (p_n, u_n, \omega_n) \subset \overline{S} \cap \mathcal{U}$ satisfies $\xi_n \to \xi = (p_0, u_0, \omega_0) \notin \mathcal{U}$, then the only viable possibility is that $u_0 = 0$, so that $(p, \mu) =$

$(p_0, \omega_0 + 1)$ is a bifurcation point from the trivial solution of (3.1)–(3.2). However, the *only* point at which such a bifurcation occurs in the set S is at the point $(u, p, \mu) = (0, 2, 1)$, so that $(u, \omega) = (0, 0)$. Hence property (v) is satisfied if $\mathcal{A}_0$ is defined to be the maximal path-connected subset of S which contains the graph of the bifurcating branch from Theorem 3.1. In this case let us note that $\lambda_0 = 2$.

This concludes the proof. □

Now define the functional on $X_e \times \mathbb{R}^2$ by

$$\overline{\nu}(u, p, \mu) = \|u\|_{C^4} + |p| + \frac{1}{|\mu|}.$$

In order to obtain a result analogous to the global Hopf bifurcation theorem of [1], we show that ν can actually become unbounded on $C(1,1)$ if and only if $\overline{\nu}$ is unbounded on $C(1,1)$.

THEOREM 4.2. *If* (**F**) *is satisfied and* $uF_u(u) \geq 0$, *then* $C(1,1)$ *contains a path-connected* $\overline{\nu}$*-unbounded subset.*

Proof. Suppose that ν is unbounded on $C(1,1)$ but that $\overline{\nu}$ is bounded on this set; it follows that there is a sequence $(u_n, p_n, \mu_n) \in C(1,1)$ such that $\|u_n\|_{C^4} + |p_n|$ is bounded, $u_n \neq 0$ for each n, and $|\mu_n| \to \infty$. However, since

$$\begin{aligned}\int_0^{2\pi} p\mu^2 (u'')^2 dt \geq \frac{1}{2\pi} \int_0^{2\pi} p\mu^2 (u')^2 &= \frac{1}{2\pi} \left(\int_0^{2\pi} \mu^4 (u'')^2 + uF_u(u) dt \right) \\ &\geq \frac{1}{2\pi} \int_0^{2\pi} \mu^4 (u'')^2 dt\end{aligned}$$

holds for any nontrivial solution in $C(1,1)$ by the Poincaré inequality, it follows that $\mu_n^2 \leq 2\pi p_n$, which is a contradiction. Therefore, by Lemma 4.1, ν is unbounded on a path-connected subset of $C(1,1)$, and the above contradiction implies that $\overline{\nu}$ must also be unbounded on this set. □

This theorem represents a partial global trichotomy for bifurcations of periodic orbits of (1.1), which says that the solution continuum $C(1,1)$ either has an unbounded sequence of orbits in phase-space or is unbounded with respect to either the parameter (p) or with the period (as occurs in the *blue-sky bifurcation* [13]). Unfortunately, due to assumption (v) of Theorem 4.1, it has not proven possible to use the same techniques to study the global existence properties of the branches $C(n,m)$ for $n > 1$.

5. Local secondary fold bifurcations. Another advantage of the approach taken in this paper as opposed to the shooting methods previously used in [35, 36] is that we can investigate the geometry of each bifurcating continuum by introducing an unfolding parameter, ϵ, into (3.1)–(3.2). We will now show that degeneracies present in (3.1)–(3.2) at $\epsilon = 0$ can unfold to give secondary fold bifurcations along the bifurcation branch when $\epsilon \neq 0$.

To illustrate this we shall consider (1.1) for the particular case given in (1.3). This has been studied in [19] (see also [12] for an asymptotic analysis of this problem using multiple scale techniques) as a model for an elastic rock layer on a *restiffening foundation*, with corresponding $\mathbb{Z}_2$-symmetric ODE

$$M_\epsilon(u, p, \mu) \equiv \mu^4 u'''' + p\mu^2 u'' + u - \epsilon(u^3 - u^5) \tag{5.1}$$

and even Hamiltonian

$$H_\epsilon(u, p, \mu) \equiv \mu^4 u' u''' - \frac{1}{2}\mu^4 u''^2 + \frac{1}{2} p\mu^2 u'^2 + \frac{1}{2}u^2 - \epsilon\left(\frac{1}{4}u^4 - \frac{1}{6}u^6\right). \tag{5.2}$$

As in the proof of Theorem 2.2, and therefore also in Theorem 3.1, we can obtain a local representation of the bifurcating branch of the zero-Hamiltonian problem associated with (5.1) from the bifurcation point $p = 2$ in the form $u_\epsilon(\beta) = \beta(k + \rho_\epsilon(\beta))$, where $\rho_\epsilon(\beta) = O(\beta)$ for fixed ϵ and is an analytic function of both β and ϵ near $(\beta, \epsilon) = (0, 0)$; here $k(t) = \cos(t)$. We now proceed with a calculation to find the Taylor expansion of $\rho_\epsilon(\beta)$ in order to determine the local geometry of the set of branches $C_\epsilon(1, 1)$. Throughout this section we shall write

$$r_\epsilon(\beta) = \beta\rho_\epsilon(\beta).$$

It is important to note that the existence of bifurcating solutions for this problem close to $p = 2$, as determined by Theorem 3.1, does not depend upon the value of ϵ. Indeed, using the implicit function theorem we simply find that for each ϵ sufficiently small and for suitable m, n, there is a bifurcating branch from $(p, \mu) = (p_{n,m}, \mu_{n,m})$ and this branch (that is, the local parametric representation of this branch) varies analytically with ϵ. This property also holds at $\epsilon = 0$, where the branches are pairs of lines. We also note that since $F_\epsilon(u)$ is an even function of u for each ϵ, $r_\epsilon(\cdot)$ is odd and $p_\epsilon(\cdot)$ and $\mu_\epsilon(\cdot)$ are even functions, forming a pitchfork bifurcation at $p = 2$.

We start our analysis by listing the Fréchet derivatives of the operator M_ϵ:

D1. $d_u M_\epsilon(u, p, \mu)[h] = \mu^4 h'''' + p\mu^2 h'' + h - \epsilon(3u^2 - 5u^4)h$,
D2. $d_u^2 M_\epsilon(u, p, \mu)[h_1, h_2] = -\epsilon h_1 h_2(6u - 20u^3)$,
D3. $d_u^3 M_\epsilon(u, p, \mu)[h_1, h_2, h_3] = -\epsilon h_1 h_2 h_3(6 - 60u^2)$,
D4. $d_u^4 M_\epsilon(u, p, \mu)[h_1, h_2, h_3, h_4] = 120\epsilon v h_1 h_2 h_3 h_4$,
D5. $d_u^5 M_\epsilon(u, p, \mu)[h_1, h_2, h_3, h_4, h_5] = 120\epsilon h_1 h_2 h_3 h_4 h_5$,

where $h_i \in X_e$ for each $i = 1, \ldots, 5$.

We denote the first derivative of $M_\epsilon(u, p, \mu)$ evaluated on the trivial solution branch $u = 0$ by $L \equiv d_u M_\epsilon(0, 2, 1)$; this operator is independent of ϵ. Suppose further that P is the projection of Y_e onto $\mathrm{ran}(L) = \langle k \rangle^\perp$ along $\langle k \rangle$; now define

$$L(p, \mu) \equiv d_u M_\epsilon(0, p, \mu).$$

We can solve the projected differential equation $P \circ N_\epsilon(\beta k + r, p, \mu) = 0$ for some function $r = r_\epsilon(\beta, p, \mu)$ near $(\beta, p, \mu, \epsilon) = (0, 2, 1; 0)$ using the implicit function theorem (we refer to the proof of Theorem 2.2 for details). From the uniqueness properties of the implicit function theorem if follows that $r_\epsilon(0, p, \mu) \equiv 0$, and if we repeatedly differentiate the identity $PM_\epsilon(\beta k + r_\epsilon(\beta, p, \mu), p, \mu) = 0$ with respect to β, then we shall obtain the Taylor coefficients of r_ϵ. This is a tedious exercise, so we omit the details, but one eventually obtains

R1. $P(d_u M_\epsilon[k + d_\beta r]) \equiv 0$,
R2. $P(d_u^2 M_\epsilon[k + d_\beta r, k + d_\beta r] + d_u M_\epsilon[d_\beta^2 r]) \equiv 0$,
R3. $P(d_u^3 M_\epsilon[k+d_\beta r]^3 + 2d_u^2 M_\epsilon[d_\beta^2 r, k+d_\beta r] + d_u^2 M_\epsilon[d_\beta^2 r, k+d_\beta r] + d_u M_\epsilon[d_\beta^3 r]) \equiv 0$,
R4. $P(d_u^4 M_\epsilon[k + d_\beta r]^4 + 6d_u^3 M_\epsilon[d_\beta^2 r, k + d_\beta r, k + d_\beta r] + 3d_u^2 M_\epsilon[d_\beta^3 r, k + d_\beta r] + 3d_u^2 M_\epsilon[d_\beta^2 r, d_\beta^2 r] + d_u M_\epsilon[d_\beta^4 r] + d_u^2 M_\epsilon[k + d_\beta r, d_\beta^3 r]) \equiv 0$,
R5. $P(d_u^5 M_\epsilon[k + d_\beta r]^5 + 10d_u^4 M_\epsilon[d_\beta^2 r, [k + d_\beta^3 r]^3] + 10d_u^3 M_\epsilon[d_\beta^3 r, k + d_\beta r, k + d_\beta r] + 3d_u^2 M_\epsilon[d_\beta^4 r, d_\beta r] + 10d_u^2 M_\epsilon[d_\beta^3 r, d_\beta^2 r] + 2d_u^2 M_\epsilon[d_\beta^4 r, d_\beta r] + d_u M_\epsilon[d_\beta^5 r]) \equiv 0$.

Evaluating these expressions at $\beta = 0$, that is, $u = r_\epsilon(0, p, \mu) = 0$, yields the following information. From R1 we have $PL(p, \mu)[k + d_\beta r(0, p, \mu; \epsilon)] = 0$, and because $L(p, \mu)k = (\mu^4 - p\mu^2 + 1)k \in \langle k \rangle$ we have $d_\beta r_\epsilon(0, p, \mu) \equiv 0$. The expression $d_\beta^2 r_\epsilon(0, p, \mu) = 0$ then follows from R2. Also, R3 gives

$$PL(p, \mu)d_\beta^3 r_\epsilon(0, p, \mu) = 6\epsilon P(k^3), \tag{5.3}$$

so that the third derivative of r_ϵ is not zero in general at $(u, p, \mu; \epsilon) = (0, 2, 1; \epsilon)$, but $d^3_\beta r_\epsilon(0, p, \mu)$ is seen to provide an $O(\epsilon)$ contribution to the Taylor expansion of r_ϵ. Using Taylor's theorem to expand $r_\epsilon(\beta, p, \mu)$ with respect to β and using the symmetry properties of r_ϵ (it is odd with respect to β), we may write $r_\epsilon(\beta, p, \mu) = \frac{\beta^3}{6} R^1_\epsilon(p, \mu) + O(\beta^5)$ for some operator $R^1_\epsilon(p, \mu)$ with range in $\langle k \rangle^\perp$.

We can determine R^1_ϵ as follows. In (5.3) seek an even Fourier series solution which is also orthogonal to k in X_e of the form $R^1_\epsilon(p, \mu) = \sum_{j=2}^\infty a_j \cos(jt)$, where the coefficients a_j remain to be determined. Since $k^3(t) = \frac{1}{4}(\cos(3t) + 3\cos(t))$, it follows that the only nonzero coefficient is a_3 and

$$R^1_\epsilon(p, \mu) = \frac{3\epsilon}{2} \frac{\cos(3\cdot)}{81\mu^4 - 9p\mu^2 + 1}.$$

Using R4 and setting $\beta = 0$, we find $d^4_\beta r_\epsilon(0, p, \mu) \equiv 0$, which of course also follows from symmetry. We may evaluate $d^5_\beta r_\epsilon(0, p, \mu)$ from R5, which simplifies to give

$$P(d^5_u M_\epsilon[k]^5 + 10 d^3_u M_\epsilon[k, k, d^3_\beta r_\epsilon] + d_u M_\epsilon[d^5_\beta r_\epsilon]) \equiv 0.$$

To find $d^5_\beta r_\epsilon(0, p, \mu)$ we solve the following linear equation for $w \in X_e \cap \langle k \rangle^\perp$,

$$\mu^4 w'''' + p\mu^2 w'' + w + P\left[60k^2 \frac{3\epsilon^2}{2} \frac{\cos(3\cdot)}{81\mu^4 - 9p\mu^2 + 1} + 120\epsilon k^5\right] = 0, \tag{5.4}$$

and then $d^5_\beta r_e(0, p, \mu) = w$. Since $k(t)^5 = \frac{1}{16}(\cos(5t) + 5\cos(3t) + 10\cos(t))$, $k(t)^2 = \frac{1}{2}(1 + \cos(2t))$, and $\cos(3t)\cos(2t) = \frac{1}{2}(\cos(5t) + \cos(t))$, we also solve (5.4) using a Fourier series expansion. Accordingly, taking $w(t) = \sum_{j=2}^\infty w_j \cos(jt)$, we find that all the coefficients w_j are zero, except when $j = 3$ or $j = 5$. In these cases

$$(81\mu^4 - 9p\mu^2 + 1)w_3 + 150\epsilon - \frac{45\epsilon^2}{81\mu^4 - 9p\mu^2 + 1} = 0$$

and

$$(625\mu^4 - 25p\mu^2 + 1)w_5 + 30\epsilon - \frac{90\epsilon^2}{2(81\mu^4 - 9p\mu^2 + 1)} = 0.$$

It follows that $r_\epsilon(\beta, p, \mu) = \frac{\beta^3}{6} R^1_\epsilon(p, \mu) + \frac{\beta^5}{120} R^2_\epsilon(p, \mu) + \frac{\beta^7}{720} d^7_\beta r_\epsilon(0, p, \mu) + O(\epsilon^2\beta^3)$, where

$$R^2_\epsilon(p, \mu) = -\epsilon\left(\frac{150}{4(81\mu^4 - 9p\mu^2 + 1)} \cos(3\cdot) + \frac{30}{4(625\mu^4 - 25p\mu^2 + 1)} \cos(5\cdot)\right) + O(\epsilon^2).$$

One can show by further differentiating that $d^6_\beta r_\epsilon(0, p, \mu) \equiv 0$, as we expect from symmetry, and the equation which determines $d^7_\beta r_\epsilon(0, p, \mu)$ shows this term to be of order $O(\epsilon^2)$. Higher derivatives of r_ϵ will also be of order $O(\epsilon^2)$.

Now $p_\epsilon(\beta)$ and $\mu_\epsilon(\beta)$ are even functions of β, and applying the zero-Hamiltonian constraint gives

$$\mu^4 \delta((k + \rho_\epsilon)''^2) = \delta\left((k + \rho_\epsilon)^2 - \frac{1}{2}\epsilon\beta^2(k + \rho_\epsilon)^4 + \frac{1}{3}\epsilon\beta^4(k + \rho_\epsilon)^6\right), \tag{5.5}$$

using a prime to denote $\frac{d}{dt}$. Seeking an expansion of the bifurcation branch about $(\epsilon, \beta) = (0, 0)$, we write

$$p_\epsilon(\beta) = 2 + \epsilon\left(P_1\beta^2 + P_2\beta^4\right) + O(\epsilon^2) \tag{5.6}$$

and

$$\mu_\epsilon(\beta) = 1 + \epsilon\left(\omega_1\beta^2 + \omega_2\beta^4\right) + O(\epsilon^2), \tag{5.7}$$

where P_1, P_2, ω_1, and ω_2 are real numbers to be determined. The highest power of β which exists in these expansions at $O(\epsilon)$ is the quartic because of the Taylor series we have found for $r_\epsilon(\beta)$. This is clear from (5.5), which contains terms of order $\epsilon\beta^2$ and $\epsilon\beta^4$ but not $\epsilon\beta^6$ or higher.

To determine ω_1 and ω_2 we substitute the expressions for $\rho_\epsilon = \frac{r_\epsilon}{\beta}$ and μ_ϵ into (5.5). Using $\delta(\rho) = \left(\frac{1}{256}\beta^2 - \frac{23}{4608}\beta^4\right)\epsilon + O(\epsilon^2)$ and $\delta(\rho'') = \left(-\frac{9}{256}\beta^2 + \frac{215}{4608}\beta^4\right)\epsilon + O(\epsilon^2)$, we then find $\frac{1}{(81\mu^4 - 9p\mu^2 + 1)} = \frac{1}{64} + O(\epsilon)$ and $\frac{1}{(625\mu^4 - 25p\mu^2 + 1)} = \frac{1}{576} + O(\epsilon)$. Setting $v = \beta(k + \rho) \in \langle k\rangle \oplus \langle k\rangle^\perp$ in (5.1) and projecting the result onto the span of $k(t) = \cos(t)$, we obtain

$$\mu^4 - p\mu^2 + 1 - \epsilon\beta^2\frac{1}{\pi}\int_0^{2\pi} k((k+\rho)^3 - \beta^2(k+\rho)^5)dt = 0. \tag{5.8}$$

We now use this information to equate coefficients at the appropriate orders to find

$$P_1 = -\frac{3}{4}, \quad P_2 = \frac{5}{8}, \quad \omega_1 = \frac{-9}{64}, \quad \text{and } \omega_2 = \frac{5}{48}. \tag{5.9}$$

5.1. Conditions for a fold bifurcation. The solution branch $C_\epsilon(1, 1)$ determined above, which branches from $p = 2$, can be continued from bifurcation in p for $p < 2$, or in μ for $\mu < 1$. When considered as a function of p, the branch has a fold bifurcation at some point, which we label p_F, and the same behavior is observed when the branch is continued in μ. The numerical calculations presented in the next section also indicate that the solution branch is restricted to the parameter range $p > p_F$ and $\mu > \mu_F$, although we have no proof of this claim.

We can now prove the following theorem.

THEOREM 5.1. *There is a neighborhood $I \subset \mathbb{R}$ of zero such that if $\epsilon \in I$, the zero-Hamiltonian branch $C_\epsilon(1, 1)$ associated with* (5.1)*, which bifurcates from $p = 2$ at $\mu = 1$, has a fold which occurs with respect to p at*

$$p_F(\epsilon) = 2 - \frac{9}{40}\epsilon + O(\epsilon^2). \tag{5.10}$$

There is also a fold in $C_\epsilon(1, 1)$ with respect to μ which occurs at

$$\mu_F(\epsilon) = 1 - \frac{243}{5120}\epsilon + O(\epsilon^2). \tag{5.11}$$

Proof. Using (5.6), (5.7), and (5.9), a fold bifurcation with respect to p occurs on the $C_\epsilon(1, 1)$ branch when the conditions $d_\beta p_\epsilon(\beta) = 0$ and $d^2_{\beta\beta} p_\epsilon(\beta) \neq 0$ are met. Applying the implicit function theorem when $\epsilon \neq 0$, these conditions are satisfied when $\beta^2 = \beta_F^2 \equiv 3/5 + \mathcal{O}(\epsilon)$, giving (5.10).

Similarly, a fold bifurcation with respect to μ occurs on the branch $C_\epsilon(1, 1)$ when $d_\beta \mu_\epsilon(\beta) = 0$ and $d^2_{\beta\beta}\mu_\epsilon(\beta) \neq 0$, and, provided $\epsilon \neq 0$, these conditions are satisfied when $\beta^2 = 27/40 + \mathcal{O}(\epsilon)$, giving (5.11). ∎

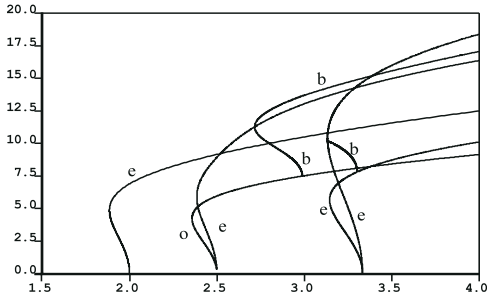

FIG. 6.1. *Bifurcations of zero-Hamiltonian periodic solutions from* $p = 2, p = 2\frac{1}{2}$, *and* $p = 3\frac{1}{3}$, *with* p *plotted horizontally against* $\|u'\|_{L^\infty}$ *vertically;* e *are even solutions,* o *are odd solutions, and* b *are solutions with* broken *symmetry.*

6. Numerical computations.

6.1. Preliminaries. We now describe a series of numerical calculations to determine solutions of the unscaled differential equation (1.1) with the restiffening foundation whose primary solution branch was studied in the previous section:

$$u'''' + pu'' + u - \epsilon(u^3 - u^5) = 0. \tag{6.1}$$

We augment this with the periodic boundary conditions $u(0) = u(T), u'(0) = u'(T)$, and $u'''(0) = u'''(T)$ and specify the phase by requiring $u'(0) = 0$. Finally, we impose the constraint that the Hamiltonian is zero, so that

$$u''(0) = \pm\sqrt{2(u(0)^2/2 - \epsilon(u(0)^4/4 - u(0)^6/6)}, \tag{6.2}$$

and we set $\epsilon = 1/2$ for the purposes of computation.

In (6.2) the positive root corresponds to the solution which is tangential to the rescaled eigensolution $e_-(x) \equiv n\cos(x\sqrt{m/n}) - m\cos(x\sqrt{n/m})$ at the bifurcation point $(u, p, \mu) = (0, p_{n,m}, \mu_{n,m})$, whereas the negative root corresponds to the solution which is tangential to the eigensolution $e_+(x) \equiv n\cos(x\sqrt{m/n}) + m\cos(x\sqrt{n/m})$. In order to follow the solution branches in p and to detect fold bifurcations, the collocation-based code *AUTO* [14] was used.

6.2. Calculation of the solution branches. We now illustrate three cases regarding the bifurcation of solutions of (6.1): the case $(n, m) = (1, 1)$, for which there is a unique bifurcating branch; the case $(n, m) = (2, 1)$, with $p_{2,1} = 2\frac{1}{2}$ and $\mu_{2,1} = \frac{1}{\sqrt{2}}$; and $(n, m) = (3, 1)$, for which $p_{3,1} = 3\frac{1}{3}$ and $\mu_{1,3} = \frac{1}{\sqrt{3}}$. Broadly speaking, higher values of n and m lead to similar solution branches.

Figure 6.1 shows the bifurcation branches which are proven to exist in Theorems 3.1 and 3.3, with p plotted against $\|u'\|_{L^\infty}$. The $(1, 1)$ branch has the form described in Theorem 5.1, and for the $p = 2\frac{1}{2}$ and $p = 3\frac{1}{3}$ cases one sees a similar geometry in that the branches initially bifurcate to the left, have fold bifurcations, and then persist for all values of p to the right of the fold point.

The following comments are in order regarding Figure 6.1 and the three points $p = 2, 2\frac{1}{2}$, and $3\frac{1}{3}$. Due to the $\mathbb{Z}_2$-symmetry of (6.1) and of the symmetry properties of the eigenfunctions when $(n, m) = (2, 1)$, if $u(t)$ is one even solution on the $(2, 1)$ branch, then so too is $-u(t + T/2)$. Therefore, in order to obtain a second *distinct* periodic solution, we apply Theorem 3.5 to give the existence of two branches of odd

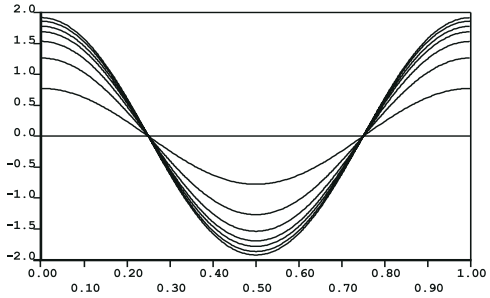

FIG. 6.2. *Solutions ($u(s)$ for $0 \le s \le 1$) bifurcating from $p = 2$, away from the bifurcation point. In accordance with Theorem* 2.2 *these are even about zero and odd about one-quarter.*

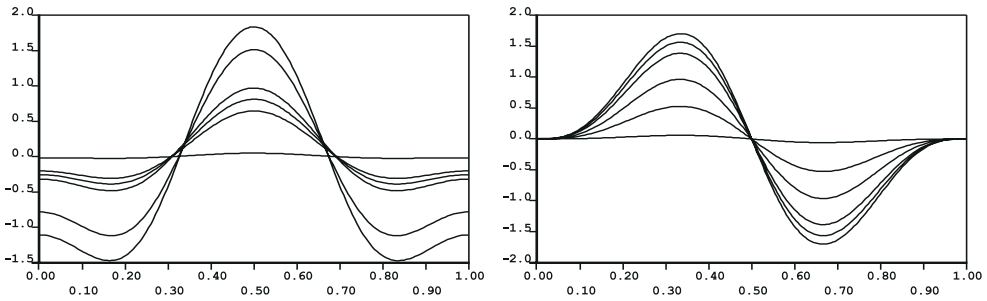

FIG. 6.3. *(left) Even solutions ($u(s)$ for $0 \le s \le 1$) bifurcating from $p = 2\frac{1}{2}$, away from the bifurcation point. (right) Odd solutions bifurcating from $p = 2\frac{1}{2}$.*

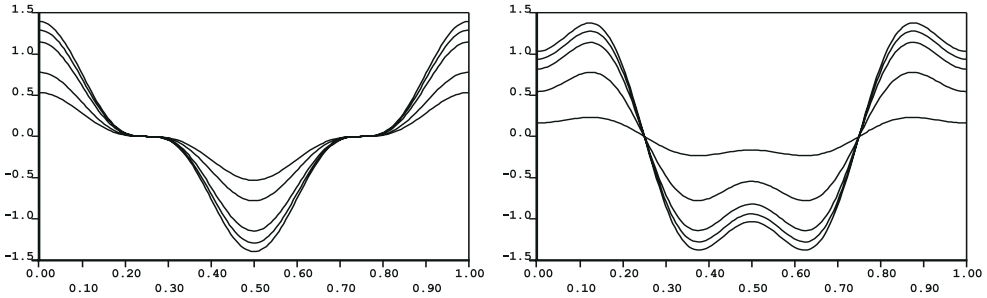

FIG. 6.4. *Solutions ($u(s)$ for $0 \le s \le 1$) bifurcating from $p = 3\frac{1}{3}$, away from the bifurcation point.*

solutions. Again, one of these branches of odd solutions can be obtained from the other by symmetry, and we therefore have plotted one of each even and odd branch in Figure 6.1. The solutions on this branch are shown in Figure 6.3.

Figures 6.2, 6.3, and 6.4 each show several solutions chosen from Figure 6.1 on the branches which connect to $p = 2, p = 2\frac{1}{2}$, and $p = 3\frac{1}{3}$, respectively, although the domain of each solution has been normalized to unity. (The information regarding the period of the solutions is given in Figure 6.5.) If we examine Figure 6.4, we notice that each of the solutions is even about zero and odd about one quarter. Consequently, the two branches of the solutions shown are in fact identical, up to a shift and a reflection, to the odd solutions which are obtained using Theorem 3.5.

Since $F(u)$ and $uF_u(u)$ are both positive for nonzero u when $\epsilon = 1/2$, the global

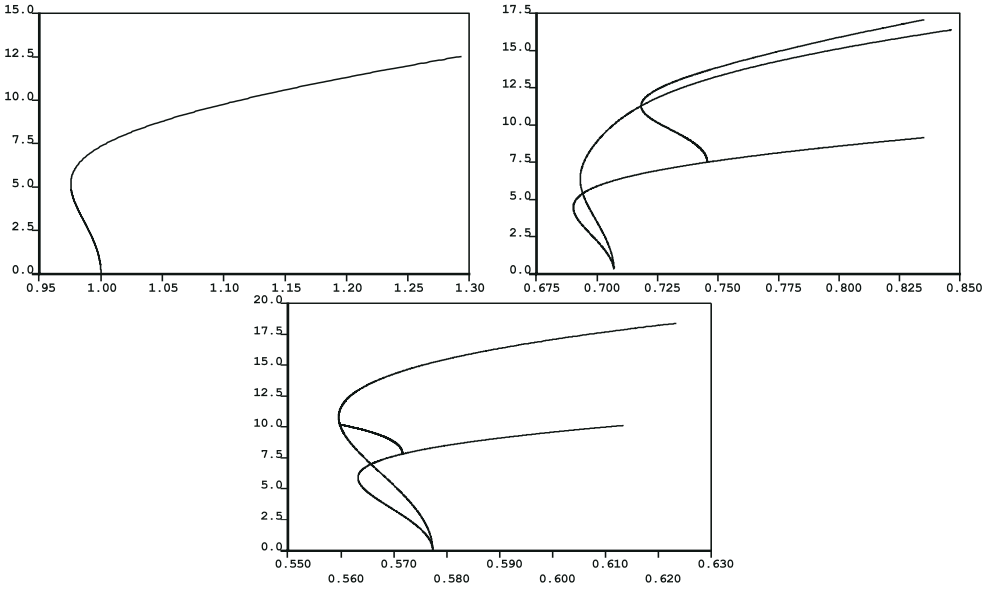

FIG. 6.5. *Period of solutions from Figure* 6.1 *(*$(1,1)$ *branch is top-left,* $(2,1)$ *top-right,* $(3,1)$ *bottom) with* $\mu = 2\pi/\text{period}$ *plotted horizontally and* $\|u'\|_{L^\infty}$ *vertically.*

bifurcation theorem (Theorem 4.2) applies to the $(1,1)$ branch, and the nodal properties are preserved along the resulting global branch in accordance with Theorems 3.6, 3.7, and 3.8. This is illustrated in each of Figures 6.2, 6.3, and 6.4.

Finally, note that the $(1,1)$ branch in Figure 6.1 appears to have no further bifurcations, whereas the $(2,1)$ and $(3,1)$ branches both have symmetry-breaking secondary bifurcation points. What is interesting about the resulting branches of unsymmetric solutions is that they form connections between the $(2,1)$ and $(3,1)$ branches. This indicates that it would be futile to seek generalizations of the results of section 3.5 to include the space of all periodic zero-Hamiltonian solutions of (1.1) and that the disjointness properties of the solution branches obtained in this paper are peculiar to spaces of symmetric solutions.

REFERENCES

[1] J. C. ALEXANDER AND J. A. YORKE, *Global bifurcations of periodic solutions*, Amer. J. Math., 100 (1978), pp. 263–292.

[2] A. AMBROSETTI AND G. PRODI, *A Primer of Nonlinear Analysis*, Cambridge Stud. Adv. Math. 34, Cambridge University Press, Cambridge, UK, 1992.

[3] C. J. AMICK AND J. F. TOLAND, *Global uniqueness of homoclinic orbits for a class of fourth-order equations*, Z. Angew. Math. Phys., 43 (1992), pp. 591–597.

[4] M. BOUGHARIOU, *Closed orbits of Hamiltonian systems on non-compact prescribed energy surfaces*, Discrete Contin. Dynam. Systems, 9 (2003), pp. 603–616.

[5] R. F. BROWN, *A Topological Introduction to Nonlinear Analysis*, Birkhäuser Boston, Cambridge, MA, 1993.

[6] B. BUFFONI, E. DANCER, AND J. TOLAND, *A variational theory of Stokes waves and their sub-harmonic bifurcations*, Arch. Ration. Mech. Anal., 2001, preprint; available online from http://www.maths.bath.ac.uk/MATHEMATICS/preprints.html — find preprint maths9801.

[7] B. BUFFONI AND J. F. TOLAND, *Global existence of homoclinic and periodic orbits for a class of autonomous Hamiltonian systems*, J. Differential Equations, 118 (1995), pp. 104–120.

[8] A. R. Champneys, *Homoclinic orbits in reversible systems and their applications in mechanics, fluids and optics*, Phys. D, 112 (1998), pp. 158–186.
[9] A. R. Champneys and J. M. T. Thompson, *A multiplicity of localized buckling modes for twisted rod equations*, Proc. Royal Soc. Ser. A, 452 (1996), pp. 2467–2491.
[10] J. V. Chaparova, L. A. Peletier, and S. A. Tersian, *Existence and nonexistence of nontrivial solutions of semilinar fourth and sixth-order differential equations*, preprint, from http://www.math.leidenuniv.nl/∼peletier.
[11] S. N. Chow and J. K. Hale, *Methods of Bifurcation Theory*, Springer-Verlag, New York, 1982.
[12] C. J. Budd, G. W. Hunt, and R. A. Kuske, *Cellular buckling close to Maxwell load*, Proc. Roy. Soc. London Ser. A, 457 (2001), pp. 2935–2964.
[13] R. L. Devaney, *Blue sky catastrophes in reversible and Hamiltonian systems*, Indiana Univ. Math. J., 26 (1977), pp. 247–263.
[14] E. J. Doedel, A. R. Champneys, T. F. Fairgrieve, Y. A. Kuznetsov, B. Sandstede, and X.-J. Wang, *AUTO*97: *Continuation and Bifurcation Software for Ordinary Differential Equations*, Tech. report, Department of Computer Science, Concordia University, Montreal, Canada, 1997; available by FTP from ftp.cs.concordia.ca in directory pub/doedel/auto.
[15] W. Eckhaus, *Singular perturbations of homoclinic orbits in* $\mathbb{R}^4$, SIAM J. Math. Anal., 23 (1992), pp. 1269–1290.
[16] M. Feckan and V. Rothos, *Bifurcations of periodics from homoclinics in singular ODE: Applications to discretizations of travelling waves of PDE*, Comm. Pure Appl. Anal., 1 (2002), pp. 475–483.
[17] H. Hofer and J. F. Toland, *On the existence of homoclinic, heteroclinic and periodic orbits for a class of indefinite Hamiltonian systems*, Math. Ann., 12 (1984), pp. 387–403.
[18] G. W. Hunt, G. J. Lord, and A. R. Champneys, *Homoclinic and heteroclinic orbits underlying the post-buckling of axially compressed cylindrical shells*, Comput. Methods Mech. Engrg., 170 (1999), pp. 239–251.
[19] G. W. Hunt, M. A. Peletier, A. R. Champneys, P. D. Woods, M. A. Wadee, C. J. Budd, and G. J. Lord, *Cellular buckling in long structures*, Nonlinear Dynamics, 1 (2000), pp. 3–29.
[20] W. D. Kalies, J. Kwapisz, J. van den Berg, and R. van der Vorst, *Homotopy classes for stable periodic and chaotic patterns in fourth-order Hamiltonian systems*, Comm. Math. Phys., 214 (2000), pp. 573–592.
[21] Y. S. Kivshar, A. R. Champneys, D. Cai, and A. R. Bishop, *Multiple states of intrinsic localised modes*, Phys. Rev. B, 58 (1988), pp. 5423–5428.
[22] J. Knobloch and A. Vanderbauwhede, *A general reduction method for periodic solutions in conservative and reversible systems*, Dyn. Difference Equations, 8 (1996), pp. 71–102.
[23] C. Lazzari, *Symmetries and exponential smallness of bifurcation functions of a class of singular, reversible systems*, Nonlinear Anal., 33 (1998), pp. 759–772.
[24] V. J. Mizel, L. A. Peletier, and W. C. Troy, *Periodic phases in second-order materials*, Arch. Ration. Mech. Anal, 145 (1998), pp. 343–382.
[25] L. Peletier and W. Troy, *Spatial patterns, higher order models in physics and mechanics*, Progress in Nonlinear Differential Equations and Their Applications, Vol. 45, Birkhäuser Boston, Cambridge, MA, 2001.
[26] L. A. Peletier, A. I. Rotariu-Bruma, and W. C. Troy, *Pulse-like spatial patterns described by higher-order model equations*, J. Differential Equations, 150 (1998), pp. 124–187.
[27] L. A. Peletier and W. C. Troy, *Chaotic spatial patterns described by the extended Fisher–Kolmogorov equation*, J. Differential Equations, 129 (1996), pp. 458–508.
[28] L. A. Peletier and W. C. Troy, *Spatial patterns described by the extended Fisher–Kolmogorov equation: Periodic solutions*, SIAM J. Math. Anal., 28 (1997), pp. 1318–1354.
[29] M. A. Peletier, *Non-existence and uniqueness for fourth-order Hamiltonian systems*, Nonlinearity, 12 (1999), pp. 1555–1570.
[30] P. H. Rabinowitz, *Periodic solutions of a Hamiltonian system on a prescribed energy surface*, Differential Equations, 33 (1979), pp. 336–352.
[31] P. H. Rabinowitz, *The prescribed energy problem for periodic solutions of Hamiltonian systems*, Contemp. Math., 81 (1988), pp. 183–191.
[32] M. Taylor, *Partial Differential Equations: Basic Theory*, Appl. Math. Sci. 1, Springer-Verlag, New York, 1996.
[33] S. Tersian and J. Chaparova, *Periodic and homoclinic solutions of some semilinear sixth-order differential equations*, J. Math. Anal. Appl., 272 (2002), pp. 223–239.
[34] J. F. Toland, *Solitary wave solutions for a model of the two-way propagation of water waves in a channel*, Math. Proc. Cambridge Philos. Soc., 90 (1981), pp. 343–360.
[35] J. B. van den Berg, *The phase-plane picture for a class of fourth-order conservative differential equations*, J. Differential Equations, 161 (2000), pp. 110–153.

[36] J. B. van den Berg, L. A. Peletier, and W. C. Troy, *Global branches of multi-bump periodic solutions of the Swift–Hohenberg equation*, Arch. Ration. Mech. Anal, 158 (2001), pp. 91–153.

[37] A. Vanderbauwhede, *Local bifurcation and symmetry*, Res. Notes in Math. 75, Pitman, London, 1982.

[38] E. Zeidler, *Nonlinear Functional Analysis and Its Applications: Fixed Point Theorems*, Vol. 1, Springer-Verlag, New York, 1986.

SIAM J. MATH. ANAL.
Vol. 36, No. 5, pp. 1489–1521

LAWS FOR THE CAPILLARY PRESSURE IN A DETERMINISTIC MODEL FOR FRONTS IN POROUS MEDIA*

BEN SCHWEIZER†

Abstract. We propose and analyze a model for sharp fronts in porous media, aiming at an investigation of the capillary pressure. Using the notion of microlocal patterns we analyze the local behavior of the system. Depending on the structure of the local patterns we can derive upscaled equations that characterize the capillary pressure and include the hysteresis effect that is known from the physical system.

Key words. homogenization, two-phase flow, front dynamics, microlocal pattern

AMS subject classifications. 74Q10, 76M50, 76T10

DOI. 10.1137/S0036141003423053

1. Introduction. The investigation of fluid motion in porous media has attracted much interest in the fields of engineering, physics, and mathematics. A particular interest concerns the case when two immiscible fluids are contained in the porous material, e.g., water and oil in rock. Different suggestions were made for averaged equations for this two-fluid motion; most are variants of the very successful Muskat–Leverett equations. In these equations the motion of the two fluids is coupled via an equation

$$p_a - p_b = p_c(s), \tag{1.1}$$

where p_a and p_b are the pressure functions in the two fluids, s is the saturation of, say, fluid a, and p_c is the capillary pressure. Our aim is to derive (1.1) for a model system.

We refer the reader to [2, 5, 7, 9] for other approaches towards the justification of the Muskat–Leverett equations. Concerning modifications of the system see [3, 4]; note that the result of this work and of [10] suggests another modification. For an analysis of the upscaled system see [8].

The far aim would be the homogenization of the geometry of Figure 1(a). This goal seems to be out of reach due to the topological changes of the free boundary during its propagation. A simpler geometry is the filter geometry of Figure 1(b). Here in every single tube the free boundary has essentially only one degree of freedom, its average height. By the laws for surface tension and contact angles, the geometry implies for every tube k a relation between average height and typical pressure, $p = \mathcal{P}_0(h, k)$. We will study this simplified model. For a homogenization result for the filter geometry with the same methods see [11].

We observe the creation of local structures: when the pressure in tube k reaches its maximal value and the height exceeds the critical point, an instability occurs. The pressure lowers; therefore, locally, the flow goes toward tube k, the height increases further, and the process accelerates. Such an event has the temporal and spatial scale of the tube distance ε and we call it an explosion. We will verify the appearance of

*Received by the editors February 19, 2003; accepted for publication (in revised form) June 11, 2004; published electronically March 25, 2005.
http://www.siam.org/journals/sima/36-5/42305.html
†Institut für Angewandte Mathematik, Universitaet Heidelberg, INF 294, D-69120 Heidelberg, Germany (schweizer@iwr.uni-heidelberg.de).

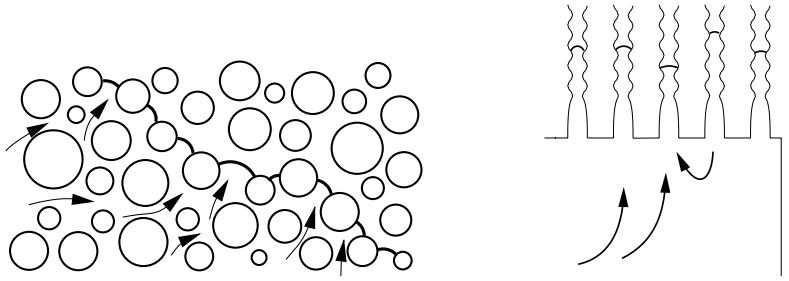

FIG. 1. *(Left) Front in porous media. (Right) Filter geometry.*

these explosions and determine their form. The distribution of the explosions can be captured using the Young measure on patterns introduced by Alberti and Müller in [1]. We will derive a conditional result for the upscaled equations, i.e., the limit $\varepsilon \to 0$: if the patterns of the limit measure all have finite length, then upscaled equations with a prescribed capillary pressure along the front are satisfied in the limit.

1.1. A model for propagating fronts.

Geometry. The fluid occupies the domain $\Omega := (-1,1)\times(-1,0)$ and is described by a Darcy law. The front is located along the upper boundary

$$\Gamma := (-1,1) \times \{0\}.$$

Γ consists of two parts. On $\Gamma_2^\varepsilon \subset \Gamma$ the fluid is in contact with the matrix; with $\gamma \in (0,1/2)$ we write

$$\Gamma_2^\varepsilon := \varepsilon \cdot (\mathbb{Z} + (\gamma, 1-\gamma)) \cap \Gamma,$$

where we used the obvious identification $\Gamma \subset \mathbb{R}$. The small parameter ε describes the pore size in the medium; we will for simplicity always assume $\varepsilon = 1/N$ with $N \in \mathbb{N}$. A fluid-gas interphase is present along

$$\Gamma_1^\varepsilon := \varepsilon \cdot (\mathbb{Z} + (-\gamma, \gamma)) \cap \Gamma.$$

The free boundary is modeled with a height function

$$h^\varepsilon : \Gamma_1^\varepsilon \to \mathbb{R}.$$

Having in mind that the free boundary in the single tube has only one degree of freedom, we reinterpret h^ε as the average height and assume that h^ε is piecewise constant. We use the space $V_\varepsilon = Q_\varepsilon L^2(\mathbb{R})$ of functions that are constant on the intervals $\varepsilon \cdot ((-\gamma,\gamma)+k)$, $k \in \mathbb{Z}$. Here Q_ε is the L^2-projection to this subspace; it is obtained by replacing a function on Γ_1^ε with its averages on the disjoint intervals of length $2\gamma\varepsilon$.

Equations. We write the Darcy law in the scaling $v^\varepsilon = -\varepsilon \nabla p^\varepsilon$. This scaling is obtained by rescaling time and is appropriate to observe microscopic processes. Note that we assumed the permeability matrix to be the identity in order to simplify

notations. We study the problem

$$\partial_t h^\varepsilon(x_1,t) = -\varepsilon Q_\varepsilon \partial_2 p^\varepsilon(x_1,0,t) \qquad \forall (x_1,0)\in\Gamma_1^\varepsilon,\ \forall t, \tag{1.2}$$

$$p^\varepsilon(x_1,0,t) = \mathcal{P}_0\left(\frac{x_1}{\varepsilon}, \frac{h^\varepsilon(x_1,t)}{\varepsilon}\right) \qquad \forall (x_1,0)\in\Gamma_1^\varepsilon,\ \forall t, \tag{1.3}$$

$$\partial_2 p^\varepsilon(x_1,0,t) = 0 \qquad \forall (x_1,0)\in\Gamma_2^\varepsilon,\ \forall t, \tag{1.4}$$

$$-\Delta_x p^\varepsilon = 0 \qquad \text{in } \Omega\times(0,T). \tag{1.5}$$

Initial values for the height function are given and we always set $h^\varepsilon(x_1) = 0$ for $(x_1,0)\in\Gamma_2^\varepsilon$. The equations are complemented with the ε-independent boundary conditions. We impose periodicity on the lateral boundaries $\Sigma_\pm := \{(x_1,x_2)\in\bar\Omega \mid x_1 = \pm 1\}$. The presented methods apply also in the case of an impermeability condition. As a driving mechanism we choose a prescribed inflow on the lower boundary $\Gamma_0 := (-1,1)\times\{-1\}$,

$$-\partial_2 p^\varepsilon(x_1,-1,t) = V_0(x_1) \qquad \forall (x_1,-1)\in\Gamma_0,\ \forall t. \tag{1.6}$$

It is also possible to prescribe the pressure at the lower boundary. In either case and throughout our investigations we demand $p^\varepsilon|_{\Gamma_0}(t) > 0$ for all t. This assumption is made to simplify notations; upscaled equations in the general case follow by symmetry.

It is left to specify the material law $\mathcal{P}_0$ which prescribes the pressure-height dependence in each cell (we will always assume that $\mathcal{P}_0(\cdot,s)$ is constant on $(k-\gamma,k+\gamma)$ for every $k\in\mathbb{Z}$). A reasonable choice is the following. The material law of the cells is the same in every cell and a sawtooth function in $s = \frac{h}{\varepsilon}$,

$$\mathcal{P}_0(k,s) \equiv \mathcal{P}_0(s) = a_0\cdot s \mod a_0 s_0. \tag{1.7}$$

Here s_0 represents the volume of the single cell. The maximal pressure that is needed to advance the free boundary is

$$p_{\max} = a_0\cdot s_0.$$

A possible modification of this model is to allow the physical parameters $a_0 = a_0(k)$ and $s_0 = s_0(k)$ to depend on the position index k. If (a_0,s_0) is periodic in k, all results remain valid. We collect some first observations on the ε-problem. The proofs are straightforward and can be found in [11].

Remark 1.1. The ε-problem has a unique solution with $p^\varepsilon(t)\in H^1(\Omega)$ for all $t\in[0,T]$. The solution sequence p^ε satisfies uniform bounds in $L^\infty((0,T),L^\infty(\Omega))$ and in $L^2((0,T),H^1(\Omega))$.

The uniform bound of $p^\varepsilon\in L^\infty(\Omega)$ allows us to choose a subsequence $\varepsilon\to 0$ such that p^ε has a limit $p^0\in L^\infty(\Omega)$ in the sense of the weak-$\star$ convergence in L^∞,

$$p^\varepsilon \to p^0 \quad \text{in } L^\infty_w.$$

1.2. Main result. Our aim is to find equations for p^0 in order to describe the averaged behavior of the solutions p^ε. It turns out that, in general, p^0 is not uniquely determined. We will have to study the microscopic behavior of the family p^ε in order to find equations for p^0.

In the model equations with relation (1.7) the pressure p^ε has discontinuities. There are points $(x_1,0,t)\in\varepsilon\mathbb{Z}\times(0,T)$ in which the pressure drops from $p_{\max}$ to 0; we call them explosion points. We have already seen that we can expect a nontrivial

behavior in an ε-space-time neighborhood of explosion points. Generically, we expect that the explosion points do not cluster and that the limit patterns of explosions are finite (see Definition 2.8 for a precise statement). In fact, in [10] we analyze a stochastic system and find that explosion points cluster only with probability 0.

If all realized explosion patterns are finite, then the limit pressure p^0 satisfies the following upscaled system in the distributional sense (see Theorem 4.6 for details):

$$\Delta p^0 = 0 \qquad \text{in } \Omega, \tag{1.8}$$

$$-\partial_2 p^0 = V_0 \qquad \text{on } \Gamma_0, \tag{1.9}$$

and p^0 is periodic across $\Sigma_\pm$. On the boundary Γ it satisfies

$$p^0 \le p_{\max}, \tag{1.10}$$

$$\partial_t(\Theta \circ p^0) \le -\partial_2 p^0, \tag{1.11}$$

$$\partial_t(\Theta \circ p^0) = -\partial_2 p^0 \quad \text{on } \{(x_1,0) \in \Gamma \mid p^0(x_1,0) < p_{\max}\}. \tag{1.12}$$

The function Θ is defined in the stochastic case as an expected value. Since we consider the deterministic equation (1.7), its definition reduces to

$$\Theta'(\rho) = \left\langle \frac{2\gamma}{\mathcal{P}_0'(k, s_k(\rho))} \right\rangle = \frac{2\gamma}{a_0}, \tag{1.13}$$

where $s_k(\rho)$ is defined by $\mathcal{P}_0(k, s_k(\rho)) = \rho$. The microlocal patterns of the functions p^ε can be described.

Remarks on Theorem 4.6. (1) An assumption on the distribution of explosions is indeed necessary. Consider a family of solutions to the ε-problems that is ε-periodic in x_1-direction. In this case, the limiting pressure p^0 is constant along Γ, but has jumps from $p_{\max}$ to 0 at discrete times. The function p^0 does not solve the upscaled equation (1.12).

(2) We study a sawtooth function as a material law in (1.7). This law is not satisfactory for all applications. Unfortunately, in our proofs we need a condition on positivity of $\mathcal{P}_0'$ away from discontinuities, and this restricts our choices for $\mathcal{P}_0$. We conjecture that the overall picture about solution sequences and the upscaled equations remain valid for continuous functions $\mathcal{P}_0$.

(3) One can formally relate the upscaled equation (1.12) to well-known effective conductivity formulae. Differentiating (1.3) with respect to time and inserting (1.3), we find that $\partial_t p^\varepsilon$ essentially equals $-a_0 \partial_2 p^\varepsilon$. Dividing (1.12) by Θ', we find that $\partial_t p^0$ essentially equals $-\bar{a}\partial_2 p^\varepsilon$, where $\bar{a}$ is the harmonic mean of the a_0.

2. Microlocal patterns and possible patterns for fronts. We study solutions $(p^\varepsilon, h^\varepsilon)$ of (1.2)–(1.6) and are interested in the averaged behavior as it is expressed by the weak limit p^0. The goal is to derive averaged equations that characterize p^0.

We already observed that the equations have a nontrivial behavior on an ε-scale in time and in space. In this scaling we expect to see the filling procedure of the single pore: fluid enters the cell $\varepsilon(k-\gamma, k+\gamma)$ until the pressure reaches its maximal value $p_{\max}$. Now the pressure is set to zero and a pressure gradient of order $1/\varepsilon$ drives a refill procedure in which mass is transported from neighboring cells to cell k. It takes a time span of order ε and a spatial area with diameter of order ε to essentially reach again the pressure $p_{\max}$.

We want to find descriptions of this microscopic process. An adequate tool is that of microlocal patterns, introduced by Alberti and Müller in [1]. We outline aspects of this tool in this section.

Notation for measures. Let E be a locally compact Hausdorff space. We denote by $\mathcal{M}(E)$ the space of all finite real Borel measures on E. Let $C(E)$ be the space of continuous functions on E with compact support. Then $\mathcal{M}(E)$ can be identified with the dual of $C(E)$. Therefore bounded sequences in $\mathcal{M}(E)$ are precompact.

2.1. Construction of microlocal patterns. Let $u^\varepsilon : S \to \mathbb{R}$ be a sequence of functions on a compact set $S \subset \mathbb{R}^m$. We assume that for $C \in \mathbb{R}$ there holds $\|u^\varepsilon\|_{L^\infty} \le C$ for all $\varepsilon > 0$. Our aim is now to study the behavior of u^ε on the length scale ε. We therefore consider the blowup of u^ε, just as in asymptotic expansions or in the theory of two-scale convergence. Together with u^ε we consider the local pattern around $s \in S$, that is, the function

$$\mathbb{R}^m \ni t \mapsto R_s^\varepsilon u^\varepsilon(t) := u^\varepsilon(s + \varepsilon t)$$

(we assume that we extended trivially the original function u^ε outside S). The pattern $R_s^\varepsilon u^\varepsilon$ is bounded in $L^\infty(\mathbb{R}^m)$ by C. As the space of patterns we use the closed ball

$$K := \bar{B}_C(0) \subset L^\infty_w(\mathbb{R}^m),$$

where L^∞_w indicates that we use the weak-$\star$ topology on K. This makes K compact.

Since the pattern depends in an oscillatory fashion on s, one proceeds as in the construction of Young measures and considers instead of the values $R_s^\varepsilon u^\varepsilon \in K$ the measure ν_s^ε, the Dirac measure on K in the point $R_s^\varepsilon u^\varepsilon$.

DEFINITION 2.1 (measure of microlocal patterns). *Given a sequence $u^\varepsilon \in \bar{B}_C(0) \subset L^\infty(S)$ and the corresponding Dirac measures ν_s^ε, we define the measure ν^ε on $S \times K$ by*

$$S \times K \supset \bar{S} \times \bar{K} \mapsto \nu^\varepsilon(\bar{S}, \bar{K}) := \int_{\bar{S}} \nu_s^\varepsilon(\bar{K})\, de^\varepsilon(s)$$

for all Borel sets $\bar{S} \subset S$ and $\bar{K} \subset K$. The energy density $e^\varepsilon(s)$ still has to be specified; we always assume $\int_S de^\varepsilon(s) \le C'$. Then we can choose a subsequence $\varepsilon \to 0$ such that for a finite limit measure ν there holds in the sense of weak-$\star$ convergence

$$\nu = \lim_{\varepsilon \to 0} \nu^\varepsilon \in \mathcal{M}(S \times K).$$

We call ν the measure of microlocal patterns of the (sub)sequence.

We will also use the following two projections of ν. The projection μ of ν to S, which coincides with the energy density of ν,

$$S \supset \bar{S} \mapsto \mu(\bar{S}) := \nu(\bar{S} \times K) = \lim_{\varepsilon \to 0} \int_{\bar{S}} de^\varepsilon(s),$$

and the projection ν_K of ν to K,

$$\bar{K} \mapsto \nu_K(\bar{K}) := \nu(S \times \bar{K}).$$

The elements in the support of ν_K are the realized patterns.

Possible micropatterns. Our aim is to characterize the realized patterns of solution sequences in our model problem. The method is to successively exclude possibilities. A first concept is that of *possible micropatterns*. In the subsequent definition we assume that the sequence u^ε is a sequence of solutions to a given family of equations.

DEFINITION 2.2 (possible micropattern). *For a point $s \in S$ we say that $U_s : \mathbb{R}^m \to \mathbb{R}$ is a possible micropattern in s if there exist a sequence $\varepsilon \to 0$, boundary and initial conditions with a solution sequence u^ε, and a sequence of points $s_\varepsilon \to s$, such that*

$$U^\varepsilon_{s_\varepsilon} := R^\varepsilon_{s_\varepsilon} u^\varepsilon \to U_s \quad \text{in } K.$$

U is a possible micropattern, if it is a possible micropattern in some point s.

The concept of possible micropatterns does not allow us to study the distribution of patterns, but it gives a necessary condition for a pattern to be contained in the support of ν_K. In fact, an elementary proof yields the following result.

Remark 2.3. Every measure on patterns ν_K of a solution sequence u^ε has its support contained in the set of possible micropatterns.

The fundamental property of possible patterns is that they satisfy the rescaled equations (note that we loose boundary and initial conditions). This fact will lead to a characterization of possible micropatterns.

2.2. Possible patterns for the motion of fronts. As described, we expect ε-scale phenomena in the neighborhood of points (x,t) where p^ε has a jump, i.e., in explosion points. In order to study the local behavior, we define for a given $(x,t) = (x_1,0,t)$ the blowup of solutions

$$\begin{aligned} P^\varepsilon(y,\tau) = P^\varepsilon_{(x,t)}(y,\tau) &:= (R^\varepsilon_{(x,t)} p^\varepsilon)(y,\tau) = p^\varepsilon(x+\varepsilon y, t+\varepsilon\tau), \\ H^\varepsilon(y_1,\tau) = H^\varepsilon_{(x_1,t)}(y_1,\tau) &:= \frac{1}{\varepsilon}(R^\varepsilon_{(x_1,t)} h^\varepsilon)(y_1,\tau) \\ &= \frac{1}{\varepsilon} h^\varepsilon(x_1+\varepsilon y_1, t+\varepsilon\tau). \end{aligned}$$

For explosion points (x,t) we expect nontrivial limits of $(P^\varepsilon, H^\varepsilon)$ and our next aim is to determine the possible limits. We will call such limits explosion patterns.

The equations for $(P^\varepsilon, H^\varepsilon)$ are identical to (1.2)–(1.5), except that the factors ε are replaced by the factor 1 in (1.2) and (1.3). We have to substitute the boundary condition (1.6) by a condition on the asymptotic behavior at infinity.

The solutions P^ε are physically meaningful on space-time domains of size $\frac{1}{\varepsilon} \times \frac{1}{\varepsilon}$. After the trivial extension of the functions, their domains are

$$\begin{aligned} \bar\Omega &:= \mathbb{R} \times \mathbb{R}_-, & \bar\Gamma &:= \mathbb{R} \times \{0\} \equiv \mathbb{R}, \\ \bar\Gamma_1 &:= \mathbb{Z} + (-\gamma, \gamma), & \bar\Gamma_2 &:= \mathbb{Z} + (\gamma, 1-\gamma). \end{aligned}$$

In order to define the microlocal patterns we consider the functions $p^\varepsilon|_\Gamma : [-1,1] \times [0,T] \to \mathbb{R}$ and set $S := [-1,1] \times [0,T]$. A limit measure is found as $\nu \in \mathcal{M}(S \times K)$ with $K = \bar B_C(0) \subset L^\infty_w(\mathbb{R}^2)$. Note that $p^\varepsilon|_\Gamma$ determines uniquely its harmonic extension p^ε and the height function h^ε up to multiples of s_0. With this identification we can consider limit patterns also as functions $P : \bar\Omega \times \mathbb{R} \to \mathbb{R}$ and $H : \bar\Gamma \times \mathbb{R} \to [0, s_0)_{\text{per}}$.

We use the projection Q_1 defined as

$$(Q_1 f)(y) = \begin{cases} \dfrac{1}{2\gamma}\displaystyle\int_{k-\gamma}^{k+\gamma} f(\zeta)\, d\zeta & \text{for } y \in (k-\gamma, k+\gamma),\ k \in \mathbb{Z}, \\ f(y) & \text{for } y \in (k+\gamma, k+1-\gamma),\ k \in \mathbb{Z}. \end{cases}$$

Using Q_1 we can write the rescaled equations as (2.1)–(2.4). The next lemma states that every possible pattern is in fact a solution of these rescaled equations.

LEMMA 2.4. *Every possible micropattern (P, H) in a point $(x_1, 0, t)$ satisfies the rescaled equations*

$$\partial_\tau H(y_1, \tau) = -Q_1 \partial_2 P(y_1, 0, \tau) \qquad \forall (y_1, 0) \in \bar{\Gamma}_1, \tag{2.1}$$

$$P(y_1, 0, \tau) \in \hat{\mathcal{P}}_0(y_1, Q_1 H(y_1, \tau)) \qquad \forall (y_1, 0) \in \bar{\Gamma}_1, \tag{2.2}$$

$$\partial_2 P(y_1, 0, \tau) = 0 \qquad \forall (y_1, 0) \in \bar{\Gamma}_2, \tag{2.3}$$

$$-\Delta_y P(\cdot, \tau) = 0 \qquad \text{in } \bar{\Omega}, \tag{2.4}$$

for almost every $\tau \in \mathbb{R}$; $\hat{\mathcal{P}}_0$ is the multivalued function

$$\hat{\mathcal{P}}_0(k, s) = \begin{cases} \mathcal{P}_0(k, s) & \text{for } s \notin s_0\mathbb{Z}, \\ \{0, p_{\max}(k)\} & \text{for } s \in s_0\mathbb{Z}. \end{cases}$$

Additionally there holds

$$\|P\|_{L^\infty(\bar{\Omega})} \le C \qquad \forall t \in \mathbb{R}. \tag{2.5}$$

Remarks. (a) Inequality (2.5) must be interpreted as a boundary condition. (b) The well-posedness of the system can be shown by a limiting procedure with the help of Remark 1.1. (c) Remark A.1 suggests that the decay for $t \to +\infty$ is like t^{-1} (see also the appendix of [10]).

Proof. Let (P, H) be a possible micropattern. By definition there exist a solution sequence $(p^\varepsilon, h^\varepsilon)$ and points $s_\varepsilon = (x_\varepsilon, t_\varepsilon) \in \Gamma \times [0, T]$, such that the rescaled solutions $(P^\varepsilon_{s_\varepsilon}, H^\varepsilon_{s_\varepsilon})$ converge in L^∞_w to (P, H). For $H \neq 0$ one verifies $\varepsilon^{-1}\mathrm{dist}(x_\varepsilon, \varepsilon\mathbb{Z}) \to 0$; it is therefore no loss of generality to assume $x_\varepsilon \in \varepsilon\mathbb{Z}$. The sequence of rescaled solutions satisfies on increasing domains the rescaled equations (2.1)–(2.4).

We use the following weak form of the rescaled equations:

$$\int_{\mathbb{R}}\int_{\bar{\Omega}} \Delta_y \Phi \cdot P^\varepsilon_{s_\varepsilon} = -\int_{\mathbb{R}}\int_{\bar{\Gamma}_1} H^\varepsilon_{s_\varepsilon} \cdot \partial_\tau \Phi + \int_{\mathbb{R}}\int_{\bar{\Gamma}} P^\varepsilon_{s_\varepsilon}(y_1, 0, \tau) \cdot \partial_2 \Phi, \tag{2.6}$$

$$\int_{\mathbb{R}} Q_1 P^\varepsilon_{s_\varepsilon}(k, 0, \tau) \cdot \varphi(\tau)\, d\tau = \int_{\mathbb{R}} \mathcal{P}_0(k, Q_1 H^\varepsilon_{s_\varepsilon}(k, \tau)) \cdot \varphi(\tau)\, d\tau, \tag{2.7}$$

$$H^\varepsilon_{s_\varepsilon} - Q_1 H^\varepsilon_{s_\varepsilon} = 0, \quad P^\varepsilon_{s_\varepsilon}(\cdot, 0, \cdot) - Q_1 P^\varepsilon_{s_\varepsilon}(\cdot, 0, \cdot) = 0, \tag{2.8}$$

satisfied for all $\Phi \in C^2_0(\bar{\Omega} \cup \bar{\Gamma} \times \mathbb{R})$ with $\partial_2 \Phi = 0$ on $\bar{\Gamma}_2$ and $Q_1 \Phi|_{\bar{\Gamma}} = \Phi|_{\bar{\Gamma}}$, and all $\varphi \in C^0_0(\mathbb{R})$, $k \in \mathbb{Z}$. The equations are satisfied for all $\varepsilon < \varepsilon_0$, a threshold that depends on the support of the test functions and on k.

We can take the limit $\varepsilon \to 0$ along the subsequence in the weak equations. We find that (P, H) again solves the weak equations (2.6) and (2.8). In particular, P is harmonic, satisfies a homogeneous Neumann condition on $\bar{\Gamma}_2$, and is piecewise constant on $\bar{\Gamma}_1$ for a.e. τ. Together with the L^∞-estimate this implies spatial continuity

of P, and an L^∞-bound for $Q_1 \partial_\tau H$. Therefore (2.1), (2.3), and (2.4) are satisfied in the strong sense.

It remains to take the limit in the nonlinear material law in (2.7). We exploit the fact that $Q_1 H^\varepsilon$ converges uniformly on compact sets to $Q_1 H$, and conclude (2.2). ∎

Remark 2.5. If the maximal pressure $p_{\max} = p_{\max}(k)$ is independent of k, then every solution of (2.1)–(2.5) satisfies

$$P(x,t) \leq p_{\max} \qquad \forall x \in \bar{\Omega},\, t \in \mathbb{R}.$$

Proof. By the comparison principle for harmonic functions, $P(\cdot,t)$ is bounded on the finite domain $(-M,M)\times(-M,0)$ by the periodic harmonic functions q_M with

$$\begin{aligned} &q_M = p_{\max} \quad \text{on } \bar{\Gamma}_1, \qquad \partial_2 q_M = 0 \quad \text{on } \bar{\Gamma}_2, \\ &q_M = C \quad \text{on } (-M,M)\times\{-M\}. \end{aligned}$$

For $M \to \infty$ the sequence of functions q_M tends to $p_{\max}$ on every bounded set. This implies the result. ∎

PROPOSITION 2.6. *We consider a solution (P,H) of system (2.1)–(2.5). On the material law we assume $\partial_s \mathcal{P}_0 > 0$ on $(0,s_0)$. If $p_{\max}$ is independent of k, then (P,H) can only be*

(a) *a constant solution, $P(\cdot) \equiv p^* \in [0,p_{\max}]$ in $\bar{\Omega}\times\mathbb{R}$,*

(b) *a solution with simultaneous explosions at an explosion time T_0 and with an outlet pattern $\alpha \in \{0,1\}^{\mathbb{Z}}$,*

$$\begin{aligned} &P(\cdot, T_0+\tau) = p_{\max} \qquad \forall \tau < 0, \\ &P(\cdot, T_0+\tau) = P_\alpha(\cdot,\tau) \quad \forall \tau > 0. \end{aligned}$$

Here P_α is the explosion solution to the opening pattern α: P_α is the unique solution of (2.1)–(2.5) to the initial values

$$P_\alpha(y_1,0,t=0) = \begin{cases} p_{\max} & \text{for } y_1 \in (k-\gamma,k+\gamma),\ \alpha(k)=1, \\ 0 & \text{for } y_1 \in (k-\gamma,k+\gamma),\ \alpha(k)=0. \end{cases}$$

Proof. Let (P,H) be given. We claim that if the solution satisfies $P(\cdot,T_0) \not\equiv p_{\max}$ for some $T_0 \in \mathbb{R}$, then no explosion can happen at a later time.

We compare P in the neighborhood of a single cell, say

$$(y_1,y_2) \in R := (-1/2,1/2)\times(-1,0),$$

with the solution q of

$$\begin{aligned} \Delta_y q(t) &= 0 && \text{in } R,\ \forall t \in (T_0,\infty), \\ q(\cdot,0,t) &= q_0(t) && \text{on } (-\gamma,\gamma),\ \forall t \in (T_0,\infty), \\ \partial_2 q &= 0 && \text{on } (-1/2,-\gamma)\cup(\gamma,1/2), \\ q &= p_{\max} && \text{on } \partial R \backslash \bar{\Gamma}, \\ \partial_t q_0 &= -c_q \frac{1}{2\gamma} \int_{-\gamma}^{\gamma} \partial_2 q. \end{aligned}$$

The initial value for q_0 and the constant c_q remain to be chosen.

We claim that for $q_0(t=0) < p_{\max}$ the function q_0 remains below $p_{\max}$ for all times. Since the minimum of q is on Γ_1, there holds $-\partial_2 q \geq 0$ on Γ_1. For some positive $\bar{c}$ we have the estimate

$$\frac{1}{2\gamma}\int_{-\gamma}^{\gamma}(-\partial_2 q) \leq \bar{c}(p_{\max} - q_0).$$

This holds since the left-hand side is finite for fixed q_0 and a linear function of the difference $p_{\max} - q_0$. The estimate implies that q_0 grows with a speed at most proportional to $p_{\max} - q_0$. This implies that a convergence of q_0 to $p_{\max}$ has at most exponential rate.

We now use q as a comparison function for P. In a cell with $P(y_1, 0, T_0) < p_{\max}$ we choose $c_q > \sup_s \partial_s \mathcal{P}_0$ and $p_{\max} > q_0(t=T_0) > P(y_1,0,T_0)$. Then $P \leq q$ holds for $t = T_0$ on R. If $P = q$ for the first time at some point in R, coincidence holds also on a point of the boundary Γ_1, and we have $P = q$ along Γ_1, since both functions are constant. Then $-\partial_2 q \geq -\partial_2 P$ along Γ_1 and we find

$$\begin{aligned}\partial_t P = \partial_s \mathcal{P}_0 \cdot \partial_t H &= -\partial_s \mathcal{P}_0 \cdot \frac{1}{2\gamma}\int_{-\gamma}^{\gamma}\partial_2 P \\ &\leq -\partial_s \mathcal{P}_0 \cdot \frac{1}{2\gamma}\int_{-\gamma}^{\gamma}\partial_2 q < -c_q \frac{1}{2\gamma}\int_{-\gamma}^{\gamma}\partial_2 q = \partial_t q_0.\end{aligned}$$

Therefore $P \leq q$ holds for all times and also P does not reach the value $p_{\max}$ in finite time. In cells with $P(y_1, t) = p_{\max}$ the Hopf lemma implies $\partial_t P(y_1, 0, t) < 0$, and we can apply the above argument for $|t - T_0|$ small.

We claim that solutions without explosions are constants. In (A.4) of Proposition A.3, we demonstrate that a solution without explosions on the time interval $(0, \infty)$ converges to a constant function for $t \to \infty$, independent of the initial values. Since the solution (P, H) is defined for all negative times we conclude that $P(t)$ is constant for all t. The uniqueness of the solution P_α follows from the linearized stability. ∎

By the above proposition we know that there exist only a few possible micropatterns. Next we want to use this as information about microlocal patterns. To this end we have to choose an energy density.

DEFINITION 2.7. *The measure of microlocal patterns ν of the sequence p^ε is defined via the energy density e^ε, which we construct as a sum of Dirac measures. We define the finite set of explosion points by*

$$M^\varepsilon := \{s = (x_1, 0, t) \in \Gamma \times (0, T) \mid p^\varepsilon(s) = 0,\ x_1 \in \varepsilon\mathbb{Z}\},$$

and set

$$e^\varepsilon(\bar{S}) := \varepsilon \sum_{(x_1,0,t)\in M^\varepsilon \cap \bar{S}} s_0(x_1 \varepsilon^{-1}).$$

The set M^ε is finite, since $\partial_t p^\varepsilon > 0$ holds in points of M^ε. The sequence of measures e^ε is bounded. Since H^ε never passes a value in $s_0\mathbb{Z}$ in the negative direction, every explosion corresponds to a loss $\varepsilon 2\gamma s_0$ of fluid mass. Therefore in the space-time volume $\bar{S} = \Gamma \times (t_1, t_2)$ there holds by conservation of mass

$$|M^\varepsilon| \leq \frac{2}{\varepsilon} + \frac{\|V_0\|_{L^1}|t_2 - t_1|}{2\gamma\varepsilon \inf\{s_0(\cdot)\}},$$

which implies

$$e^\varepsilon(\bar S) \le \left(2 + \frac{\|V_0\|_{L^1}|t_2 - t_1|}{2\gamma \inf\{s_0(\cdot)\}}\right) \sup\{s_0(\cdot)\}.$$

Remark. The disadvantage of the above energy density e^ε is that it can be defined only in the case that explosions can be localized to a point; for a continuous $\mathcal{P}_0$-function we cannot define an analogue of it.

In the continuous case one could introduce an energy density by setting

$$\tilde e^\varepsilon(x,t) := \frac{1}{2\gamma}(v_2^\varepsilon(x,0,t))_+.$$

The two energy densities have many similarities; note the disadvantage of $\tilde e^\varepsilon$ that it can become positive also because of oscillations without explosions.

We saw that the qualitative behavior of the limit p^0 is different in the periodic situation and in the (expected) physical situation of explosions that are far from each other. The limit measure ν allows us to distinguish the two situations in terms of the explosion patterns. In the periodic case, the patterns will also be periodic; i.e., ν is supported on periodic and therefore infinite patterns. With the subsequent definition we distinguish the two cases.

DEFINITION 2.8. *We say that ν is of* finite type, *if for some constant $C_f \in \mathbb{N}$ there holds the following: every observable explosion pattern has at most C_f explosion points, i.e.,*

$$P_\alpha \in \operatorname{supp}(\nu_K) \Rightarrow |\alpha| := |\{k \in \mathbb{Z} : \alpha(k) = 1\}| \le C_f.$$

Consider the energy density e^ε and assume that ν is of finite type. Then

$$\operatorname{supp}(\nu_K) \subset \bigcup_{\alpha\in A} \{P_\alpha\} \quad \text{with } A \subset \{\alpha \in \{0,1\}^{\mathbb{Z}} : |\alpha| \le C_f\}.$$

The limit measure ν can then be written as

$$\nu(\bar S \times \bar K) = \sum_{\alpha\in A} \int_{\bar S} \eta_\alpha(\bar K) d\mu_\alpha(s), \tag{2.9}$$

where μ_α are measures on $\Gamma \times [0,T]$ and η_α is the Dirac measure on K on the explosion solution P_α.

We exploited that the constant pressure solutions do not contribute to the energy and in (2.9) that A is countable.

3. Limit measures of finite type and regions without explosions. We continue our study of a sequence of solutions $(p^\varepsilon, h^\varepsilon)$ and their L^∞_w-limit p^0. In order to derive and even to formulate upscaled equations for p^0 we need some regularity result for p^ε and p^0. In the case that the limit measure ν is of finite type, a fundamental regularity statement holds: loosely speaking, spatial averages of the pressure p^0 cannot have jumps in time. This, in turn, helps us to find regions without explosions: points with $p^0 < p_{\max}$ have neighborhoods in which the ε-system is without explosions for all small ε. Note that the pointwise statement $p^0 < p_{\max}$ has to be interpreted in an appropriate way for the L^∞-function p^0. Regularity properties of p^ε in regions without explosions will be exploited in section 4.

We start with a crucial observation: if the average pressure of the ε-system is below $p_{\max}$ in a small area, then there cannot be explosions.

LEMMA 3.1 (quantitative Hopf lemma/near field effect). *Let $q : [-1,1]\times[-1,0] \to \mathbb{R}$ be a harmonic function, periodic in the first variable, and continuous up to the boundary, with $0 \le q \le p_{\max}$. Let $x \in [-1,1] \times \{0\} \equiv [-1,1]_{\mathrm{per}}$ be a point with $q(x) = p_{\max}$ and*

$$\frac{1}{2\delta}\int_{x-\delta}^{x+\delta} q(\xi)\, d\xi = \rho < p_{\max}.$$

Then the Neumann derivative in x has a lower bound

$$\partial_2 q(x) \ge c_H(\delta,\rho)$$

with $c_H(\delta,\rho) \to +\infty$ for fixed $\rho < p_{\max}$ and $\delta \to 0$.

COROLLARY 3.2. *There exists $\delta_0 > 0$ depending only on $\|V_0\|_{L^\infty}$, $p_{\max}$, and ρ, such that for small ε every solution p^ε of (1.2)–(1.6) with*

$$p^\varepsilon_\delta(x,t) := \frac{1}{2\delta}\int_{x-\delta}^{x+\delta} p^\varepsilon(\xi,t)\, d\xi \le \rho \qquad \forall t \in (t_1,t_2)$$

satisfies the following for $\delta < \delta_0$: p^ε has no explosion in $(x-\delta, x+\delta)\times(t_1,t_2)$.

Proof. In order to have an explosion in (x,t) we must have $\lim_{\tau \nearrow t} p^\varepsilon(x,\tau) = p_{\max}$ and $\lim_{\tau\nearrow t}(-\partial_2 p^\varepsilon(x,t)) \ge 0$. If we assume that $p^\varepsilon \le p_{\max}$ holds in the whole domain, Lemma 3.1 yields that $-\partial_2 p^\varepsilon$ is negative for small δ, and no explosion is possible. In the general case we decompose p^ε into one part p_A with the boundary values p^ε on Γ_1^ε and $\partial_2 p_A = 0$ on Γ_0, and a remainder p_B with vanishing values on Γ_1^ε and $-\partial_2 p_B = V_0$ on Γ_0, both with $\partial_2 p_{A,B} = 0$ on Γ_2^ε. Then $|\partial_2 p_B|$ is uniformly bounded and p_A is bounded by $p_{\max}$, so Lemma 3.1 applies to p_A. We conclude that $\partial_2 p_A(x,t)$ is large and that no explosion is possible. ∎

Proof of Lemma 3.1. We can assume $(x,t) = (0,0)$. The lemma follows from an argument that is related to rearrangement. For given ρ and δ, the Neumann derivative is minimal if $q = p_{\max}$ in a neighborhood $(-s,s)$ of x and $q = 0$ in the region $s < |x| < \delta$, where s is chosen such that $2sp_{\max} = \int_{-\delta}^{\delta} q = 2\delta\rho$. This can be seen as follows. The quantity $\partial_2 q(0,0)$ is decreased if we modify the Dirichlet boundary values on $(-\delta,\delta)$ by adding a nonnegative multiple of the function $v : (-1,1) \to \mathbb{R}$,

$$v(x) := \begin{cases} +1 & \text{for } |x| \in (x_1, x_1+r),\\ -1 & \text{for } |x| \in (x_2, x_2+r),\\ 0 & \text{else} \end{cases}$$

for $0 < x_1 < x_1 + r < x_2$ or $0 > x_1 > x_1 - r > x_2$. The harmonic extension $\bar v$ of v satisfies $\partial_2 \bar v(0) \le 0$, since $\bar v$ is nonnegative in a neighborhood of 0 by the monotonicity of the Green's function.

The above modifications allow a redistribution of the boundary values of q. It allows us to compare $\partial_2 q(0)$ with $\partial_2 q_\delta(0)$, where q_δ is the solution of the periodic problem

$$\begin{aligned} \Delta q &= 0 && \text{in } (-1,1)\times(-1,0),\\ q &= w && \text{on } (-1,1)\times\{0\},\\ q &= p_{\max} && \text{on } (-1,1)\times\{-1\}, \end{aligned}$$

with the boundary values

$$w(x) := \begin{cases} 0 & \text{for } s < |x| < \delta, \\ p_{\max} & \text{else.} \end{cases}$$

The number $s := \sigma\delta := \frac{\rho}{p_{\max}}\delta$ is determined by the integral condition $\int_{-\delta}^{\delta} w = 2\delta\rho$. From the Hopf lemma we know that $c_H(\delta,\rho) := \partial_2 q_\delta(0) > 0$.

In order to show $c_H(\delta,\rho) \to +\infty$ for $\delta \to 0$ it remains to consider the family of harmonic functions q_δ for $s = \sigma\delta$ and σ fixed. On the domain $\mathbb{R} \times \mathbb{R}_-$ we find the general solution q_δ by rescaling q_1: $q_\delta(x) = q_1(x/\delta)$. We calculate

$$\partial_2 q_\delta(0) = \frac{1}{\delta}\partial_2 q_1(0) \to \infty$$

for $\delta \to 0$. On the bounded domain the asymptotic behavior remains unchanged and we find the result. □

LEMMA 3.3. *Assume that the measure ν is of finite type. Let $0 < c < 1$ be an arbitrary number that we interpret as an explosion density. Then there is no subsequence $\varepsilon_k \to 0$ such that in the ε_k-systems there happen $\frac{c}{\varepsilon_k}$ explosions simultaneously.*

Proof. We present here the proof in the case that ν_K is supported only on the pattern P_0 of the single explosion. The proof in the general case requires only additional notational effort.

Step 1. We consider a sequence $\beta_N \in \{0,1\}^{\{-N,\dots,N\}}$ (patterns) with $|\beta_N| := \big|\{x \in \{-N,\dots,N\} \mid \beta_N(x) = 1\}\big| \geq c \cdot 2N$.

Claim. There exist a number $\rho > 0$, a distance $d \in \mathbb{N}$, and a subsequence $N \to \infty$ such that all β_N realize the distance d at least with density ρ, i.e.,

$$\big|\{x \in \{-N,\dots,N\} \mid \beta_N(x) = 1,\, \beta_N(x+d) = 1\}\big| \geq \rho \cdot 2N \quad \forall N.$$

We argue by contradiction and assume that the claim is not true. With d_0 applications we find that for every $\rho > 0$ and $d_0 \in \mathbb{N}$, there exists $N_0 > 0$ such that for all $N \geq N_0$ the distances $d = 1, 2, \dots, d_0$ are realized with density less than ρ. We calculate for large N the density of β_N. On $d_0\rho \cdot 2N$ points we have no restriction, on the remaining places we have at most $2N/(d_0+1)$ values 1. We calculate for the density

$$c \leq \frac{|\beta_N|}{2N} \leq \frac{d_0\rho \cdot 2N + 2N/(d_0+1)}{2N} = d_0\rho + \frac{1}{d_0+1}.$$

Since ρ and d_0 were arbitrary we arrive at a contradiction.

Step 2. We assume that for a subsequence $\varepsilon_k \to 0$ there are $\frac{c}{\varepsilon_k}$ simultaneous explosions. We claim that this contradicts the fact that no pattern of length 2 is contained in the limit measure ν.

We set $N = 1/\varepsilon$ and denote by t^ε the time instance of the simultaneous explosions. For $x \in \mathbb{Z}$ we set $\beta_N(x) = 1$ if at position $(\varepsilon x, t^\varepsilon)$ the ε-problem has an explosion, i.e., $p^\varepsilon(\varepsilon x, t^\varepsilon) = 0$, $\beta_N(x) = 0$ otherwise. By the above claim, for some $\rho > 0$, $d \in \mathbb{N}$, and a subsequence $N \to \infty$, β_N realizes the distance d of values 1 at least ρN times, say at positions $\{x_i\}$. Let $\bar{K}$ be a neighborhood of the single explosion P_0 in K, such that all patterns $k \in K$ with an explosion in $x_1 = d$ are not contained in $\bar{K}$. Then there are ρ/ε points $s_i = (\varepsilon x_i, t^\varepsilon)$ at which $R^\varepsilon_s p^\varepsilon$ is not contained in $\bar{K}$. Therefore

$$\nu^\varepsilon(S \times \bar{K}) \leq \nu^\varepsilon(S \times K) - \rho \inf\{s_0\}.$$

On the other hand, by assumption, we have

$$\nu^\varepsilon \to \mu \otimes \delta_{P_0} \in \mathcal{M}(S \times K),$$

which implies

$$\begin{aligned}\mu(S) &= \lim_{\varepsilon\to 0} \nu^\varepsilon(S \times \bar{K}) \le \lim_{\varepsilon\to 0} \nu^\varepsilon(S \times K) - \rho \inf\{s_0\} \\ &= \mu(S) - \rho \inf\{s_0\},\end{aligned}$$

a contradiction. □

Note that the result of the above lemma could be improved: not only can we not have $O(N)$ explosions at the same time instance, but it is also impossible to realize $O(N)$ explosions in a time span of length $O(\varepsilon)$. Even if in Proposition 2.6 we showed that in the limit patterns explosions happen simultaneously, this need not be true for the ε-problem. Nevertheless, the statement of Lemma 3.3 will be sufficient for our purposes.

We next prove an auxiliary result on the maximal gain of fluid-mass in a test volume in a short time.

LEMMA 3.4. *In every domain $W = B_\delta(x) \times (-1, 0)$, the maximal total inward flow through the lateral boundaries $\Sigma_\pm = \{x \pm \delta\} \times (-1, 0)$ can be estimated with arbitrary $\delta_0 > \delta$ by*

$$\begin{aligned}\int_{t_1}^{t_2} \int_\Sigma \partial_n p^\varepsilon \le{}& 2p_{\max} \frac{t_2 - t_1}{\delta_0 - \delta} \\ &+ 4\gamma(\delta_0 - \delta) \sup\{s_0\} + 2(t_2 - t_1)(\delta_0 - \delta)\|V_0\|_\infty.\end{aligned} \tag{3.1}$$

In particular, the bound can be chosen arbitrarily small for $t_2 - t_1$ small.

Proof. We do the calculations for the right boundary $\Sigma_{x+\delta} := \{x + \delta\} \times (-1, 0)$. The average total flow to the left between $x + \delta$ and $x + \delta_0$ is

$$\begin{aligned}\int_{t_1}^{t_2} \frac{1}{\delta_0 - \delta} \int_{x+\delta}^{x+\delta_0} \int_{-1}^{0} \partial_1 p^\varepsilon &\le \frac{1}{\delta_0 - \delta} \int_{t_1}^{t_2} \left(\int_{\Sigma_{x+\delta_0}} p^\varepsilon - \int_{\Sigma_{x+\delta}} p^\varepsilon \right) \\ &\le p_{\max} \frac{t_2 - t_1}{\delta_0 - \delta}.\end{aligned}$$

Then there is an intermediate value $z \in (x + \delta, x + \delta_0)$ where the average total flow is realized,

$$\int_{\Sigma_z} \int_{t_1}^{t_2} \partial_1 p^\varepsilon \le p_{\max} \frac{t_2 - t_1}{\delta_0 - \delta}.$$

By incompressibility, the maximal total flow through $\Sigma_{x+\delta}$ is the sum of two contributions: (1) the flow through Σ_z in the time interval (t_1, t_2), (2) the total inward flow through the upper boundary $(x+\delta, z) \times \{0\}$ and the lower boundary $(x+\delta, z) \times \{-1\}$. The total volume that can be released on Γ between $x + \delta$ and z is bounded by $2\gamma(\delta_0 - \delta) \cdot \sup\{s_0\}$. Formula (3.1) follows, with the factor 2 we include the left boundary. □

If the limit measure ν is of finite type we expect that the typical distance between explosions is large compared to ε. The following proposition verifies and sharpens this statement. Let us imagine that at temporal distances $O(\sqrt{\varepsilon})$ there happen $1/\sqrt{\varepsilon}$

explosions. Then the spatio-temporal distance between two explosions is always large compared to ε. Nevertheless, $1/\varepsilon$ explosions happen in a given spatio-temporal region. With such a construction it is also possible to have $1/\varepsilon$ explosions in an arbitrarily short time span Δt. The next proposition excludes this possibility for our evolution equations.

PROPOSITION 3.5. *If the measure ν is of finite type, then the marginal $\mu(\Gamma, \cdot) \in \mathcal{M}([0,T])$ of the measure ν, $\mu(\Gamma, (t_1, t_2)) := \nu(\Gamma \times (t_1, t_2) \times K)$, contains no atoms,*

$$\mu(\Gamma \times \{t\}) = 0 \quad \forall t \in (0, T]. \tag{3.2}$$

Proof. We assume that the limit measure $\mu(\Gamma, \cdot)$ contains an atom. Then for some $\bar{t}$,

$$e := \liminf_{\Delta t \to 0} \mu(\Gamma \times (\bar{t} - \Delta t, \bar{t} + \Delta t)) > 0. \tag{3.3}$$

The interpretation is that along the subsequence $\varepsilon \to 0$ we find $O(e/\varepsilon)$ explosions in the time span $(\bar{t} - \Delta t, \bar{t} + \Delta t)$.

To arrive at a contradiction we first choose $\Delta\rho$ small compared to e. The smallness will be specified towards the end of our calculations. We now fix $\delta > 0$. The required smallness for δ depends only on the numbers e and $\Delta\rho$, and on the boundary data V_0. Note that choosing $\delta > 0$ small we find, by Corollary 3.2,

$$\begin{aligned} &p_\delta^\varepsilon(x,t) < p_{\max} - \Delta\rho \quad \forall t \in (t_1, t_2) \\ &\quad \Rightarrow \text{no explosions happen in } B_\delta(x) \times (t_1, t_2). \end{aligned} \tag{3.4}$$

Given δ, we find a position $x \in [-1, 1]$ such that

$$\limsup_{\Delta t \to 0} \mu(B_\delta(x) \times (\bar{t} - \Delta t, \bar{t} + \Delta t)) \geq e \cdot \delta.$$

Using periodicity of the domain we can assume $x \in (-1 + \delta, 1 - \delta)$. We now fix Δt small, such that $\mu(B_\delta(x) \times (\bar{t} - \Delta t, \bar{t} + \Delta t)) \geq \frac{e\delta}{2}$, and such that additionally for a given constant $c_g > 0$ (depending on δ) there holds

$$\frac{e\delta}{16} \geq \|V_0\|_\infty \cdot 4\delta |\Delta t| + c_g \sqrt{\Delta t}.$$

We next choose ε small enough to have

$$\mu^\varepsilon(B_\delta(x) \times (\bar{t} - \Delta t, \bar{t} + \Delta t)) \geq \frac{1}{2} \mu(B_\delta(x) \times (\bar{t} - \Delta t, \bar{t} + \Delta t)),$$

and that there are no c/ε explosions at the same time instance (a small c is chosen in dependence of e and δ). The latter property is insured for small ε by Lemma 3.3. For this ε, we set $t_1 \in [\bar{t} - \Delta t, \bar{t} + \Delta t]$ to be the moment of the first explosion in $B_\delta(x)$, and set $t_2 = \bar{t} + \Delta t$. We introduce the time instance t^ε at which half of the explosions in $B_\delta(x) \times [t_1, t_2]$ have happened. At this point Lemma 3.3 guarantees that some explosions must happen after this time instance.

We can estimate the number of explosions of the ε-system in the test volume by

$$\begin{aligned} \mu^\varepsilon(B_\delta(x) \times (t_1, t^\varepsilon]) &\geq \frac{1}{2} \mu^\varepsilon(B_\delta(x) \times (\bar{t} - \Delta t, \bar{t} + \Delta t)) \\ &\geq \frac{1}{4} \mu(B_\delta(x) \times (\bar{t} - \Delta t, \bar{t} + \Delta t)) \geq \frac{e\delta}{8}. \end{aligned}$$

This means that in the ε-system at least $(e\delta)/(8\varepsilon \sup\{s_0\})$ explosions happen in $B_\delta(x) \times [t_1, t^\varepsilon]$. We will show that this implies that the average pressure is below $p_{\max}$ in t^ε and will conclude with Corollary 3.2. It remains to verify the implication

$$\text{loss of mass in explosions } \Rightarrow \text{ loss of pressure.}$$

In what follows we will use the estimate for the lateral inflow

$$\int_{t_1}^{t_2} \int_\Sigma \partial_n p^\varepsilon \le c_g \sqrt{\Delta t} \tag{3.5}$$

for small Δt, which follows from Lemma 3.4 if we choose $\delta_0 = \delta + \sqrt{t_2 - t_1}$. We calculate the gain of fluid mass in the ε-system by adding the inflow into the box from below and the loss due to explosions; in the following we consider only values of h^ε in $[0, s_0]$; i.e., in an explosion we set h^ε to zero. For all $t \in (t^\varepsilon, \bar{t} + \Delta t)$,

$$\begin{aligned}
&\int_{B_\delta(x) \cap \Gamma_{\bar{1}}^\varepsilon} \varepsilon^{-1} h^\varepsilon(\xi, \tau)\, d\xi \Big|_{\tau = t_1}^{t} \\
&\quad \le \|V_0\|_\infty \cdot 2\delta 2 |\Delta t| + c_g \sqrt{\Delta t} - \mu^\varepsilon(B_\delta(x) \times (t_1, t^\varepsilon)) \\
&\quad \le \|V_0\|_\infty \cdot 4\delta |\Delta t| + c_g \sqrt{\Delta t} - \frac{e\delta}{8} \le -\frac{e\delta}{16}.
\end{aligned}$$

We see that a decrease of fluid mass of order $O(1)$ took place in the test volume. We want to conclude from this that the average pressure also decreased by the order $O(1)$. We have to compare the effect of loss of mass with an effect that has the potential to increase the average pressure: redistribution of mass.

This effect is controlled in the following. We know that the average pressure at time t_1 satisfies $p_\delta^\varepsilon(x) \ge p_{\max} - \Delta\rho$, since an explosion happens at this time instance. Until time t the values of h^ε change, but although some of them might increase, we verify that this is not a large contribution. We sum over $x_i \in \varepsilon\mathbb{Z}$ with $x_i \in B_\delta(x)$,

$$\begin{aligned}
\frac{1}{2\delta} \sum_i (h^\varepsilon(x_i, t) - h^\varepsilon(x_i, t_1))_+ &\le \frac{1}{2\delta} \sum_i \left(\varepsilon \frac{p_{\max}}{a_0(i)} - h^\varepsilon(x_i, t_1) \right) \\
&= \frac{\varepsilon}{2\delta} \sum_i \frac{1}{a_0(i)} (p_{\max} - p^\varepsilon(x_i, t_1)) \\
&\le \frac{1}{\inf\{a_0\}} \frac{1}{2\delta} \frac{1}{2\gamma} \int_{B_\delta(x) \cap \Gamma_{\bar{1}}^\varepsilon} (p_{\max} - p^\varepsilon(\cdot, t_1)) \\
&\le \frac{1}{\inf\{a_0\}} \cdot \Delta\rho + o(1).
\end{aligned}$$

In the last line we used that in the limit $\varepsilon \to 0$ the two averages $\frac{1}{2\delta \cdot 2\gamma} \int_{B_\delta(x) \cap \Gamma_{\bar{1}}^\varepsilon} p$ and $\frac{1}{2\delta} \int_{B_\delta(x)} p$ coincide. Except for boundary effects, both expressions define linear and translation invariant averages of the values $p(x_i)$. The boundary effects vanish for $\varepsilon \to 0$ and the two expressions asymptotically coincide with the arithmetic mean of the values $p(x_i)$.

We can now calculate an upper bound for the average pressure at an arbitrary

time instance $t \in (t^\varepsilon, t_2]$,

$$\begin{aligned}\int_{B_\delta(x)} p^\varepsilon(\xi,t)\ d\xi &\le \int_{B_\delta(x)} p^\varepsilon(\xi,t_1)\ d\xi \\ &\quad + \sum_i (h^\varepsilon(x_i,t) - h^\varepsilon(x_i,t_1))_- \cdot \inf\{a_0\} \\ &\quad + \sum_i (h^\varepsilon(x_i,t) - h^\varepsilon(x_i,t_1))_+ \cdot \sup\{a_0\} + o(1) \\ &\le 2\delta p_{\max} + \sum_i (h^\varepsilon(x_i,t) - h^\varepsilon(x_i,t_1)) \cdot \inf\{a_0\} \\ &\quad + \sum_i (h^\varepsilon(x_i,t) - h^\varepsilon(x_i,t_1))_+ \cdot (\sup\{a_0\} - \inf\{a_0\}) + o(1) \\ &\le 2\delta p_{\max} - \frac{1}{2\gamma}\frac{e\delta}{16} \cdot \inf\{a_0\} + 2\delta \cdot C\Delta\rho + o(1)\end{aligned}$$

for $\varepsilon \to 0$. The corrector $o(1)$ takes into account that the p-average over $B_\delta(x)$ coincides only asymptotically with the p-average over $B_\delta(x) \cap \Gamma_1^\varepsilon$. Dividing by 2δ we find for small ε

$$\frac{1}{2\delta}\int_{B_\delta(x)} p^\varepsilon(\xi,t)\ d\xi \le p_{\max} - \Delta\rho \qquad \forall t \in (t^\varepsilon, t_2),$$

if $\Delta\rho$ was chosen small compared to e. We know that in the ball $B_\delta(x)$ there happen explosions in the time interval (t^ε, t_2). This is in contradiction with the fact that for our choice of δ, below the average pressure $p_{\max} - \Delta\rho$, there can be no explosions by Corollary 3.2. □

In case that ν is of finite type, the measure μ has a direct physical interpretation. For every set $\bar{S} = \Gamma \times (t_1, t_2)$ the number $\mu(\bar{S})$ is the weighted number of explosions in $\bar{S}$. If $s_0(k) = 1$ for all k, then

$$\mu(\bar{S}) = \nu(\bar{S} \times K) = \lim_{\varepsilon \to 0} \left(\varepsilon \cdot \#\{\text{explosions in } \bar{S}\}\right),$$

the limit taken along the chosen subsequence. In the general case μ measures the total mass of fluid that is lost in explosions.

Our next aim is to show the following relation between the measure of limit patterns and spatial averages of the limit pressure $p_\delta^0(x,t)$. Loosely speaking, we show

$$\nu \text{ of finite type } \Rightarrow\ p_\delta^0 \text{ has no jumps.}$$

In order to show this statement, by Proposition 3.5, it remains to verify that if $\mu(\Gamma, \cdot)$ contains no atoms, then p_δ^0 has no jumps. By definition, the functions p_δ^0 are Lipschitz continuous in x for every t. In order to state regularity properties in time, we choose as a representative of p_δ^0 the temporal maximal function,

$$p_\delta^0(x,t) = \limsup_{r \searrow 0} \frac{1}{2r}\int_{t-r}^{t+r} p_\delta^0(x,\tau)\ d\tau.$$

In the following we work with this representative of p_δ^0.

PROPOSITION 3.6. *Assume that the measure $\mu(\Gamma,\cdot)$ contains no atoms. If in a point $(x,\bar t\,)$ the average pressure is not maximal,*

$$p^0_\delta(x,\bar t\,)=\rho<p_{\max}, \tag{3.6}$$

then there exist $\rho_0<p_{\max}$, $\varepsilon_0>0$, and $t_1<\bar t<t_2$ with

$$p^\varepsilon_\delta(x,t)\le\rho_0 \quad \forall t\in(t_1,t_2),\ \varepsilon<\varepsilon_0, \tag{3.7}$$

$$p^0_\delta(x,t)\le\rho_0 \quad \forall t\in(t_1,t_2). \tag{3.8}$$

The number ρ_0 does not depend on δ.

Proof. Our emphasis in this proposition lies on finding $t_1<\bar t$; in this part the assumption on μ is used. The proof has similarity with the proof of Proposition 3.5, but this time we use the converse implication

$$\text{loss of pressure } \Rightarrow \text{ loss of mass in explosions.}$$

We choose $\Delta\rho$ small compared to $p_{\max}-\rho$. Let us assume that for a subsequence $\varepsilon\to0$ and a sequence $t_1^\varepsilon\nearrow\bar t$ there holds

$$p^\varepsilon_\delta(x,t_1^\varepsilon)\ge p_{\max}-\Delta\rho.$$

From (3.6) we conclude that there exists $t^\varepsilon>t_1^\varepsilon$ arbitrarily close to $\bar t$ with

$$p^\varepsilon_\delta(x,t^\varepsilon)\le\rho+\Delta\rho.$$

This follows from the fact that spatio-temporal averages of p^ε converge to the corresponding averages of p^0. Exploiting that the pressure in t_1^ε is large, we verify that redistribution of mass between the cells is a small effect,

$$\begin{aligned}\frac{1}{2\delta}\sum_i(h^\varepsilon(x_i,t^\varepsilon)-h^\varepsilon(x_i,t_1^\varepsilon))_+ &\le \frac{1}{2\delta}\sum_i\left(\varepsilon\frac{p_{\max}}{a_0(i)}-h^\varepsilon(x_i,t_1^\varepsilon)\right)\\ &=\frac{\varepsilon}{2\delta}\sum_i\frac{1}{a_0(i)}(p_{\max}-p^\varepsilon(x_i,t_1^\varepsilon))\le c_h.\end{aligned}$$

For small ε the constant c_h can be chosen as $c_h=C\Delta\rho$ with C independent of δ. Our next aim is to conclude that the average height is decreased. We calculate

$$\begin{aligned}\rho-p_{\max}+2\Delta\rho &\ge \frac{1}{2\delta}\int_{B_\delta(x)}p^\varepsilon(\xi,\tau)\,d\xi\Big|_{t_1^\varepsilon}^{t^\varepsilon}\\ &\ge\frac{1}{2\delta 2\gamma}\int_{B_\delta(x)\cap\Gamma_1^\varepsilon}p^\varepsilon(\xi,\tau)\,d\xi\Big|_{t_1^\varepsilon}^{t^\varepsilon}+o(1)\\ &\ge\frac{1}{2\delta}\sum_i(h^\varepsilon(x_i,t^\varepsilon)-h^\varepsilon(x_i,t_1^\varepsilon))_+\cdot\inf\{a_0\}\\ &\quad+\frac{1}{2\delta}\sum_i(h^\varepsilon(x_i,t^\varepsilon)-h^\varepsilon(x_i,t_1^\varepsilon))_-\cdot\sup\{a_0\}+o(1)\\ &\ge c_h\cdot(\inf\{a_0\}-\sup\{a_0\})+o(1)\\ &\quad+\frac{1}{2\delta}\sum_i(h^\varepsilon(x_i,t^\varepsilon)-h^\varepsilon(x_i,t_1^\varepsilon))\cdot\sup\{a_0\}.\end{aligned}$$

For small ε and small $\Delta\rho$ we conclude for the change in the average height

$$\frac{1}{2\delta}\sum_i (h^\varepsilon(x_i,t^\varepsilon) - h^\varepsilon(x_i,t_1^\varepsilon))\cdot \sup\{a_0\} \le \rho - p_{\max} + (3+C)\Delta\rho < 0,$$

an order $O(1)$ loss of volume. As in the preceding proposition, for $t^\varepsilon - t_1^\varepsilon$ small enough, this cannot be induced by outflow on lateral or lower sides. Therefore, there are $O(1/\varepsilon)$ explosions and we have $\mu^\varepsilon(\Gamma\times(t_1,t_2)) \ge c > 0$ for all $t_2 > \bar t$ and for a subsequence $\varepsilon \to 0$. In the limit we find $\mu(\Gamma\times(t_1,t_2)) \ge c$; since t_1 and t_2 are arbitrary we found an atom of $\mu(\Gamma,\cdot)$ and thus a contradiction.

The construction of t_2 follows the same pattern. We calculate that an increase in pressure requires an increase in volume of order $O(1)$. This cannot be compensated by lateral inflow, inflow from below, or redistribution effects. Then there must be a gain of volume through the upper boundary Γ—a contradiction since no "negative explosions" are possible.

Property (3.8) follows from (3.7). Note that at first we find

$$p_\delta^0(x,t) \le \rho_0 \qquad \text{a.e. } t\in(t_1,t_2).$$

By the choice of the representative p_δ^0 we conclude that the inequality holds for all t. ∎

COROLLARY 3.7. *Assume that ν is of finite type. Let $s=(x,\bar t)\in\Gamma\times(0,T)$ be a point with*

$$\limsup_{\delta\to 0} p_\delta^0(s) = \rho < p_{\max}.$$

Then there exist $\delta_0>0$, $\varepsilon_0>0$, and $\Delta t>0$ such that for all $\varepsilon<\varepsilon_0$ there are no explosions in $B_{\delta_0}(x)\times(\bar t-\Delta t,\bar t+\Delta t)$, i.e.,

$$\mu^\varepsilon((x-\delta_0,x+\delta_0)\times(\bar t-\Delta t,\bar t+\Delta t)) = 0.$$

Proof. The result follows from inequality (3.7) using Corollary 3.2. Note that we have to pick a small δ in dependence of $\rho_0 < p_{\max}$; here we use that ρ_0 in Proposition 3.6 does not depend on δ. The conclusion remains valid if lim sup is replaced by lim inf in the assumption. ∎

4. Upscaled equations. Our aim is to derive the physical laws for the averaged pressure. On the boundary we expect a law relating the increase of pressure with the parameters pressure and normal velocity, and we write

$$p^0 < p_{\max} \Rightarrow \partial_t p^0 = \alpha(p^0, -\partial_2 p^0),$$
$$p^0 = p_{\max} \Rightarrow \partial_t p^0 = \alpha(p^0, -\partial_2 p^0)_-$$

for some function α. The first implication expresses that, as long as the maximally sustained pressure is not yet achieved, there is an increase of the pressure according to the local rules of filling pores. A flow towards the boundary increases the filling height of the single pore and, due to the monotonicity of the law $\mathcal{P}_0$, the pressure will increase. The second implication describes the situation once the maximally sustained pressure is achieved. A backward flow lowers the pressure according to the averaged law. A further flow towards the boundary results in explosions and cannot increase the pressure.

For the mathematical interpretation of the above equations some care must be applied. For $p^0 \in L^\infty$ we will interpret the condition $p^0 < p_{\max}$ as

$$\limsup_{\delta \to 0} p^0_\delta(x,t) < p_{\max}.$$

On the right-hand side of the equations $\partial_2 p^0|_\Gamma$ has a meaning as a distribution. But, for the second implication, it is not clear how to take the negative part of this distribution in some parts of the boundary, the full expression in others. We should show the following relations along Γ:

$$p^0 \le p_{\max} \quad \text{a.e. on } \Gamma, \tag{4.1}$$

$$p^0(x,t) < p_{\max} \Rightarrow \partial_t p^0 = \alpha(p^0, -\partial_2 p^0), \tag{4.2}$$

$$\partial_t p^0 \le \alpha(p^0, -\partial_2 p^0) \quad \text{as distributions on } \Gamma. \tag{4.3}$$

The first inequality is immediate, since every p^ε satisfies the inequality. The evolution equation is (4.2) and it is interpreted in the sense of distributions in a neighborhood of (x,t). Inequality (4.3) is a lift-off condition. In the case of linear laws it can be shown just as (4.2). In the nonlinear case the proof is more involved, since in the situation of (4.3) the solution has less regularity properties than in the situation of (4.2).

Our aim is to homogenize the law $p^\varepsilon(\varepsilon i) = \mathcal{P}_0(i, \varepsilon^{-1} h^\varepsilon(\varepsilon i))$. The function $\mathcal{P}_0$ depends in an oscillatory fashion on x, and the function h^ε will in general also have an oscillatory behavior. Therefore a homogenization limit has to be performed. The key in the proofs is to assure regularity properties of p^ε. We want to analyze not only linear laws as in (1.7), but also more general nonlinear models.

DEFINITION 4.1. *We speak of a nonlinear model if the laws $\mathcal{P}_0(i,\cdot)$ are nonlinear $s_0(i)$-periodic functions with $\max \mathcal{P}_0(i,\cdot) = \mathcal{P}_0(i, s_0(i)) = p_{\max}$ and $0 < a_1 \le \mathcal{P}_0'(i,\cdot) \le a_2 < \infty$ for all i, $s \in (0, s_0(i))$. We say that the nonlinear model satisfies the linear regularity properties if the statement of Lemma 4.2 for the linear law holds also for the laws $\mathcal{P}_0(i,\cdot)$.*

An example of a nonlinear model that satisfies the linear regularity is given by the choice $\mathcal{P}_0(i,\cdot) = \mathcal{P}_0(s)$ with some strictly monotonically increasing function $\mathcal{P}_0 : [0, s_0) \to \mathbb{R}$.

We modify the function p^0 on a set of vanishing measure such that

$$p^0(x,t) = \limsup_{\delta \to 0} p^0_\delta(x,t).$$

LEMMA 4.2. *We consider linear laws $\mathcal{P}_0(i,\cdot)$ and we assume that the box $(x-\delta_0, x+\delta_0) \times (t-\Delta t, t+\Delta t)$ contains no explosions. Then there is a neighborhood $U \subset \Gamma \times \mathbb{R}$ of (x,t) in which the pressure $p^\varepsilon(\cdot,t)$ is continuous with modulus of continuity independent of ε and t.*

For every $\Delta\rho > 0$ there exist $\delta, \varepsilon_0 > 0$ such that for all $(x_1,\tau), (x_2,\tau) \in U$,

$$|p^\varepsilon(x_1,\tau) - p^\varepsilon(x_2,\tau)| \le \Delta\rho \qquad \forall |x_1 - x_2| < \delta,\ \varepsilon < \varepsilon_0, \tag{4.4}$$

$$|p^0(x_1,\tau) - p^0(x_2,\tau)| \le \Delta\rho \qquad \forall |x_1 - x_2| < \delta, \tag{4.5}$$

$$|p^\varepsilon(\xi,\tau) - p^\varepsilon_{\delta'}(x_1,\tau)| \le \Delta\rho \qquad \forall \xi \in B_{\delta'}(x_1),\ \varepsilon < \varepsilon_0,\ 0 < \delta' < \delta. \tag{4.6}$$

In the nonlinear case, (4.4) *and* (4.6) *hold for all τ except for an exceptional set of arbitrary small measure which can be prescribed together with $\Delta\rho$.*

Proof. In Proposition A.3 we show (4.4). It implies that local averages of p^0 satisfy the same inequality and we can conclude (4.5) by the theorem of Lebesgue. Inequality (4.6) is a direct consequence of (4.4).

In the nonlinear case, given $\Delta\rho$, we choose first κ and ε_0 such that the errors introduced by p_B and $p_{A,2}$ are small compared to $\Delta\rho$. By the uniform continuity of $p_{A,1}$ for most of the time we can choose δ small in order to satisfy (4.4). ∎

Remark 4.3 (partial continuity of p^0). Assume finiteness of ν and linearity of the laws $\mathcal{P}_0$. Let $s_0 = (x_0, t_0)$ be a boundary point with $p^0(s_0) < p_{\max}$. Then in a neighborhood U of s_0 the function p^0 is continuous in (x, t). Everywhere holds the equality

$$p^0(x,t) = \lim_{\delta\to 0} p^0_\delta(x,t). \tag{4.7}$$

Proof. By definition of the representative p^0 there holds

$$\limsup_{\delta\to 0} p^0_\delta(s_0) < p_{\max}.$$

Then Corollary 3.7 yields the existence of a neighborhood without explosions. Note that this holds also in points with $\liminf_{\delta\to 0} p^0_\delta(s_0) < p_{\max}$. Lemma 4.2 yields the existence of a smaller neighborhood U of (x_0, t_0) such that p^0 is uniformly continuous in x for a.e. t, with modulus of continuity independent of t. Furthermore Proposition A.3 yields uniform estimates for $\partial_t p^0_A|_\Gamma \in L^2$ in a space-time neighborhood of (x_0, t_0). They imply that

$$p^0_\delta(x,t) \quad \text{is continuous in } t.$$

We conclude that p^0 is continuous in (x, t).

We can now conclude (4.7). In points s with $\liminf_{\delta\to 0} p^0_\delta(s) < p_{\max}$ it is a consequence of the continuity of p^0. In the other case we have

$$p_{\max} = \liminf_{\delta\to 0} p^0_\delta(s) \le \limsup_{\delta\to 0} p^0_\delta(s) \le p_{\max}.$$

This implies again (4.7). ∎

With the above regularity properties of p^ε and p^0 we can now homogenize the law $\mathcal{P}_0$. As a model we have chosen a uniform law with $s_0(i) = s_0$ and $a_0(i) = a_0$ independent of the position i. In this case the expression (4.8) can be trivially calculated and equals $1/a_0$. We use the general expressions below in order to include stochastic and nonlinear models.

ASSUMPTION 4.4 (ergodicity). *Consider for $\rho \in (0, p_{\max})$ the expression*

$$\lim_{\varepsilon\to 0} \frac{\varepsilon}{2\delta} \sum_i \frac{1}{\mathcal{P}_0'(i, s_i)}, \tag{4.8}$$

where s_i are the unique solutions of $\mathcal{P}_0(i, s_i) = \rho$. The sum is taken over all indices i with $\varepsilon i \in B_\delta(x)$.

We assume on the function $\mathcal{P}_0$ that the above limit exists for all x and δ and that it is independent of x and δ.

An example of an ergodic material is a k_0-periodic function $\mathcal{P}_0$.

PROPOSITION 4.5 (a law for p^0 in regions without explosions). *Let the ergodicity assumption, Assumption 4.4, be satisfied and $R = (x - \delta_0, x + \delta_0) \times (t - \Delta t, t + \Delta t)$*

be a region without explosions. Then in R there holds in the sense of distributions in (x,t)

$$\partial_t \left[\Theta(p^0(x,t))\right] = -\partial_2 p^0(x,t), \tag{4.9}$$

where the function Θ satisfies

$$\Theta'(\rho) = \lim_{\varepsilon\to 0} 2\gamma \frac{\varepsilon}{2\delta} \sum_i \frac{1}{\mathcal{P}_0'(i,s_i)} \quad \forall \rho \in (0, p_{\max}). \tag{4.10}$$

On the right-hand side appears the expression of Assumption 4.4.

We emphasize that the above proposition holds also in the case of a nonlinear model without additional regularity assumptions.

Proof. We start the proof from the differentiated version of the microscopic pressure law $p^\varepsilon = \mathcal{P}_0(\varepsilon^{-1}h^\varepsilon)$ in a point $x_i = \varepsilon i$,

$$\partial_t p^\varepsilon(x_i,t) = \mathcal{P}_0'(i, \varepsilon^{-1}h^\varepsilon(x_i,t)) \cdot \varepsilon^{-1}\partial_t h^\varepsilon(x_i,t). \tag{4.11}$$

We have a one-to-one correspondence between pressure p^ε and height h^ε in every point x_i,

$$p^\varepsilon(x_i) = \mathcal{P}_0(i, \varepsilon^{-1}h^\varepsilon(x_i)) \quad \text{or} \quad h^\varepsilon(x_i) = \varepsilon H_0(i, p^\varepsilon(x_i)).$$

We introduce the functions Φ_i satisfying $\Phi_i(0) = 0$ and

$$\Phi_i'(\rho) = \frac{1}{\mathcal{P}_0'(i, H_0(i,\rho))}.$$

We now divide (4.11) by $\mathcal{P}_0'$ and, using (1.2), write the equation as

$$\frac{d}{dt}\Phi_i(p^\varepsilon(x_i,t)) = \varepsilon^{-1}\partial_t h^\varepsilon(x_i,t) = \frac{1}{\varepsilon 2\gamma}\int_{x_i-\gamma\varepsilon}^{x_i+\gamma\varepsilon} (-\partial_2 p^\varepsilon(\xi,t))\, d\xi.$$

Since we do not have knowledge on limits of time derivatives, we use the time integrated form. We additionally have to average over the spatial variable and use therefore the following time and space integrated equation:

$$\begin{aligned} &\frac{1}{\Delta t}\frac{\varepsilon}{2\delta}\sum_i [\Phi_i(p^\varepsilon(x_i, t+\Delta t)) - \Phi_i(p^\varepsilon(x_i,t))] \\ &\quad = \frac{1}{\Delta t}\frac{1}{2\gamma}\int_t^{t+\Delta t} (-\partial_2 p_\delta^\varepsilon(x,\tau))\, d\tau. \end{aligned} \tag{4.12}$$

We used here that taking x_1-spatial averages and the operator ∂_2 can be interchanged. The right-hand side converges for $\varepsilon \to 0$ as a distribution,

$$\frac{1}{\Delta t}\int_t^{t+\Delta t} (-\partial_2 p_\delta^\varepsilon)\, d\tau \to \frac{1}{\Delta t}\int_t^{t+\Delta t} (-\partial_2 p_\delta^0)\, d\tau.$$

Here the convergence of the integrand is interpreted as

$$\begin{aligned} &\int_\Gamma (-\partial_2 p_\delta^\varepsilon(x,\tau)) \cdot \varphi(x)\, dx \\ &\quad := -\int_\Gamma p_\delta^\varepsilon(x,\tau)\cdot \partial_2\varphi(x)\, dx + \int_\Omega p_\delta^\varepsilon(x,y,\tau)\cdot \Delta\varphi(x,y)\, dx\, dy \\ &\quad \to -\int_\Gamma p_\delta^0(x,\tau)\cdot \partial_2\varphi(x)\, dx + \int_\Omega p_\delta^0(x,y,\tau)\cdot \Delta\varphi(x,y)\, dx\, dy \\ &\quad =: \int_\Gamma (-\partial_2 p_\delta^0(x,\tau))\cdot\varphi(x)\, dx \end{aligned}$$

for all periodic $\varphi \in C^2(\Omega)$ with compact support in $\Omega \cup \Gamma$. In the convergence we used that $p^\varepsilon|_\Gamma \to p^0|_\Gamma$ in L^∞_w, and accordingly the convergence of $p^\varepsilon_\delta|_\Gamma$.

We next consider the left-hand side of (4.12) and its limit as $\varepsilon \to 0$. We choose the function $\Theta(\rho)$ of (4.10) as the average

$$\Theta(\rho) := \lim_{\varepsilon\to 0} 2\gamma \frac{\varepsilon}{2\delta} \sum_i \Phi_i(\rho).$$

Since averages of Φ_i' exist, averages of Φ_i also exist. We now have to use the fact that p^ε has no oscillations in x for most values of t. We pick a small $\Delta\rho > 0$ and choose $\delta_0 > 0$ small to satisfy

$$|p^\varepsilon(x_i,t) - p^\varepsilon_\delta(x,t)| \le \Delta\rho \qquad \forall x_i = \varepsilon i \in B_\delta(x),\ \delta < \delta_0,\ t \in \mathcal{T}_{\delta_0}.$$

Here we use Lemma 4.2. In the linear case we can choose $\mathcal{T}_{\delta_0} = (t_1, t_2)$; in the nonlinear case $\mathcal{T}_{\delta_0}$ is an (ε-dependent) subset of (t_1, t_2). For small $\delta_0 > 0$ the measure $|\mathcal{T}_{\delta_0}|$ is arbitrarily close to $|t_2 - t_1|$.

We next exploit that the averages p^ε_δ are uniformly continuous (Proposition A.3), and that we can choose a subsequence $\varepsilon \to 0$ such that $p^\varepsilon_\delta \to p^0_\delta$ uniformly in R. We use this to write

$$|p^\varepsilon(x_i,t) - p^0_\delta(x,t)| \le \Delta\rho + o(1) \qquad \forall i,\ \forall t \in \mathcal{T}_{\delta_0}.$$

With our knowledge on oscillations of p^ε we can now use

$$\begin{aligned} |\Phi_i(p^\varepsilon(x_i,t)) - \Phi_i(p^0_\delta(x,t))| &\le \sup_i \|\Phi_i'\|_\infty \cdot |p^\varepsilon(x_i,t) - p^0_\delta(x,t)| \\ &\le \frac{1}{a_1}(\Delta\rho + o(1)) \qquad \forall i,\ \forall t \in \mathcal{T}_{\delta_0} \end{aligned}$$

to perform the replacement

$$\frac{\varepsilon}{2\delta}\sum_i \Phi_i(p^\varepsilon(x_i,t)) = \frac{\varepsilon}{2\delta}\sum_i \Phi_i(p^0_\delta(x,t)) + O(\Delta\rho) + o(1) \quad \forall t \in \mathcal{T}_{\delta_0}.$$

In what follows we have to consider the expressions as distributions in time and use a test function $\phi(t)$ with compact support in (t_1, t_2). We conclude

$$\begin{aligned} &\int_{t_1}^{t_2} \frac{\varepsilon}{2\delta}\sum_i \Phi_i(p^\varepsilon(x_i,t))\phi(t)\,dt \\ &\quad= \int_{t_1}^{t_2} \frac{\varepsilon}{2\delta}\sum_i \Phi_i(p^0_\delta(x,t))\phi(t)\,dt + O(\Delta\rho) + o(1) + o_\delta(1) \\ &\quad\to \frac{1}{2\gamma}\int_{t_1}^{t_2} \Theta(p^0_\delta(x,t))\phi(t)\,dt + O(\Delta\rho) + o_\delta(1), \end{aligned}$$

with $o_\delta(1) \to 0$ for $\delta_0 \to 0$, since averages of Φ_i are bounded and $\mathcal{T}_{\delta_0}$ has large measure. In taking the limit $\varepsilon \to 0$ we used the ergodicity assumption.

We write (4.12) now with a discrete integration by parts,

$$\begin{aligned} &-\int_{t_1}^{t_2} \Theta(p^0_\delta(x,t))\frac{\phi(t) - \phi(t-\Delta t)}{\Delta t}\,dt \\ &\quad= -\int_{t_1}^{t_2} \partial_2 p^0_\delta(x,t)\ \frac{1}{\Delta t}\int_t^{t+\Delta t} \phi(\tau)\,d\tau\,dt + \frac{2}{\Delta t}O(\Delta\rho) + o_\delta(1). \end{aligned}$$

We take the limit $\delta \to 0$ using that along a subsequence $\delta \to 0$ the functions p^0_δ converge to p^0 pointwise a.e. Since now δ_0 can also be chosen small, we find in the sense of distributions in x

$$-\int_{t_1}^{t_2} \Theta(p^0(\cdot,t)) \frac{\phi(t)-\phi(t-\Delta t)}{\Delta t}\,dt = -\int_{t_1}^{t_2} \partial_2 p^0(\cdot,t)\ \frac{1}{\Delta t}\int_t^{t+\Delta t} \phi(\tau)\,d\tau\,dt + \frac{2}{\Delta t}O(\Delta\rho).$$

Since $\Delta\rho$ was arbitrary, the equation holds also without the error term. Since Δt also was arbitrary, we find the result. □

In the following theorem we collect all the upscaled equations. The principal assumption of the theorem is that ν is of finite type. Thinking of periodic solutions of period $k\varepsilon$ in x_1, we know that such an assumption is necessary. The assumption can be replaced by "with probability 1" in the case of stochastic equations.

THEOREM 4.6. *We consider a subsequence p^ε of solutions to* (1.2)–(1.6) *with a limit measure ν of finite type. Let p^0 be the limit of p^ε in L^∞_w and in the weak topology of $L^2((0,T),H^1(\Omega))$. Then there exists a representative p^0 that is harmonic for all t and which satisfies*

$$0 \le p^0(x,t) \le p_{\max} \qquad \forall x \in \Gamma,\ \forall t. \tag{4.13}$$

Every point $(x,t) \in \Gamma \times (0,T)$ with

$$p^0(x,t) < p_{\max}$$

has a neighborhood in $\Gamma \times (0,T)$ on which

$$\partial_t \Theta(p^0) = -\partial_2 p^0 \tag{4.14}$$

holds in the sense of distributions. Everywhere on $\Gamma \times (0,T)$ holds the corresponding inequality

$$\partial_t \Theta(p^0) \le -\partial_2 p^0. \tag{4.15}$$

In the nonlinear case the same properties hold. To conclude (4.15) *we have to assume that the nonlinear equations satisfy the linear regularity property of Definition* 4.1. *Without this assumption we only have the weak lift-off condition* (4.22).

We make some remarks on this theorem.

Boundary values on Γ. The formal definition of the limit function is as follows. We choose a subsequence $\varepsilon \to 0$ such that $p^\varepsilon|_\Gamma$ converge in L^∞_w to some limit $p^0_\Gamma \in L^\infty(\Gamma)$. The weak limit $p^0 \in L^2((0,T),H^1(\Omega))$ satisfies

$$\int_\Gamma p^0_\Gamma \cdot \partial_2 \varphi = \int_\Omega p^0 \cdot \Delta\varphi \qquad \text{a.e. } t \in (0,T)$$

for all $\varphi \in C^2_0(\Omega \cup \Gamma)$ with $\varphi = 0$ on Γ. This shows that p^0 is a harmonic function with boundary values p^0_Γ.

Initial values. If $p^\varepsilon(t=0) = P_0$ is continuous and satisfies $P_0|_\Gamma < p_{\max}$, then by the regularity results all functions p^0_δ are continuous in a neighborhood and have the initial values $(P_0)_\delta$. Therefore $p^0(t=0) = P_0$.

The lift-off condition (4.15). In all points $p^0 < p_{\max}$, (4.14) describes the evolution. Without a lift-off condition p^0 could remain on the level $p_{\max}$ even if the fluid flows backward into the domain. Therefore a condition of lift-off is necessary in order to close the system. We emphasize that weaker conditions may be sufficient; relevant is that the left-hand side in (4.15) is negative in regions where the right-hand side is negative.

In the derivation of (4.15) we face the problem that a pointwise analysis is necessary. Loosely speaking, in some points we have to argue with the help of regularity of p^ε in order to find the law, in other points we use $p^\varepsilon \geq p_{\max} - \Delta\rho$ in order to find the law. Such pointwise analysis is in conflict with the use of distributional limits as in the proof of Proposition 4.5.

The analysis proceeds in three steps. In section 2.2 we characterized the possible patterns of the system. In section 3 we showed that, if all limit patterns are finite, averages of the pressure cannot have jumps (Proposition 3.6) and that every point with nonmaximal limit pressure has a neighborhood without explosions (Corollary 3.7). Based on these observations we derive the upscaled equations.

Proof. All assertions of the theorem are already shown except for (4.15). In the case of linear laws it can easily be derived following the lines of Proposition 4.5, starting from the inequality in (4.12). The point is that for linear laws the information that p^ε is close to p^ε_δ is not needed.

In the general case the subsequent proposition establishes with (4.16) a pointwise inequality for the pressure decay. It describes on a microscopic scale the condition of lift-off and is the key for the proof.

Using (4.16) in the first inequality and the Lebesgue convergence theorem in the second (the boundedness from below of the integrand is verified in Proposition 4.7), we calculate for nonnegative smooth test functions ϕ with compact support in $(\Omega \cup \Gamma) \times (0,T)$

$$\begin{aligned}
&\int_0^T \int_\Gamma \frac{\Theta(p^0(x,0,t+\Delta t)) - \Theta(p^0(x,0,t))}{\Delta t} \phi(x,t)\,dx\,dt \\
&\quad \leq \int_0^T \int_\Gamma \liminf_{\delta\to 0} \liminf_{y\nearrow 0} \frac{1}{\Delta t} \int_t^{t+\Delta t} (-\partial_2 p^0_\delta(x,y,\tau))\,d\tau \phi(x,t)\,dx\,dt \\
&\quad \leq \liminf_{\delta\to 0} \liminf_{y\nearrow 0} \int_0^T \int_\Gamma \frac{1}{\Delta t} \int_t^{t+\Delta t} (-\partial_2 p^0_\delta(x,y,\tau))\,d\tau \phi(x,t)\,dx\,dt \\
&\quad = \liminf_{\delta\to 0} \liminf_{y\nearrow 0} \int_0^T \left\{ \int_{\Omega\cap\{x_2<y\}} \frac{1}{\Delta t} \int_t^{t+\Delta t} p^0_\delta(\tau)\,d\tau \cdot \Delta\phi\,dx_1\,dx_2 \right. \\
&\qquad\qquad \left. - \int_{\Omega\cap\{x_2=y\}} \frac{1}{\Delta t} \int_t^{t+\Delta t} p^0_\delta(\tau)\,d\tau \cdot \partial_2\phi\,dx_1 \right\} dt \\
&\quad = \liminf_{\delta\to 0} \int_0^T \left\{ \int_\Omega \frac{1}{\Delta t} \int_t^{t+\Delta t} p^0_\delta(\tau)\,d\tau \cdot \Delta\phi\,dx_1\,dx_2 \right. \\
&\qquad\qquad \left. - \int_\Gamma \frac{1}{\Delta t} \int_t^{t+\Delta t} p^0_\delta(\tau)\,d\tau \cdot \partial_2\phi\,dx_1 \right\} dt
\end{aligned}$$

$$= \int_0^T \left\{ \int_\Omega \frac{1}{\Delta t} \int_t^{t+\Delta t} p^0(\tau)\ d\tau \cdot \Delta\phi\, dx_1\, dx_2 \right.$$
$$\left. - \int_\Gamma \frac{1}{\Delta t} \int_t^{t+\Delta t} p^0(\tau)\, d\tau \cdot \partial_2\phi\, dx_1 \right\} dt$$
$$= \int_0^T \left\langle \frac{1}{\Delta t} \int_t^{t+\Delta t} (-\partial_2 p^0(\tau))\, d\tau, \phi \right\rangle dt.$$

Since Δt was arbitrary, this proves the claim in the case of linear regularity. The general case is treated in Corollary 4.8. □

PROPOSITION 4.7 (pointwise lift-off condition). *Let $\Delta t > 0$ be given. We assume that the equations satisfy the linear regularity. Then for a.e. point $(x, 0, t_0)$ there holds with*

$$V := \liminf_{\delta\to 0} \liminf_{y\nearrow 0} \frac{1}{\Delta t} \int_{t_0}^{t_0+\Delta t} \left(-\partial_2 p^0_\delta(x, y, t)\right) dt$$

the inequality

$$\frac{\Theta(p^0(x, 0, t_0 + \Delta t)) - \Theta(p^0(x, 0, t_0))}{\Delta t} \le V. \tag{4.16}$$

The expression $\liminf_{y\nearrow 0} \frac{1}{\Delta t} \int_{t_0}^{t_0+\Delta t} (-\partial_2 p^0_\delta(x, y, t))\, dt$ is bounded from below independent of δ and (x, t_0), and $\frac{1}{\Delta t} \int_{t_0}^{t_0+\Delta t} (-\partial_2 p^0_\delta(x, y, t))\, dt$ is bounded from below for every fixed δ by a constant independent of y.

Proof. An inspection of the subsequent proof and in particular inequality (4.19) shows the bounds

$$\frac{\Theta(0) - \Theta(p_{\max})}{\Delta t} \le \liminf_{y\nearrow 0} \frac{1}{\Delta t} \int_{t_0}^{t_0+\Delta t} (-\partial_2 p^0_\delta(x, y, t))\, dt,$$
$$\frac{\Theta(0) - \Theta(p_{\max})}{\Delta t} - \sup_{-1<y<0} g_\delta(y) \le \frac{1}{\Delta t} \int_{t_0}^{t_0+\Delta t} (-\partial_2 p^0_\delta(x, y, t))\, dt.$$

These imply the uniform estimates that are necessary in order to apply the Lebesgue convergence theorem. In particular V is bounded from below and we can use in the following $V > -\infty$.

We will consider here the most interesting case of maximal pressure in (x, t_0), $\liminf_{\delta\to 0} p^0_\delta(x, 0, t_0) = p_{\max}$. In the other case we find a region without explosions and can base the proof on the regularity of p^0 in $(x, 0, t_0)$. The claimed inequality is immediate in the case $V \ge 0$, since Θ is monotonically increasing. We can therefore assume from now on that $V < 0$.

To show (4.16) we first fix ΔV small compared to $|V|$, and $\Delta\rho$ small compared to $|V| \cdot \Delta t$. Next we choose $\delta > 0$ small enough to have

$$\liminf_{y\nearrow 0} \frac{1}{\Delta t} \int_{t_0}^{t_0+\Delta t} (-\partial_2 p^0_\delta(x, y, t))\, dt \le V + \Delta V,$$

$p^0_\delta(x, 0, t_0) \ge p_{\max} - \Delta\rho$, and a third smallness condition which depends on V and Δt and is made explicit later. Now we choose $y < 0$ close to 0 to have $g_\delta(y)$ small

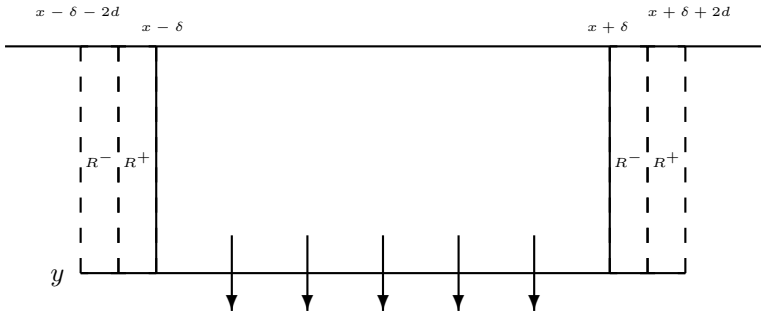

FIG. 2. *For the proof of Proposition* 4.7.

compared to $|V|$ ($g_\delta(y)$ is introduced later), and

$$\frac{1}{\Delta t}\int_{t_0}^{t_0+\Delta t}(-\partial_2 p_\delta^0(x,y,t))\,dt \le V + 2\Delta V. \tag{4.17}$$

We finally choose ε_0 small in order to have $p_\delta^\varepsilon(x,0,t_0^\varepsilon) \ge p_{\max} - 2\Delta\rho$ for all $\varepsilon < \varepsilon_0$ along the subsequence and in points $t_0^\varepsilon \to t_0$. Furthermore, ε_0 is chosen small enough to have

$$\frac{1}{\Delta t}\int_{t_0^\varepsilon}^{t_0+\Delta t}(-\partial_2 p_\delta^\varepsilon(x,y,t))\,dt \le V + 3\Delta V \tag{4.18}$$

for all $\varepsilon < \varepsilon_0$. At this point we exploited to have $y < 0$ fixed; in the interior of Ω spatial derivatives of time averages of the pressure p^ε converge.

By calculating the total flow into the box $R_{\delta,y}(x) := \{(x_1,x_2) : x-\delta < x_1 < x+\delta, y < x_2 < 0\}$ as illustrated in Figure 2, we verify the inequality

$$2\gamma\frac{1}{\Delta t}\left[\frac{\varepsilon}{2\delta}\sum_i \Phi_i(p^\varepsilon(x_i,\tau))\right]_{\tau=t_0^\varepsilon}^{t_0+\Delta t} \le V + 3\Delta V + g_\delta(y) + o(1) \tag{4.19}$$

for $\varepsilon \to 0$. The left-hand side measures the increase of volume on the upper boundary. On the right-hand side $V + 3\Delta V$ measures the maximal inflow into $R_{\delta,y}$ through the lower boundary as it was calculated in (4.18). $g_\delta(y)$ shall be a bound for ε-limits of the inflow through the lateral boundaries, multiplied by γ/δ. Then (4.19) follows as (4.12), since explosions only lower the left-hand side. The crucial point is now to find a bound $g_\delta(y)$ with $\lim_{y\nearrow 0} g_\delta(y) = 0$.

Construction of the bound $g_\delta(y)$. In the case of C^0 flows, that is, C^1 pressure fields, one concludes immediately that the flow through a slice of length $|y|$ is of order $|y|$. Here we only have some kind of continuity of the pressure field. Therefore, the estimate will be only of lower order, and the proof becomes more involved.

The basis for the proof is the following observation. For every (x,t) we consider the rectangles $R^- := R^-_{d,y}(x) := \{(x_1,x_2) : x < x_1 < x+d,\ y < x_2 < 0\}$ and $R^+ := R^+_{d,y}(x) := \{(x_1,x_2) : x+d < x_1 < x+2d,\ y < x_2 < 0\}$. Then for every aspect ratio $r := d/y$ the following limits coincide:

$$\lim_{y\to 0,\ d=ry}\frac{1}{|R^-|}\int_{R^-}\int_{t_0}^{t_0+\Delta t} p^0 = \lim_{y\to 0,\ d=ry}\frac{1}{|R^+|}\int_{R^+}\int_{t_0}^{t_0+\Delta t} p^0. \tag{4.20}$$

This follows from $p^0 \in L^2(0,T;H^1(\Omega))$. Time integrals of p^0 are in $H^1(\Omega)$, and for $q \in H^1(\Omega)$ holds (we use $R^- + te_1$ to denote the box R^-, translated to the right by t),

$$\begin{aligned}\frac{1}{d\cdot y}\left(\int_{R^+} q - \int_{R^-} q\right) &= \frac{1}{d\cdot y}\int_0^d dt \int_{R^- + te_1} (\partial_1 q)\\ &\le \frac{1}{y}\left(\int_{R^-\cup R^+} |\partial_1 q|^2\right)^{1/2} |2d\cdot y|^{1/2} \to 0\end{aligned}$$

for $y \to 0$ and fixed aspect ratio. We used that $|\partial_1 q|^2$ is in $L^1(\Omega)$ and therefore integrals over vanishing domains vanish.

We now conclude from (4.20) the estimate on $g_\delta(y)$. A weighted average of the horizontal flow in the box $R := \{(x_1,x_2) : x+\delta < x_1 < x+\delta+2d,\ y < x_2 < 0\}$ is (with $x_0 := x+\delta+d$)

$$\begin{aligned}\frac{1}{d}\int_R \partial_1 p^\varepsilon(x_1,x_2)\cdot\frac{d-|x_1-x_0|}{d}\,dx_1\,dx_2 &= \frac{1}{d^2}\left[\int_{R^+} p^\varepsilon - \int_{R^-} p^\varepsilon\right]\\ \to \frac{1}{r}\left[\frac{1}{|R^+|}\int_{R^+} p^0 - \frac{1}{|R^-|}\int_{R^-} p^0\right].&\end{aligned}$$

By (4.20) time integrals over this term become arbitrarily small for small y. For every small $|y|$ we find a corresponding d' such that the flow through the lateral side $\{(x_1,x_2) \mid x_1 = x+\delta+d',\ y < x_2 < 0\}$ is small.

We are not allowed to change the parameter δ, but we must show that the inflow through the lateral side $x_1 = x+\delta$ is small. To this end we use the aspect ratio r that can be chosen freely. The pressure along Γ is bounded and therefore the vertical velocity satisfies in the interior an estimate $|\partial_2 p^\varepsilon(x,y)| \le C_1|y|^{-1}$ by the representation formula (see, e.g., [6, p. 22]). We find that the total vertical inflow through $\{(x_1,x_2) \mid x+\delta < x_1 < x+\delta+2d,\ x_2 = y\}$ can be bounded by $d\cdot C_1/|y| = C_1\cdot r$. Choosing a small aspect ratio r we have a bound for the vertical inflow from below.

The vertical inflow from above is bounded by $C_2\cdot d$ since no negative explosions can occur. This contribution to $g_\delta(y)$ is therefore also small for small d (independent of the aspect ratio). Putting the results together we find that choosing r and then $|y|$ small, the total inflow through $x+\delta$ is bounded by a small number $g_\delta(y)$.

Conclusions from (4.19). As a first step we write (4.19) as

$$\begin{aligned}\frac{\varepsilon}{2\delta}\sum_i \Phi_i(p^\varepsilon(x_i,t_0+\Delta t)) \le \frac{\varepsilon}{2\delta}\sum_i \Phi_i(p^\varepsilon(x_i,t_0^\varepsilon))&\\ + \frac{\Delta t}{2\gamma}\left(V + 3\Delta V + g_\delta(y)\right) + o(1).&\end{aligned}$$

Since the derivatives of Φ_i are bounded, we conclude

$$p_\delta^\varepsilon(x,0,t_0+\Delta t) \le \rho < p_{\max},$$

where ρ depends only on $V<0$ and Δt, and not on δ. As in the proof of Proposition 3.6 we conclude that there is some $\rho_0 < p_{\max}$ independent of δ and $t_1 < t_0+\Delta t$ such that

$$p_\delta^\varepsilon(x,0,t) \le \rho_0 \qquad \forall t \in [t_1, t_0+\Delta t].$$

We said that a third smallness condition on δ should be satisfied. We demand that δ is small compared to $\delta_H(\rho_0)$ with δ_H from Lemma 3.1. With this choice we know that there are no explosions in the region $(x-\delta, x+\delta)\times(t_1, t_0+\Delta t)$.

We now fix a new small parameter $\sigma > 0$. Given σ we pick a new, smaller $\delta > 0$, repeat the above steps of the proof and find new $y < 0$ and $\varepsilon_0 > 0$. For the new δ we can assume by Lemma 4.2 that

$$|p^\varepsilon(\xi, 0, t_0+\Delta t) - p^\varepsilon_\delta(x, 0, t_0+\Delta t)| \le \sigma \qquad \forall \xi \in (x-\delta, x+\delta).$$

Here we used the linear regularity property.

The functions p^ε_δ are uniformly continuous in a neighborhood of $(x, 0, t_0+\Delta t)$ by Proposition A.3 and therefore $p^\varepsilon_\delta \to p^0_\delta$ is a uniform convergence for a subsequence. Along this subsequence we now take limits in (4.19),

$$\begin{aligned}
&\limsup_{\varepsilon\to 0} 2\gamma\frac{\varepsilon}{2\delta}\sum_i \Phi_i(p^\varepsilon(x_i, t_0+\Delta t)) \\
&\le \limsup_{\varepsilon\to 0} 2\gamma\frac{\varepsilon}{2\delta}\sum_i \Phi_i(p^\varepsilon_\delta(x, 0, t_0+\Delta t)) + C\sigma \\
&= \limsup_{\varepsilon\to 0} 2\gamma\frac{\varepsilon}{2\delta}\sum_i \Phi_i(p^0_\delta(x, 0, t_0+\Delta t)) + C\sigma \\
&= \Theta(p^0_\delta(x, 0, t_0+\Delta t)) + C\sigma.
\end{aligned}$$

For the second term we have, by $p^\varepsilon_\delta(x, t^\varepsilon_0) \ge p_{\max} - 2\Delta\rho$,

$$\liminf_{\varepsilon\to 0} 2\gamma\frac{\varepsilon}{2\delta}\sum_i \Phi_i(p^\varepsilon(x_i, t^\varepsilon_0)) \ge \Theta(p_{\max}) - C\Delta\rho.$$

We take $\limsup_{\varepsilon\to 0}$ in (4.19) and find

$$\begin{aligned}
&\frac{\Theta(p^0_\delta(x, 0, t_0+\Delta t)) - \Theta(p^0_\delta(x, 0, t_0))}{\Delta t} \\
&\qquad \le V + 3\Delta V + g_\delta(y) + C\sigma + C\Delta\rho.
\end{aligned} \tag{4.21}$$

We take the limit $y \nearrow 0$ and then $\delta \to 0$ using that $p^0_\delta \to p^0$ for a.e. $(x, 0, t)$. We find inequality (4.16) up to the error terms. Since ΔV, σ, and $\Delta\rho$ were arbitrary, we find the result. ∎

The assumption of linear regularity was used in the above proof only towards the end in order to replace p^ε by p^ε_δ in the evaluation of the laws Φ_i. We can also restrict ourselves with the conclusion that

$$\frac{\varepsilon}{2\delta}\sum_i \Phi_i(p^\varepsilon(x_i, t_0+\Delta t)) \le \Theta(p_{\max}) - C_1$$

for all small ε implies for some $c > 0$

$$p^0_\delta(x, t_0+\Delta t) \le p_{\max} - c\,C_1,$$

since the derivatives of Φ_i are bounded. We find the following corollary to the above proof.

Corollary 4.8 (weak lift-off condition). *We consider the nonlinear case and do not assume linear regularity. There exists $c > 0$ such that in every point (x, t_0) and for every $\Delta t > 0$,*

$$\liminf_{\delta\to 0} p^0_\delta(x, t_0+\Delta t) \le p_{\max} + c\ V\Delta t \tag{4.22}$$

with V as in Proposition 4.7.

With Theorem 4.6 we have derived a system of upscaled equations that is satisfied by every weak limit p^0 of the pressure functions p^ε. We have to verify that we have found all necessary information on the limit system. To this end we showed in [11] that solutions of the upscaled system of Theorem 4.6 are unique, at least for a linear function Θ. The uniqueness also implies the weak convergence of the initial sequence p^ε to the solution p^0 of the limit system.

5. Conclusions and outlook. We performed an analysis of a deterministic model for the motion of fronts in porous media. Upscaled equations were found under the hypothesis that limit patterns are finite. The limit equations include a hysteresis effect of the system: during imbibition, i.e., under inflow conditions and after a transition time, the pressure along the boundary coincides everywhere with $p_{\max}$; this value can therefore be interpreted as the capillary pressure of imbibition. When changing the boundary conditions to drainage, i.e., an outflow condition along the bottom, the system undergoes again a transitional regime before the capillary pressure of drainage (zero in our case) is reached.

An open question concerns the uniqueness of solutions of the upscaled system in the nonlinear case, that is, with the weak lift-off condition. It is desirable to extend the results to more general geometries and to more general equations for the fluid. We expect that in such systems the principle feature, the appearance of isolated explosions, remains the same.

We emphasize that the upscaled equations derived in this work form a mesoscopic model of a two-phase flow since the position of the front is still resolved. Desirable is the derivation of macroscopic laws from our mesoscopic results.

Appendix A. Regularity properties away from explosions. In this appendix we consider only regions without explosions. We expect that in this case the solution is regular. In order to get a feeling for the smoothing properties of the equations, we first consider the above equations omitting the projection Q_ε.

Remark A.1. The unique classical solution u^0 of

$$\Delta u^0(t) = 0 \qquad \text{in } \mathbb{R} \times \mathbb{R}_-, \tag{A.1}$$

$$\partial_t u^0 = -\lambda \partial_2 u^0 \qquad \text{on } \mathbb{R}, \tag{A.2}$$

with initial condition $u^0(x_1, 0, 0) = \operatorname{sgn}(x_1)$ and satisfying the uniform bounds $0 \le u^0 \le 1$ is given by

$$u^0(t, x) = \frac{2}{\pi} \arctan\left(\frac{x_1}{\lambda t}\right). \tag{A.3}$$

Proof. We demonstrate how to find u^0 in the self-similar form $u^0(x, t) = U(\frac{x}{t})$. The function U is harmonic in the lower half-plane and on $\{(x_1, x_2) \mid x_2 = 0\}$ it satisfies

$$x_1 \cdot \partial_1 U = \partial_2 U.$$

We use a complex differentiable function $f : \mathbb{C}^- \to \mathbb{C}$ to find $U = \operatorname{Re} f$. The condition on the real line translates into

$$\operatorname{Re}(z \cdot f' - i f') = 0.$$

We set $z \cdot f' - i f' = ci$ for a real number c and find

$$f' = c\frac{i}{z - i} = c\frac{-1}{1 + iz} = c\frac{1 - i\bar{z}}{|1 + iz|^2}.$$

This implies

$$\partial_1 U = \operatorname{Re} f' = \frac{c}{1+|x_1|^2}$$

on the real line. Using $U(x_1) \to \pm 1$ for $x_1 \to \pm\infty$ determines the constant c to be $2/\pi$ and yields the result. The x_1-derivative of U inside the domain is calculated as $\partial_1 U(x_1, x_2) = \frac{1-x_2}{(1-x_2)^2+x_1^2}$; this yields the complete form of U,

$$U(x_1, x_2) = \frac{2}{\pi} \arctan\left(\frac{x_1}{1-x_2}\right). \qquad \square$$

The explicit solution above gives us an idea of how solutions to the original equations (1.2)–(1.5) behave qualitatively. Unfortunately, the picture may change in many respects once the coefficients in $\partial_t p^\varepsilon(\cdot, 0, t) = -a(\cdot) Q_\varepsilon \partial_2 p^\varepsilon(\cdot, 0, t)$ depend on x_1. But one useful property remains valid as one can see by using the exact solution as a comparison function.

Remark A.2. Consider the original equations (1.2)–(1.5) with initial values

$$p^\varepsilon(x_1, 0, 0) = \begin{cases} 0 & \text{for } x_1 < 0, \\ 1 & \text{for } x_1 > 0. \end{cases}$$

Then for every $\delta > 0$ there exists a constant C such that

$$p^\varepsilon(-\delta, 0, t) \le Ct \qquad \forall t.$$

PROPOSITION A.3. *We study solutions $p^\varepsilon : (-1,1) \times (-1,0) \times (0,t) \to \mathbb{R}$ of the original equations (1.2)–(1.5) for $V_0 \in C^1$ and with (piecewise) linear laws $\mathcal{P}_0$. If no explosions happen on $(-\delta_0, \delta_0) \times (0,t)$, then, for every $0 < \delta < \delta_0$, the family $p^\varepsilon(\cdot, 0, t)$ is uniformly continuous in $(-\delta, \delta)$.*

In the case of nonlinear laws, for every $\delta < \delta_0$, $0 < t_1 < t$, and arbitrary $\kappa > 0$ we can write $p^\varepsilon|_\Gamma$ as $p^\varepsilon|_\Gamma = p_A + p_B = p_{A,1} + p_{A,2} + p_B$ with

$$\begin{aligned} \partial_t p_A &\in L^2((t_1,t) \times (-\delta,\delta)), & \|p_A\| &\le C(\kappa), \\ p_{A,1} &\in L^2((t_1,t), H^1(-\delta,\delta)), & \|p_{A,1}\| &\le C(\kappa), \\ p_{A,2} &\in L^2((t_1,t), L^\infty(-\delta,\delta)), & \|p_{A,2}\| &\le C\varepsilon^\alpha, \\ p_B &\in L^\infty((t_1,t) \times (-\delta,\delta)), & \|p_B\| &\le \kappa. \end{aligned}$$

In particular, all spatial averages p^ε_δ of p^ε are uniformly continuous on $(t_1,t) \times (-\delta,\delta)$. The constant $\alpha > 0$ is independent of ε, p^ε, δ, and κ.

A solution for $\varepsilon = 1$ on the extended domain $\mathbb{R} \times (-\infty, 0) \times (0, \infty)$ without explosions satisfies

$$|p^\varepsilon(x_1, 0, t) - p^\varepsilon(0, 0, t)| \to 0 \quad \text{for} \quad t \to \infty, \tag{A.4}$$

independent of the initial values.

Proof. Idea: Assume that there are no explosions at all. Then we write the linear laws as

$$\partial_t p^\varepsilon(\cdot, 0, \tau) = -a^\varepsilon(\cdot) Q_\varepsilon \partial_2 p^\varepsilon(\cdot, 0, \tau) \quad \text{on } \Gamma_1^\varepsilon, \ \forall \tau \in (0, t).$$

In this case we can show uniform estimates for $\partial_t p^\varepsilon|_\Gamma \in L^\infty((3t/4,t),L^2(\Gamma))$. These are at the same time bounds for $Q_\varepsilon \partial_2 p^\varepsilon|_\Gamma = -\frac{1}{a^\varepsilon(\cdot)}\partial_t p^\varepsilon|_\Gamma$. We can decompose p^ε into the "macroscopic" part, the harmonic, periodic function p_m satisfying $\partial_2 p_m = Q_\varepsilon \partial_2 p^\varepsilon$ on Γ, and a remainder $u^\varepsilon = p^\varepsilon - p_m$. In Lemma A.4 we show that u^ε is small in L^∞. The elliptic regularity theory yields uniform estimates for $p_m \in L^\infty((t/2,t),C^\alpha(\Gamma))$ for $\alpha < 1/2$. This yields the claim.

We now show the estimate for $\partial_t p^\varepsilon$ in the case of $\delta_0 = 1$ (no explosions along Γ). By the energy estimate for $p^\varepsilon \in L^2((0,t),H^1(\Omega))$ we find a time instance $t_1 < t/2$ such that $p^\varepsilon(t_1) \in H^1(\Omega)$ satisfies an ε-independent bound. We multiply $\Delta p^\varepsilon = 0$ by $\partial_t p^\varepsilon$ and find

$$\begin{aligned}\int_{\Gamma_0} V_0 \partial_t p^\varepsilon &= \int_\Omega \nabla p^\varepsilon \cdot \partial_t \nabla p^\varepsilon - \int_{\Gamma_1^\varepsilon} \partial_2 p^\varepsilon \cdot \partial_t p^\varepsilon \\ &= \int_\Omega \partial_t \frac{1}{2}|\nabla p^\varepsilon|^2 + \int_{\Gamma_1^\varepsilon} \frac{1}{a^\varepsilon(\cdot)}|\partial_t p^\varepsilon|^2.\end{aligned}$$

This yields an estimate for $\partial_t p^\varepsilon|_{\Gamma_1^\varepsilon} \in L^2((t_1,T)\times\Gamma_1^\varepsilon)$. We find a time instance t_2, $t_1 < t_2 < 3t/4$, with bounded (by an ε-independent constant) $\partial_t p^\varepsilon(t_2)|_{\Gamma_1^\varepsilon} \in L^2$. We multiply the differentiated equation $\partial_t \Delta p^\varepsilon = 0$ by $\partial_t p^\varepsilon$ and find

$$\begin{aligned}\int_{\Gamma_0} \partial_t V_0 \partial_t p^\varepsilon &= \int_\Omega |\partial_t \nabla p^\varepsilon|^2 - \int_{\Gamma_1^\varepsilon} \partial_t \partial_2 p^\varepsilon \cdot \partial_t p^\varepsilon \\ &= \int_\Omega |\partial_t \nabla p^\varepsilon|^2 + \partial_t \int_{\Gamma_1^\varepsilon} \frac{1}{2}\frac{1}{a^\varepsilon(\cdot)}|\partial_t p^\varepsilon|^2 - \int_{\Gamma_1^\varepsilon} \frac{1}{2}\frac{\partial_t(a^\varepsilon(p^\varepsilon))}{(a^\varepsilon)^2}|\partial_t p^\varepsilon|^2.\end{aligned}$$

In the case of a linear law a^ε is independent of p^ε and therefore the last term vanishes. An integration yields the claimed estimate for $\partial_t p^\varepsilon|_\Gamma \in L^\infty((t_2,t),L^2(\Gamma))$.

In the case of a nonlinear law the coefficient $a(x_1,t)$ depends on $p^\varepsilon(x_1,0,t)$, and the term containing $\partial_t a^\varepsilon |\partial_t p^\varepsilon|^2$ cannot be controlled by the other two terms. Nevertheless, the argument leading to the estimate for

$$\partial_t p^\varepsilon|_{\Gamma_1^\varepsilon} \in L^2((t_1,T)\times\Gamma_1^\varepsilon)$$

remains valid. With $p_A = p^\varepsilon$ and $p_B = 0$ we found the claimed decomposition. The estimates for the x_1-derivative of $p_{A,1} := p_m$ follow for harmonic functions from the estimates for the Neumann boundary values.

We now study the general case $\delta_0 < 1$. We choose $\Delta t < t$ small (depending on δ_0 and ρ), and consider from now on the solution only on the time interval $(t-\Delta t, t)$, the coefficients a^ε are always given by the original solution. We now decompose the solution into a part p_A with the initial values of p^ε and without explosions on Γ, and a second part p_B that captures the evolution of the explosions. Then on p_A the above arguments for $\delta_0 = 1$ can be applied. The function p_B is small on $(t-\Delta t, t)\times(-\delta_0/2, \delta_0/2)$ for Δt small by the maximum principle of Remark A.2.

Formula (A.4) follows from a scaling argument,

$$|p^\varepsilon(x_1,0,t) - p^\varepsilon(0,0,t)| = |p^{\varepsilon/t}(x_1/t,0,1) - p^{\varepsilon/t}(0,0,1)|,$$

where the scaled solution $p^{\varepsilon/t}$ has initial values $p^{\varepsilon/t}(x_1/t,0,0) = p^\varepsilon(x_1,0,0)$. The result follows from the uniform continuity of $p^{\varepsilon/t}$ at time 1, since x_1/t is arbitrarily close to 0. Note that we have to modify the solutions $p^{\varepsilon/t}|_{(-1,1)\times(-1,0)}$ on the boundary

in order to guarantee their periodicity. Applying the above argument with an initial time close to 1, this change introduces only a small error term in the expression $p^{\varepsilon/t}(x_1/t,0,1)-p^{\varepsilon/t}(0,0,1)$. ∎

LEMMA A.4. *Let u^ε be a sequence of periodic harmonic functions on $\Omega=(-1,1)\times(-1,0)$ satisfying a harmonic Neumann condition on the lower boundary and with the following properties on the upper boundary $\Gamma=(-1,1)\times\{0\}$:*

$$\begin{aligned} Q_\varepsilon\partial_2 u^\varepsilon &= 0,\\ g^\varepsilon := u^\varepsilon - Q_\varepsilon u^\varepsilon &\quad \text{satisfies } \|\partial_{x_1} g^\varepsilon|_{\Gamma_1^\varepsilon}\|_{L^2}\le C. \end{aligned}$$

Then $u^\varepsilon|_\Gamma\to 0$ in $L^\infty(\Gamma)$ for $\varepsilon\to 0$ independent of the sequence g^ε.

Proof. Note that we have the technical difficulty of $g^\varepsilon\notin H^1$ in general. We therefore introduce a new projection $\tilde Q_\varepsilon$ such that $\tilde Q_\varepsilon v=Q_\varepsilon v$ on Γ_1^ε with $\tilde Q_\varepsilon$ bounded in $\mathcal{L}(H^1(\Gamma),H^1(\Gamma))$. For $v\in L^2(\Gamma)$ we can use $\tilde v$, the harmonic function on Ω that satisfies $\tilde v|_{\Gamma_1^\varepsilon}=Q_\varepsilon v|_{\Gamma_1^\varepsilon}$ and $\partial_2\tilde v=0$ on Γ_2^ε, together with periodicity and a harmonic Neumann condition on the lower boundary. We set $\tilde Q_\varepsilon v:=\tilde v$.

With this modified projection we can consider the bounded sequence $w^\varepsilon:=u^\varepsilon|_\Gamma-\tilde Q_\varepsilon u^\varepsilon|_\Gamma\in H^1(\Gamma)$. We multiply $\Delta u^\varepsilon=0$ by u^ε to find

$$\begin{aligned} \int_\Omega|\nabla u^\varepsilon|^2 &= \int_\Gamma \partial_2 u^\varepsilon\ u^\varepsilon = \int_{\Gamma_1^\varepsilon}\partial_2 u^\varepsilon\ u^\varepsilon\\ &= \int_{\Gamma_1^\varepsilon}\partial_2 u^\varepsilon\ (u^\varepsilon-\tilde Q_\varepsilon u^\varepsilon) = \int_\Gamma \partial_2 u^\varepsilon\ w^\varepsilon. \end{aligned}$$

Since the family w^ε is bounded in $H^1(\Gamma)$ and $\partial_2 u^\varepsilon|_\Gamma\in H^{-1}(\Gamma)$ is bounded by $u^\varepsilon|_\Gamma\in L^2(\Gamma)$, we find an a priori bound for $u^\varepsilon\in H^1(\Omega)$.

In order to show the L^∞-convergence of u^ε we use again the above calculation and the fact that the family of functions $\partial_2 u^\varepsilon|_\Gamma\in H^{-1/2}(\Gamma)$ is bounded. We claim that the functions w^ε vanish in $H^{1/2}(\Gamma)$ at the rate $\varepsilon^{1/4}$. The functions w^ε are bounded in $C^{1/2}(\Gamma)$ and have vanishing averages on all intervals $\varepsilon(k-\gamma,k+\gamma)$. Therefore they satisfy an L^∞-estimate $\|w^\varepsilon\|_{L^\infty}\le C\sqrt{\varepsilon}$. Additionally $w^\varepsilon\in H^1(\Gamma)$ is bounded. By an interpolation between $L^2(\Gamma)$ and $H^1(\Gamma)$ we conclude that $\|w^\varepsilon\|_{H^{1/2}(\Gamma)}\le C\varepsilon^{1/4}$. We conclude that $u^\varepsilon\in H^1(\Omega)$ vanishes at the rate $\varepsilon^{1/8}$. We use an inverse estimate of an L^∞-norm in terms of an L^q-norm (exploiting that u^ε is constant on ε-intervals up to the error w^ε of order $O(\sqrt{\varepsilon})$), and a trace theorem with $q>8$ to find

$$\begin{aligned} \|u^\varepsilon\|_{L^\infty(\Gamma)} &\le C\left[\sqrt{\varepsilon}+\varepsilon^{-1/q}\|u^\varepsilon\|_{L^q(\Gamma)}\right]\\ &\le C(q)\left[\sqrt{\varepsilon}+\varepsilon^{-1/q}\|u^\varepsilon\|_{H^1(\Omega)}\right]\le C\varepsilon^{1/8-1/q}. \end{aligned}$$

This shows the assertion. ∎

REFERENCES

[1] G. ALBERTI AND S. MÜLLER, *Two-scale Young measures for variational problems with multiple scales*, Comm. Pure Appl. Math., 54 (2001), pp. 761–825.

[2] J.-L. AURIAULT AND E. SANCHEZ-PALENCIA, *Remarques sur la loi de Darcy pour les écoulements biphasiques en milieu poreux*, J.M.T.A., (supplement) (1986), pp. 141–156.

[3] G. BARENBLATT, J. GARCIA-AZORERO, A. DEPABLO, AND J.-L. VAZQUEZ, *Mathematical model of non-equilibrium water-oil displacement in porous strata*, Appl. Anal., 65 (1997), pp. 18–45.

[4] A. Y. Beliaev and R. J. Schotting, *Analysis of a new model for unsaturated flow in porous media including hysteresis and dynamic effects*, J. Comput. Geosci., 5 (2001), pp. 345–368.
[5] A. Bourgeat, *Two-phase flow*, in Homogenization and Porous Media, U. Hornung, ed., Springer, New York, 1997, pp. 95–127.
[6] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Grundlehren Math. Wiss. 224, Springer, Berlin, 1983.
[7] M. C. Leverett, *Steady flow of gas-oil-water mixtures through unconsolidated sands*, Trans. AIME, 132 (1938), p. 149.
[8] S. Luckhaus and P. I. Plotnikov, *Entropy solutions to the Buckley-Leverett equations*, Siberian Math. J., 41 (2000), pp. 329–348.
[9] A. Mikelic and L. Paoli, *On the derivation of the Buckley-Leverett model from the two fluid Navier-Stokes equations in a thin domain*, Comput. Geosci., 1 (1997), pp. 59–83.
[10] B. Schweizer, *A stochastic model for fronts in porous media*, Ann. Mat. Pura Appl. (4), to appear.
[11] B. Schweizer, *Laws for the Capillary Pressure via the Homogenization of Fronts in Porous Media*, Habilitationsschrift, Ruprecht-Karls-Universität Heidelberg, Heidelberg, Germany, 2002.

SIAM J. MATH. ANAL.
Vol. 36, No. 5, pp. 1522–1539

DELAYED LOSS OF STABILITY IN NONAUTONOMOUS DIFFERENTIAL EQUATIONS WITH RETARDED ARGUMENT*

BERNHARD LANI-WAYDA† AND KLAUS R. SCHNEIDER‡

Abstract. Assume that zero is a stable equilibrium of an ODE $\dot{x} = f(x, \lambda)$ for parameter values $\lambda < \lambda_0$ and becomes unstable for $\lambda > \lambda_0$. If we suppose that $\lambda(t)$ varies slowly with t, then, under some conditions, the trajectories of the nonautonomous ODE $\dot{x} = f(x, \lambda(t))$ stay close to zero even long after $\lambda(t)$ has crossed the value λ_0. This phenomenon is called "delayed loss of stability" and is well known for ODEs. In this paper, we describe an analogous phenomenon for delay equations of the form $\dot{x}(t) = f(t, x(t-1))$. We study an example which requires combining linearization at zero with estimates on the nonlinear behavior away from zero, and where we obtain an explicit estimate on the time until the growth of $|x(t)|$ becomes "visible."

Key words. nonautonomous delay equations, slowly changing parameters, delayed loss of stability, decay and growth of oscillations

AMS subject classifications. Primary, 34K12; Secondary, 34K06

DOI. 10.1137/S0036141002409532

1. Introduction.

Dynamical systems as mathematical models of real life processes depend on several parameters which are assumed to be fixed within some time period (see, e.g., [4], [12], [1]). The influence of a parameter λ on the behavior of a dynamical system is studied within the framework of bifurcation theory. Suppose now that a relevant system parameter λ changes very slowly in time, for example, because of an aging process. In the model equation, one can then replace the parameter λ by $\lambda(\varepsilon t)$, where $\varepsilon > 0$ is a small. (The new equation is then nonautonomous.) The so-called dynamic bifurcation theory is concerned with the investigation of the corresponding changes of the system behavior [1]. A special phenomenon, known as *delayed loss of stability*, can lead to dramatic consequences (e.g., thermal explosion [7]). For ordinary differential equations (ODEs), this effect is well known and has been studied from different points of view [3], [13], [2], [14].

Let us illustrate the phenomenon by considering the simple linear equation

$$\dot{y}(t) = k(\varepsilon t) y(t), \tag{1.1}$$

where $\varepsilon > 0$ is small, such that the coefficient $k(\varepsilon t)$ in (1.1) changes slowly in time. Setting $\varepsilon t = \tau$, $y(t) = y(\tau/\varepsilon) = x(\tau)$, we get from (1.1)

$$\varepsilon \frac{dx}{d\tau} = k(\tau) x. \tag{1.2}$$

Concerning the function k, we suppose the following.

(A) $k : \mathbb{R} \longrightarrow \mathbb{R}$ is continuous, strictly increasing, and there exist numbers $\tau_- < \tau_0 < \tau_+$ such that

$$k(\tau) < 0 \quad \text{for } \tau < \tau_0, \qquad k(\tau) > 0 \quad \text{for } \tau > \tau_0, \quad \int_{\tau_-}^{\tau_+} k(\tau) d\tau = 0. \tag{1.3}$$

*Received by the editors June 12, 2002; accepted for publication (in revised from) April 16, 2004; published electronically March 25, 2005.
http://www.siam.org/journals/sima/36-5/40953.html
†Mathematisches Institut der Universität Giessen, Arndtstr. 2, 35392 Giessen, Germany (Bernhard.Lani-Wayda@math.uni-giessen.de).
‡Weierstraß-Institut für Angewandte Analysis und Stochastik, Mohrenstrasse 39, 10117 Berlin, Germany (schneider@wias-berlin.de).

The so-called associated system to (1.2) reads

$$\frac{dx}{d\sigma} = k(\tau)x(\sigma), \tag{1.4}$$

where τ in the right-hand side has to be considered as a parameter and σ is the independent variable. From hypothesis (A) it follows that the equilibrium $x = 0$ of the associated equation (1.4) is stable for $\tau < \tau_0$ and unstable for $\tau > \tau_0$; that is, it changes its stability at $\tau = \tau_0$.

The solution $x(\cdot, \tau_-, x_-)$ of (1.1) satisfying $x(\tau_-, \tau_-, x_-) = x_-$ is explicitly given by

$$x(\tau, \tau_-, x_-) = x_- \exp\left\{\frac{1}{\varepsilon}\int_{\tau_-}^{\tau} k(s)ds\right\}.$$

We see that if k satisfies assumption (A), then $x(\tau, \tau_-, x_-)$ is exponentially decaying for $\tau_- < \tau < \tau_0$ and stays near $x = 0$ also for some time interval $\tau_0 < \tau < \hat{\tau}$ with $\hat{\tau} < \tau_+$, during which $x = 0$ is an unstable equilibrium of (1.4). Obviously, analyticity of the equation is inessential for this simple phenomenon (but plays a role in connection with more refined results as, e.g., [13]).

The main goal of this paper is to describe a similar effect for differential-delay equations, where we restrict ourselves to simplest cases.

As a preparation, we consider in section 2 the constant coefficient equation

$$\dot{x}(t) = cx(t-1), \tag{1.5}$$

with $c \in [-3\pi/4, -\pi/4]$. It is well known that the zero solution of (1.5) is stable for $c \in (-\pi/2, 0)$, and unstable if $c < -\pi/2$. Contrary to the ODE case, the exponential rate of growth or decay is not directly given by c but has to be estimated. We provide such estimates. As further preparatory background material, section 2 contains a simple version of the variation-of-constants formula for the case of inhomogeneous equations with nonconstant coefficient.

In section 3 we compare solutions of

$$\dot{x}(t) = g(t, x(t-1))$$

(with nonlinear g) on successive time intervals I_i to solutions of the equation

$$\frac{dx}{dt} = c_i x(t-1),$$

with constants c_i which are values of $\partial_2 g(\cdot, 0)$ on I_i. In Theorem 3.2, we obtain estimates that express the phenomenon of delayed loss of stability for differential-delay equations. In section 4 we treat the example equation

$$\dot{x}(t) = (-\pi/4 - \varepsilon t)\arctan(x(t-1)).$$

Here, we study the initial value problem with the initial segment identically 1, and estimate the time until the solution is close enough to zero by a method that is not based on linearization. Theorem 3.2 is then applicable to the motion close to zero, and we obtain a lower bound for the time t_1 until the solution reaches absolute value 1 again. (This is to be understood in the sense that certainly $|x(t)| \le 1$ for $t \le t_1$.)

We do not rigorously prove the (essentially obvious) fact that the solution grows again after the coefficient has crossed the stability border, but give a heuristic argument for this in section 4.

Notation. For bounded functions φ on $[-1,0]$, the sup-norm is denoted by $|\varphi|$. Generally, we use the symbol $\|\ \|_\infty$ for the sup-norm of bounded functions on some domain.

Let $\mathbf{C}$ denote the space of continuous functions on $[-1,0]$ with the max-norm. Assume that $G : \mathbb{R}\times\mathbf{C} \to \mathbb{R}$ is continuous, is locally Lipschitz continuous with respect to the second argument, and satisfies a linear growth condition

$$|G(t,\varphi)| \le L(t)(1+|\varphi|) \quad (t\in\mathbb{R},\ \varphi\in\mathbf{C})$$

with $L : \mathbb{R} \longrightarrow \mathbb{R}_0^+$ continuous. Then, for $\varphi \in \mathbf{C}$ and $\tau \in \mathbb{R}$, there is a unique continuous function $x^{G,\varphi,\tau} : [\tau-1,\infty] \longrightarrow \mathbb{R}$ such that

$$x_\tau^{G,\varphi,\tau} = \varphi, \quad \dot{x}^{G,\varphi,\tau}(t) = G(t, x_t^{G,\varphi,\tau}) \quad \text{for } t \ge \tau.$$

(At $t=\tau$, the derivative is to be read as right-side derivative.) The symbol x_t, as usual, denotes the segment of the function x at time t, that is, $x_t(\theta) = x(t+\theta)$, $-1 \le \theta \le 0$.

We shall need solutions of *linear* equations also for discontinuous initial values; let $\mathbf{J}$ denote the space of functions $\varphi : [-1,0] \longrightarrow \mathbb{R}$ which are continuous on $[-1,0)$ but possibly have a jump discontinuity at 0 (i.e., $\lim_{t\longrightarrow 0, t<0}\varphi(t)$ exists). We use the sup-norm $|\ |$ also on this space, and we introduce the weaker norm $|\ |_*$ on $\mathbf{J}$ defined by

$$|\psi|_* := |\psi(0)| + \int_{-1}^{0} |\psi(s)|\,ds.$$

Using the space $\mathbf{J}$ is a simple approach to variation of constants which suffices for our purposes.

2. Linear equations. First we consider linear equations of the type

$$\dot{x}(t) = a(t)x(t-1). \tag{a}$$

The proof of Proposition 2.1 below uses mainly simple standard arguments, in combination with known results for initial segments in $\mathbf{C}$. With a view to the paper's length, it is omitted. A detailed proof is available from the authors.

PROPOSITION 2.1. *Let $\tau, T \in \mathbb{R}$, $\tau < T$, and let $a : [\tau,T] \longrightarrow \mathbb{R}$ be continuous.*

(a) *For $\psi \in \mathbf{J}$ and $s \in [\tau,T]$ there exists a unique solution $x^{a,\psi,s} : [s-1,T] \longrightarrow \mathbb{R}$ of the initial value problem $\dot{x}(t) = a(t)x(t-1)$, $x_s = \psi$.*

(b) *The map*

$$F : (\mathbf{J}, |\ |_*) \times \{(s,t)\in[\tau,T]^2 \mid s\le t\} \ni (\psi,s,t) \mapsto x^{a,\psi,s}(t) \in \mathbb{R}$$

is continuous.

(c) *If $a \in C^1$ and $T \ge \tau+3$, then for $t \ge \tau+3$, $t \le T$ the segment $x_t^{a,\psi,\tau}$ is C^2.*

Our aim is to express solutions of (a) with a slowly varying coefficient by solutions of the constant coefficient equation

$$\dot{x}(t) = c\cdot x(t-1) \quad (c\in\mathbb{R}). \tag{c}$$

First, we provide more detailed information on (c) for values of c around the stability border $-\pi/2$. For $c \in \mathbb{R}$, let $\Sigma_c \subset \mathbb{C}$ denote the set of zeroes of the characteristic function $\lambda \mapsto \lambda - c \cdot \exp(-\lambda)$ associated with (c).

PROPOSITION 2.2. *For $c \in (-\infty, -e^{-1})$, the set Σ_c has the form*

$$\Sigma_c = \{\lambda_k(c) \mid k \in \mathbb{N}_0\} \cup \{\overline{\lambda_k(c)} \mid k \in N_0\},$$

where $\lambda_k(c) = \rho_k(c) + i\omega_k(c)$, $\overline{\lambda_k(c)} = \rho_k(c) - i\omega_k(c)$ $(k \in \mathbb{N}_0)$, and $\omega_k(c) \in (2k\pi, (2k+1)\pi)$. The following properties hold:

(a) *$\rho_k(c) > \rho_{k+1}(c)$ $(k \in \mathbb{N}_0)$, so that $\rho_0(c) = \max \operatorname{Re} \Sigma_c$.*

(b) *$\rho_0(-\pi/2) = 0$.*

(c) *For $c \in [-3\pi/4, -\pi/4]$, $\rho_0'(c)$ exists and $\rho_0'(-\pi/2) = \frac{-2\pi}{4+\pi^2}$. Further, $\omega_0(c) \in (\pi/4, \pi)$, and*

$$-\frac{4(\pi+2)}{\pi^2} \le \rho_0'(c) \le -\frac{4(\pi-2)}{3\pi^2}, \quad \text{and}$$

$$\rho_0(c) \le \begin{cases} -|c+\pi/2|\dfrac{4(\pi-2)}{3\pi^2} & \text{if } c > -\pi/2, \\ |c+\pi/2|\dfrac{4(\pi+2)}{\pi^2} & \text{if } c \le -\pi/2. \end{cases}$$

(d) *$|\rho_0(c)| \le (\pi+2)/\pi \le 2$ for $c \in [-3\pi/4, -\pi/4]$.*

Proof. The assertions on Σ_c and property (a) follow from Theorem 5 in [15]. Writing $\lambda = \rho + i\omega$, the characteristic equation $\lambda = c\exp(-\lambda)$ is equivalent to the equations

$$\rho = ce^{-\rho}\cos\omega, \quad \omega = -ce^{-\rho}\sin\omega.$$

Note that $\sin\omega = 0$ would imply $\omega = 0$, but we know already that for $c < -e^{-1}$ there exist no real roots of the characteristic equation. Hence, we can restrict ourselves to the case $\sin\omega \neq 0$, and we obtain from the above two equations $\omega = -c\exp(\omega\cot\omega)\sin\omega$. Setting

$$\chi(\omega) := \frac{\omega}{\sin\omega}\exp(-\omega\cot\omega) \quad \text{for } \omega \in \mathbb{R} \setminus \{k\pi \mid k \in \mathbb{Z}\},$$

the last equation is equivalent to

$$\chi(\omega) = -c. \tag{2.1}$$

The function χ is discussed in [15]. One has

$$\chi'(\omega) = \frac{\chi(\omega)}{\omega}[(1-\omega\cot\omega)^2 + \omega^2]; \tag{2.2}$$

χ and χ' are positive on $(0,\pi)$, with $\chi(\omega) \longrightarrow e^{-1}$ as $\omega \longrightarrow 0$, and $\chi(\omega) \longrightarrow \infty$ as $\omega \longrightarrow \pi$, $\omega < \pi$. For $c \in (-\infty, -e^{-1})$, the number $\omega_0(c)$ is the unique solution of (2.1) in $(0,\pi)$ and $\rho_0(c) = -\omega_0(c)\cot\omega_0(c) = \log\frac{-c\sin\omega_0(c)}{\omega_0(c)}$, so we have

$$\rho_0(c) = \log(-c) + \log\sin\omega_0(c) - \log\omega_0(c). \tag{2.3}$$

Obviously $\chi(\pi/2) = \pi/2$, so $\omega_0(-\pi/2) = \pi/2$ and $\rho_0(-\pi/2) = 0$. Properties (a) and (b) are proved.

(c) It follows from the inverse function theorem and from (2.3) that ω_0 and ρ_0 are differentiable functions on $(-\infty, -e^{-1})$, in particular, on $[-3\pi/4, -\pi/4]$. Using (2.2) we obtain, for $c \in (-\infty, -e^{-1})$,

$$\begin{aligned}\omega_0'(c) &= -\frac{1}{\chi'(\omega_0(c))} = -\frac{\omega_0(c)}{\chi(\omega_0(c))[(1-\omega_0(c)\cot\omega_0(c))^2+\omega_0(c)^2]}\\ &= \frac{\omega_0(c)}{c[(1-\omega_0(c)\cot\omega_0(c))^2+\omega_0(c)^2]},\end{aligned}$$

and from (2.3) we get

$$\begin{aligned}\rho_0'(c) &= \frac{1}{c} + \omega_0'(c)\left(\cot\omega_0(c) - \frac{1}{\omega_0(c)}\right) = \frac{1}{c} + \frac{\omega_0(c)\cot\omega_0(c)-1}{c[(1-\omega_0(c)\cot\omega_0(c))^2+\omega_0(c)^2]}\\ &= \frac{1}{c}\left(1 + \frac{\omega_0(c)\cot\omega_0(c)-1}{[(1-\omega_0(c)\cot\omega_0(c))^2+\omega_0(c)^2]}\right).\end{aligned}$$

In particular, we see that $\rho_0'(-\pi/2) = \frac{-2}{\pi}(1+\frac{-1}{1+\pi^2/4}) = \frac{-2\pi}{4+\pi^2}$, which is the first assertion of (c). Note now that $\chi(\pi/4) = \frac{\pi/4}{\sqrt{2}/2}\exp(-\pi/4) = \frac{\pi\sqrt{2}}{4}\exp(-\pi/4) < \frac{\pi}{4}\frac{\sqrt{2}}{1+\pi/4} < \pi/4$, so $\omega_0(-\pi/4) > \pi/4$. It follows that

$$\omega_0([-3\pi/4, -\pi/4]) \subset (\pi/4, \pi). \tag{2.4}$$

Further, for all $\omega > 0$ and $u \in \mathbb{R}$, one has $|\frac{u}{u^2+\omega^2}| \le \frac{1}{2\omega}$. With (2.4) we conclude that

$$\left|\frac{\omega_0(c)\cot\omega_0(c)-1}{[(1-\omega_0(c)\cot\omega_0(c))^2+\omega_0(c)^2]}\right| \le \frac{1}{2\omega_0(c)} \le \frac{2}{\pi}.$$

With the above expression for $\rho_0'(c)$, we now obtain that $\rho_0'(c) \in \frac{1}{c}[1-2/\pi, 1+2/\pi]$ for $c \in [-3\pi/4, -\pi/4]$, so for these c one has $(1+2/\pi)(-4/\pi) \le \rho_0'(c) \le (1-2/\pi)(-4/3\pi)$, or

$$-\frac{4(\pi+2)}{\pi^2} \le \rho_0'(c) \le -\frac{4(\pi-2)}{3\pi^2}.$$

The estimates on $\rho_0(c)$ in part (c) follow by integration.

(d) It follows from (b) and (c) that for $c \in [-3\pi/4, -\pi/4]$ one has

$$|\rho_0(c)| \le \frac{\pi}{4}\frac{4(\pi+2)}{\pi^2} = \frac{\pi+2}{\pi} \le 2. \qquad \square$$

It is known that for $c < -e^{-1}$ and $\rho > \rho_0(c)$, there exists $K > 0$ such that all solutions $x^{c,\varphi,\tau}$ of (c) satisfy an estimate of the form $|x^{c,\varphi,\tau}(t)| \le K\exp(\rho(t-\tau))|\varphi|$ for $t \ge \tau$. (Compare, e.g., Corollary 6.1, page 215 of [9], and the definition of the constant K given in the proof of Lemma 6.2, page 213 of the same reference.) Analogous results hold for much more general linear equations. We now derive a similar estimate with an explicit value for K, and with $\rho = \rho_0(c)$, for the special case of (c).

PROPOSITION 2.3. *Set* $K := [4 + 15(3\pi/4) + 24(3\pi/4)^2]e^4$, *and let* $c \in [-3\pi/4, -\pi/4]$ *and* $t, s \in \mathbb{R}$, $t \ge s$.

(a) *For* $\psi \in C^2([-1,0],\mathbb{R})$, *one has* $|x^{c,\psi,s}(t)| \le (4|\psi| + 7|\psi'| + 13|\psi''|)e^{\rho_0(c)(t-s)}$.

(b) *For* $\varphi \in \mathbf{J}$, *one has* $|x^{c,\varphi,s}(t)| \le |\varphi|K\exp[\rho_0(c)(t-s)]$.

Proof. Since (c) is autonomous, it suffices to prove the assertions for the case $s = 0$. For $t > 0$, we have for $\varphi \in \mathbf{C}$ the series expansion

$$x^{c,\varphi,0}(t) = \sum_{\mu \in \Sigma_c} (\mathrm{pr}_\mu\, \varphi) \exp(\mu t),$$

where $\mathrm{pr}_\mu\, \varphi = \frac{1}{1+\mu}[\varphi(0) + \mu \int_{-1}^{0} e^{-\mu s}\varphi(s)ds]$ (see [15, Theorem 6], or [10, Lemma 6.8], or [5, Theorem 6.3, and Corollary 6.4 of Chapter V and formula (3.3) on p. 106]).

Claim: For $\psi \in C^2([-1,0],\mathbb{R})$ and for all $\mu \in \Sigma_c$, one has

$$|\,\mathrm{pr}_\mu\, \psi| \le \frac{(3\pi/4)|\psi| + 4|\psi'| + e^2|\psi''|}{|\mu(1+\mu)|}.$$

Proof. If $\mu \in \Sigma_c$, then $\mu = ce^{-\mu}$, so $e^\mu = c/\mu$. Using partial integration twice, we calculate

$$\mu \int_{-1}^{0} e^{-\mu s}\psi(s)ds = -\psi(0) + \frac{c}{\mu}\psi(-1) + \frac{1}{\mu}\left[\frac{c}{\mu}\psi'(-1) - \psi'(0)\right] + \frac{1}{\mu}\int_{-1}^{0} e^{-\mu s}\psi''(s)\,ds.$$

It follows that

$$|\,\mathrm{pr}_\mu\, \psi| \le \frac{1}{|1+\mu|}\left[\frac{|c|}{|\mu|}|\psi| + \frac{1}{|\mu|}\left(\frac{|c|}{|\mu|}+1\right)|\psi'| + \frac{1}{|\mu|}\int_{-1}^{0}|e^{-\mu s}|\,ds\cdot|\psi''|\right].$$

Since $c \in [-3\pi/4, \pi/4]$, we know from Proposition 2.2 that

$$\Sigma_c = \{\rho_k(c) \pm i\omega_k(c) \mid k \in \mathbb{N}_0\},$$

that $\omega_k(c) \ge 2k\pi$ for $k \ge 1$, and that $\omega_0(c) \ge \pi/4$. In particular, $|c|/|\mu| \le (3\pi/4)/(\pi/4) = 3$ for $\mu \in \Sigma_c$. Further, it follows from Proposition 2.2(d) that for $s \in [-1,0]$ we have $|e^{-\mu s}| \le e^{|\rho_0(c)|} \le e^2$. Thus we obtain

$$|\,\mathrm{pr}_\mu\, \psi| \le \frac{1}{|1+\mu|}\left[\frac{3\pi/4}{\mu}|\psi| + \frac{1}{|\mu|}4|\psi'| + \frac{e^2}{|\mu|}|\psi''|\right].$$

The claim is proved. For $\psi \in C^2([-1,0],\mathbb{R})$, we have

$$\begin{aligned}
\sum_{\mu\in\Sigma_c} |\,\mathrm{pr}_\mu\, \psi| &\le 2\cdot \sum_{\substack{\mu\in\Sigma_c,\\ \mathrm{Im}\,\mu>0}} |\,\mathrm{pr}_\mu\, \psi| \\
&\le 2\sum_{k=0}^{\infty}[(3\pi/4)|\psi| + 4|\psi'| + e^2|\psi''|]\frac{1}{|\omega_k(c)|(|\omega_k(c)|+1)} \\
&\le 2\left(\frac{3\pi}{4}|\psi| + 4|\psi'| + e^2|\psi''|\right)\left\{\frac{1}{\pi/4(1+\pi/4)} + \sum_{k=1}^{\infty}\frac{1}{(2k\pi)^2}\right\} \\
&\le 2\left(\frac{3\pi}{4}|\psi| + 4|\psi'| + e^2|\psi''|\right)\left\{\frac{1}{(3/4)\cdot(7/4)} + \frac{1}{4\pi^2}\frac{\pi^2}{6}\right\} \\
&\le 2\left(\frac{3\pi}{4}|\psi| + 4|\psi'| + e^2|\psi''|\right)\frac{17}{21} \le \frac{19/2}{2}\frac{17}{21}|\psi| + \frac{8\cdot 17}{21}|\psi'| + \frac{15\cdot 17}{21}|\psi''| \\
&\le 4|\psi| + 7|\psi'| + 13|\psi''|.
\end{aligned}$$

Now we obtain from the series expansion, and from $|e^{\mu t}| \le e^{\rho_0(c)t}$ for $\mu \in \Sigma_c$, that

$$|x^{c,\psi,0}(t)| \le \sum_{\mu\in\Sigma_c} |\operatorname{pr}_\mu \psi| e^{\rho_0(c)t} \le (4|\psi| + 7|\psi'| + 13|\psi''|)e^{\rho_0(c)t}.$$

Assertion (a) is proved.

(b) We know from Proposition 2.2(d) that $|\rho_0(c)| \le 2$, and hence we have

$$e^{|\rho_0(c)|} \le e^2. \tag{2.5}$$

Let $\varphi \in \mathbf{J}$. For $t \in [0,1]$, we have (using (2.1))

$$\begin{aligned} |x^{c,\varphi,0}(t)| &\le |\varphi|(1+|c|t) \le |\varphi|(1+|c|)e^{-\rho_0(c)t}e^{\rho_0(c)t} \\ &\le |\varphi|(1+|c|)e^2 e^{\rho_0(c)t}. \end{aligned} \tag{2.6}$$

Moreover, one has $x_1^{c,\varphi,0} \in C^1$, although $\dot{x}^{c,\varphi,0}$ may have a jump discontinuity at 1.

Similarly, we have for $t \in [1,2]$

$$|x^{c,\varphi,0}(t)| \le |\varphi|(1+|c|)^2 e^{-\rho_0(c)t}e^{\rho_0(c)t} \le |\varphi|(1+|c|)^2 e^4 e^{\rho_0(c)t}. \tag{2.7}$$

Set $\psi := x_2^{c,\varphi,0}$; then $\psi \in C^2$, since $x_1^{c,\varphi,0} \in C^1$, and we have

$$|\psi| \le (1+|c|)^2|\varphi|, \qquad |\psi'| \le |c| \cdot |x_1^{c,\varphi,0}| \le |c|(1+|c|)|\varphi|, \qquad |\psi''| \le c^2|\varphi|. \tag{2.8}$$

Using part (a), and inequality (2.5) for the last step, we obtain for $t \ge 2$

$$\begin{aligned} |x^{c,\varphi,0}(t)| &= |x^{c,\psi,2}(t)| = |x^{c,\psi,0}(t-2)| \\ &\le (4|\psi| + 7|\psi'| + 13|\psi''|)e^{\rho_0(c)(t-2)} \\ &\le [4(1+|c|)^2|\varphi| + 7|c|(1+|c|)|\varphi| + 13|c|^2|\varphi|]e^{\rho_0(c)(t-2)} \\ &\le [4 + 15|c| + 24|c|^2]e^4 e^{\rho_0(c)t}|\varphi|. \end{aligned}$$

We see from (2.6) and (2.7) that this estimate also holds for $t \in [0,2]$. The assertion of (b) now follows from $|c| \le 3\pi/4$. □

Next, we need some preparations concerning the inhomogeneous linear equation

$$\dot{x}(t) = a(t)x(t-1) + h(t). \tag{a,h}$$

We assume that a and h are continuous on an interval $[\tau, T]$. For $t \in [\tau, T]$ we define a segment $\hat{h}(t) \in \mathbf{J}$ by setting

$$\hat{h}(t)(\theta) := \begin{cases} h(t), & \theta = 0, \\ 0, & \theta \in [-1, 0). \end{cases}$$

Note that $|\hat{h}(t) - \hat{h}(s)| = |h(t) - h(s)|$ for $s, t \in [\tau, T]$, so that the map $\hat{h} : [\tau, T] \longrightarrow (\mathbf{J}, |\,|)$, $t \mapsto \hat{h}(t)$ is continuous.

Recall the notation $x^{a,\psi,s}$ for the solution of $\dot{y}(t) = a(t)y(t-1)$ starting with ψ at time s. We now see from continuity of $\hat{h}$ and from Proposition 2.1(b) that, for $t \in [\tau, T]$, the function $[\tau, T] \ni s \mapsto x^{a,\hat{h}(s),s}(t)$ is continuous. In particular, the integral $\int_\tau^t x^{a,\hat{h}(s),s}(t)ds$ exists, simply as a Riemann integral of a continuous function.

The following result is known but stated in somewhat different notation in the literature (see, e.g., [8, Theorem 16.3 and formula (16.17)] or [9, formula (2.2), p. 173]).

LEMMA 2.4 (variation of constants). *For $\psi \in \mathbf{C}$, the solution $x^{a,h,\psi,\tau}$ of (a,h) with $x^{a,h,\psi,\tau}_\tau = \psi$ satisfies*

$$x^{a,h,\psi,\tau}(t) = x^{a,\psi,\tau}(t) + \int_\tau^t x^{a,\hat{h}(s),s}(t)ds \quad \text{for } t \geq \tau.$$

In the next section we use Lemma 2.4 to compare solutions of nonautonomous and nonlinear equations of the type

$$\dot{x}(t) = g(t, x(t-1)), \tag{g}$$

to solutions of constant coefficient equations. This is done on successive intervals, where we "adapt" the constant coefficient to the coefficient $a(t) := \partial_2 g(t,0)$ of the linearization of (g) on each interval. Thus we obtain Theorem 3.2, which provides upper estimates on the values of solutions close to zero in terms of exponential functions. Here the real part $\rho_0(a(r))$ of the leading eigenvalue for the "frozen" coefficient equation $\dot{y}(t) = a(r)y(t-1)$ enters in a formula which resembles the expression for the solution of a scalar linear ODE with varying coefficient.

Theorem 3.2 is applied to the example equation $\dot{x}(t) = (-\pi/4 - \varepsilon t)\arctan(x(t-1))$ in section 4, where it serves to control the motion near zero.

3. Nonlinear nonautonomous equations. We consider (g) from above, where we assume that $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is continuous, and has two continuous derivatives w.r. to the second argument. Further, we assume that for all t one has $g(t,0) = 0$ and that $|\partial_2^2 g|$ has a finite supremum which we denote by $\|\partial_2^2 g\|$. For a bounded function a on an interval $[s,t]$, we use the notation

$$V_a(s,t) := \sup_{\tau\in[s,t]} a(\tau) - \inf_{\tau\in[s,t]} a(\tau).$$

Using Lemma 2.4, we can now obtain an estimate on solutions of nonautonomous and nonlinear equations.

LEMMA 3.1. *Let $\varphi \in \mathbf{C}$, $T \geq 1$, $s \in \mathbb{R}$, and let $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be as above. Set $a(t) := \partial_2 g(t,0)$ for $t \in \mathbb{R}$. Assume that $c \in a([s, s+T]) \cap [-3\pi/4, -\pi/4]$. Set $V := V_a(s, s+T)$, and with K from Proposition 2.3 set*

$$K_V := \max\{K, 1 + 3\pi/4 + V\}, \quad L_V := K_V e^2.$$

Let $x : [s-1, \infty) \to \mathbb{R}$ be the solution of (g) with $x_s = \varphi$, and assume that $\tau \in [s, s+T]$ and $\xi \geq 0$ are such that $|x_t| \leq \xi$ for all $t \in [s,\tau]$. Then, for all $t \in [s,\tau]$, one has

$$|x_t| \leq L_V|\varphi| \exp[(\rho_0(c) + L_V V + L_V\|\partial_2^2 g\|\xi/2) \cdot (t-s)].$$

Proof. For $t \in [s, s+T]$, there exists $r_t \in (0,1)$ such that

$$\begin{aligned} g(t, x(t-1)) &= \partial_2 g(t,0)x(t-1) + [\partial_2^2 g(t, r_t x(t-1))/2]x(t-1)^2 \\ &= cx(t-1) + (\partial_2 g(t,0) - c)x(t-1) + [\partial_2^2 g(t, r_t x(t-1))/2]x(t-1)^2 \\ &= cx(t-1) + (a(t) - c)x(t-1) + [\partial_2^2 g(t, r_t x(t-1))/2]x(t-1)^2. \end{aligned}$$

Thus, with $h(t) := (a(t)-c)x(t-1) + [\partial_2^2 g(t, r_t x(t-1))/2]x(t-1)^2$ for $t \in [s, s+T]$, one has for these t

$$\dot{x}(t) = cx(t-1) + h(t).$$

Further, for $t \in [s, \tau]$ one has

$$|\hat{h}(t)| \le V|x(t-1)| + (\|\partial_2^2 g\|/2)x(t-1)^2 \le [V + \|\partial_2^2 g\|\xi/2] \cdot |x(t-1)|.$$

It follows from Lemma 2.4 and from Proposition 2.3(b) that, for $t \in [s, \tau]$,

$$\begin{aligned}
|x(t)| &= \left| x^{c,\varphi,s}(t) + \int_s^t x^{c,\hat{h}(\sigma),\sigma}(t)\, d\sigma \right| \\
&\le K|\varphi| \exp[\rho_0(c)(t-s)] + K \int_s^t \exp[\rho_0(c)(t-\sigma)]|\hat{h}(\sigma)|\, d\sigma \\
&\le K \left\{ |\varphi| \exp[\rho_0(c)(t-s)] + [V + \|\partial_2^2 g\|\xi/2] \int_s^t \exp[\rho_0(c)(t-\sigma)]|x(\sigma-1)|\, d\sigma \right\}.
\end{aligned}$$

Set $W := [V + (\|\partial_2^2 g\|\xi/2)]$. If now $t \in [s+1, \tau]$ (in case $s+1 \le \tau$) and $\theta \in [-1, 0]$, then

$$\begin{aligned}
|x(t+\theta)| &\le K \left\{ |\varphi| \exp[\rho_0(c)(t+\theta-s)] + W \int_s^{t+\theta} \exp[\rho_0(c)(t+\theta-\sigma)]|x(\sigma-1)|\, d\sigma \right\} \\
&\le K \exp(|\rho_0(c)|) \left\{ |\varphi| \exp[\rho_0(c)(t-s)] + W \int_s^{t+\theta} \exp[\rho_0(c)(t-\sigma)]|x_\sigma|\, d\sigma \right\} \\
&\le K \exp(|\rho_0(c)|) \left\{ |\varphi| \exp[\rho_0(c)(t-s)] + W \int_s^{t} \exp[\rho_0(c)(t-\sigma)]|x_\sigma|\, d\sigma \right\}.
\end{aligned}$$

Hence, for $t \in [s+1, \tau]$, it follows trivially that with $K_V := \max\{K, 1 + 3\pi/4 + V\}$ we have

(3.1)
$$|x_t| \le K_V \exp(|\rho_0(c)|) \left\{ |\varphi| \exp[\rho_0(c)(t-s)] + W \int_s^t \exp[\rho_0(c)(t-\sigma)]|x_\sigma|\, d\sigma \right\}.$$

For $t \in [s, s+1] \cap [s, \tau]$, we obtain (using the differential equation and the definition of K_V) that

$$\begin{aligned}
|x_t| &\le |\varphi| + \int_{-1}^0 (|c| + V)|\varphi|\, ds = |\varphi|(1 + |c| + V) \\
&\le |\varphi|(1 + 3\pi/4 + V) \le K_V|\varphi|.
\end{aligned}$$

The right-hand side of (3.1) is, for $t \in [s, s+1]$, obviously bounded below by $K_V|\varphi|$. Hence, (3.1) holds also for $t \in [s, s+1]$.

Now, setting $y(t) := \exp[-\rho_0(c)t]|x_t|$ for $t \in [s, \tau]$, we obtain from (3.1) and (2.5) that

$$\begin{aligned}
y(t) &\le K_V \exp(|\rho_0(c)|) \left\{ |\varphi| \exp[-\rho_0(c)s] + W \int_s^t y(\sigma)\, d\sigma \right\} \\
&\le K_V e^2 \left\{ |\varphi| \exp[-\rho_0(c)s] + W \int_s^t y(\sigma)\, d\sigma \right\} \\
&= L_V \left\{ |\varphi| \exp[-\rho_0(c)s] + W \int_s^t y(\sigma)\, d\sigma \right\}.
\end{aligned}$$

It follows from Gronwall's lemma that for $t \in [s, \tau]$ one has

$$y(t) \le L_V|\varphi| \exp[-\rho_0(c)s] \exp[L_V W(t-s)].$$

Hence we conclude

$$\begin{aligned} |x_t| &\le L_V|\varphi| \exp[\rho_0(c)(t-s)] \exp[L_V W(t-s)] \\ &= L_V|\varphi| \exp[(\rho_0(c) + L_V V + L_V \|\partial_2^2 g\|\xi/2) \cdot (t-s)]. \end{aligned}$$ □

We are now prepared for the proof of a delayed loss of stability estimate for nonlinear nonautonomous equations of type (g). Again, we restrict attention to the case where $\partial_2 g(\cdot, 0)$ takes values in $[-3\pi/4, -\pi/4]$. Recall the definition of $V_a(s,t)$ for $s \le t$.

THEOREM 3.2. *Let $t_- \in \mathbb{R}$, let g be as above, and assume that the function defined by $a(t) := \partial_2 g(t, 0)$ takes values in $[-3\pi/4, -\pi/4]$. Assume that there exists $T \ge 1$ and $V \ge 0$ such that one has for all $s \ge t_-$*

$$V_a(s, s+T) \le V. \tag{3.2}$$

Let $\varphi \in \mathbf{J}$, and let $x : [t_- - 1, \infty) \to \mathbb{R}$ be the solution of (g) with $x_{t_-} = \varphi$. Assume that $t_+ \ge t_-$ and $\xi \ge 0$ are such that

$$\forall t \in [t_-, t_+] : |x_t| \le \xi.$$

Define L_V as in Lemma 3.1*, and set*

$$C := C(V, T, \xi) := L_V V + L_V \|\partial_2^2 g\|\xi/2 + \log(L_V)/T.$$

Finally, for $t, s \in \mathbb{R}$, $t \ge s \ge t_-$, set

$$u(t,s) := \exp\left[\int_s^t (\rho_0(a(r)) + C)\, dr\right].$$

(a) *Then one has for all $t \in [t_-, t_+]$*

$$|x_t| \le |\varphi| L_V u(t, t_-).$$

(b) *With $c_- := \frac{4(\pi-2)}{3\pi^2}$, $c_+ := \frac{4(\pi+2)}{\pi^2}$, the following estimates hold:*

If $t, s \in [t_-, t_+]$, $s \le t$, and $a(\cdot) \ge -\pi/2$ on $[s,t]$, then

$$u(t,s) \le \exp\left[\int_s^t (-c_-|a(s) + \pi/2| + C)\, ds\right].$$

If $t, s \in [t_-, t_+]$, $s \le t$, and $a(\cdot) \le -\pi/2$ on $[s,t]$, then

$$u(t,s) \le \exp\left[\int_s^t (c_+|a(s) + \pi/2| + C)\, ds\right].$$

Remarks. 1. The first estimate in (b) implies (not necessarily monotonous) decay of $|x_t|$, as long as $a(s) \ge -\pi/2$ and $c_-|a(s) + \pi/2| \ge C$. One can expect the second inequality to hold only if the term $\|\partial_2^2 g\|\xi/2$ is small enough, i.e., if the solution x takes sufficiently small values. This is natural since the decay is an effect of the linearization at zero.

If one wants to obtain decay for "large" initial values φ, it is necessary to combine the estimate of Theorem 3.2(a) with different methods, as we do in the example in section 4. While the method of linearization close to zero is "universal," the possibilities to obtain a decay estimate far from zero depend both on the initial value and on the specific nonlinearity. We explain the basic idea for doing this in our example at the beginning of section 4.

2. If one obtains $|x_{t_0}| < |\varphi|$ for some $t_0 \in [t_-, t_+]$, then the second inequality in b) can be used to give a lower estimate for the length of the time interval on which $|x_t| \le |\varphi|$. Note that, as soon as $\rho_0(a(t)) > 0$ (in fact, as soon as $\rho_0(a(t)) + C > 0$), the function $t \mapsto u(t, t_-)$ starts to increase exponentially with t, and typically one has to expect that the solution x does the same. We do not prove this rigorously, but give a heuristic argument: For the constant coefficient equation with an unstable (complex) leading eigenvalue, all solutions, except the ones starting in a subspace S of codimension 2, exhibit the growth associated with that unstable eigenvalue. The functions in S are rapidly oscillating (have two or more zeroes per time unit). The same behavior is to be expected from equations with slowly varying coefficient. This could be proved by comparison to the constant coefficient case on successive "long" intervals, which is also our method of proving Theorem 3.2. In the example of section 4, the solution under consideration is slowly oscillating (has zeroes further apart than 1) and hence will certainly exhibit growth behavior corresponding to the leading eigenvalue of the constant coefficient approximation.

Proof of Theorem 3.2. Set $\tilde{C} := \tilde{C}(V, \xi) := L_V V + L_V \|\partial_2^2 g\| \xi/2$. For $t \ge t_-$, set $\eta(t) := \exp[\int_{t_-}^t (\rho_0(a(s)) + \tilde{C})\, ds] = \exp[\int_{t_-}^t (\rho_0(a(s))\, ds] \exp[\tilde{C}(t - t_-)]$. Consider φ and x as in the theorem.

Claim. If $t \in [t_- + (j-1)T, t_- + jT]$ for some $j \in \mathbb{N}$, and $t \le t_+$, then

$$|x_t| \le |\varphi| L_V^j \eta(t).$$

Proof. (induction on j.) The case $j = 1$: Assume $t \in [t_-, t_- + T]$. From the mean value theorem, there exists $\tau = \tau(t) \in [t_-, t]$ such that

$$\int_{t_-}^t \rho_0(a(s)) ds = (t - t_-)\rho_0(a(\tau)).$$

Applying Lemma 3.1 with $s := t_-$, $\tau := T$, $c := a(\tau)$, one obtains

$$\begin{aligned} |x_t| &\le L_V |\varphi| \exp[(\rho_0(c) + L_V V + L_V \|\partial_2^2 g\| \xi/2)(t - t_-)] \\ &= L_V |\varphi| \exp\left[\int_{t_-}^t (\rho_0(a(s)) + \tilde{C})\, ds\right] \\ &= |\varphi| L_V \eta(t), \end{aligned}$$

which is the assertion for $j = 1$.

Assume now that the assertion holds for some $j \in \mathbb{N}$ and that $t \in [t_- + jT, t_- + (j+1)T]$, $t \le t_+$. Set $\psi := x_{t_- + jT}$. Then the induction hypotheses gives $|\psi| \le |\varphi| L_V^j \eta(t_- + jT)$. From the case $j = 1$, applied with $t_- + jT$ in place of t_-, one obtains for the solution $y : [t_- + jT - 1, \infty) \to \mathbb{R}$ of (g) with $y_{t_- + jT} = \psi$ that

$$|y_t| \le |\psi| L_V \exp\left[\int_{t_- + jT}^t (\rho_0(a(s)) + \tilde{C})\, ds\right].$$

Together with the estimate on $|\psi|$, we conclude

$$|x_t| = |y_t| \le |\varphi| L_V^j \eta(t_- + jT) L_V \exp\left[\int_{t_-+jT}^{t} (\rho_0(a(s)) + \tilde{C})\, ds\right]$$
$$= |\varphi| L_V^{j+1} \eta(t).$$

The claim is proved.

Now let $t \in [t_-, t_+]$, and set $j := \min\{n \in \mathbb{N} \mid t_- + nT > t\}$. Then $t_- + (j-1)T \le t < t_- + jT$, and from the above claim we get $|x_t| \le |\varphi| L_V^j \eta(t)$. Note that

$$L_V^{j-1} = \exp\left[\frac{(j-1)T\log(L_V)}{T}\right] \le \exp\left[(t-t_-)\frac{\log(L_V)}{T}\right] = \exp\left[\int_{t_-}^{t} \frac{\log(L_V)}{T} ds\right].$$

Recalling the definition of η, and noting that $\tilde{C} + \log(L_V)/T = C$, we obtain

$$|x_t| \le |\varphi| L_V \exp\left[\int_{t_-}^{t} (\rho_0(a(s)) + \tilde{C} + \log(L_V)/T)\, ds\right] = |\varphi| L_V u(t, t_-),$$

that is, assertion (a). Assertion (b) follows from the estimates on ρ_0 from Proposition 2.2(c). □

4. An example. For $\varepsilon \in (0, 0.01]$, we set

$$g(t,x) := (-\pi/4 - \varepsilon t)\arctan(x),$$

and we consider the solution $x : [-1, \infty) \longrightarrow \mathbb{R}$ of (g) with the constant function equal to 1 as initial segment. (The dependence of all objects on ε is not denoted.) Note that $a(t) := \partial_2 g(t, 0)$ satisfies $a(t) \in [-3\pi/4, -\pi/4]$ as long as $t \in [0, \pi/2\varepsilon]$. Further, for these t and for $y \in \mathbb{R}$, one has

$$|\partial_2^2 g(t,y)| \le |-3\pi/4| \sup_{z\in\mathbb{R}} |2z/(1+z^2)^2| \le 2 \cdot 3\pi/4 \le 5.$$

(It is not essential that these properties do not hold for t outside the interval $[0, \pi/2\varepsilon]$, in which we will be interested.)

As already remarked, we need to prove decay of our solution x from the constant value 1 to values near zero first, before Theorem 3.2 becomes applicable to the motion near zero. We briefly explain our basic approach in achieving this: The solution x is oscillating about zero, with zeroes z_i, $i = 1, 2, \dots$, and extremal values m_i occurring at the times $z_i + 1$. It is easy to obtain an estimate of the form $|g(t,y)| \le q|y|$ for our g, where $q \in (0,1)$. We conclude that the m_i form a geometrically decreasing sequence: $m_{i+1} \le q m_i$. This implies exponential decay of $|x(t)|$, if the z_i are not too far apart. However, since it would be difficult to obtain an upper bound on $z_{i+1} - z_i$, we also consider the theoretically possible (although practically not occurring) case of "long" distances between z_i and z_{i+1} (in Proposition 4.3 below). On such intervals, the negative feedback forces the solution to decay monotonously, and we prove an exponential estimate also in this case. While the first argument is based on the estimate $|\arctan(y)| \le |y|$ (the feedback is "weak" enough), this second argument uses the estimate $|\arctan(y)| \ge (\pi/4)|y|$ if $|y| \le 1$ (the negative feedback is "strong" enough).

We carry out the details now.

PROPOSITION 4.1. *The solution x is slowly oscillating; that is, there exists a sequence $(z_1, z_2, \dots)$ in $\mathbb{R}$ such that $0 < z_1 < z_2 < \cdots$ and such that the z_i are precisely the zeroes of x, and $z_{i+1} - z_i > 1$ for all $i \in \mathbb{N}$. The extrema of x on $(0, \infty)$ occur at the times $\mu_i := z_i + 1 \in (z_i, z_{i+1})$, so we have*

$$z_1 < \mu_1 = z_1 + 1 < z_2 < \mu_2 = z_2 + 1 < \cdots .$$

Further, one has $z_1 \le 2$.

Proof. Assume that x has no zero on some interval of the form $[t_0, \infty)$. Then the negative feedback property $g(t, y)y < 0$ $(t > 0, y \in \mathbb{R} \setminus \{0\})$ implies that $x(t) \longrightarrow 0$ $(t \longrightarrow \infty)$, so there exists $t_1 \ge t_0$ with $|x(\cdot)| \le 0.1$ on $[t_1, \infty)$. Now setting $\alpha(t) := \int_0^1 \partial_2 g(t, sx(t-1))\, ds$, the function x satisfies $\dot{x}(t) = \alpha(t)x(t-1)$ for $t \ge t_1$, and for these t one has

$$\alpha(t) \le -(\pi/4) \min_{|y| \le 0.1} \arctan'(y) = -(\pi/4) \cdot 100/101 < -\exp(-1).$$

We can now apply Theorem 8 in [6] (with $n := 1$, $r := 1$, $\eta(t, -1) := 0$, $\eta(t, \theta) := \alpha(t)$ for $\theta \in (-1, 0]$, and with $q(t, \theta) := -\eta(t, \theta)$); in particular, the last inequality shows that condition (A4) of that theorem is satisfied. It follows that x has infinitely many zeroes on $[t_1, \infty)$, in contradiction to our assumption.

We know now that x must have infinitely many zeroes. It follows from the fact that the segment x_0 has no zero, and from the fact that the zero-counting Liapunov functional used in [11] does not increase in time, that x is slowly oscillating (see [11], Theorem 2.1). The assertion about extrema is now clear, in view of the differential equation.

We now prove $z_1 \le 2$: On $[0, 1]$, we have

$$\dot{x}(t) = (-\pi/4 - \varepsilon t) \arctan(1) \le (-\pi/4)(\pi/4) = -\pi^2/16,$$

and hence $x(1) \le 1 - \pi^2/16$. On the other hand, for $t \in [0, 1]$, one obtains (using $\varepsilon \le 0.01$) that $\dot{x}(t) \ge (-\pi/4 - \varepsilon)(\pi/4) \ge -10/16 = -5/8$, so $x(t) \ge 1 - (5/8)t$ for these t. It follows from $|\arctan(y)| \ge |(\pi/4)y|$ if $|y| \le 1$ that for $t \in [1, 2]$ one has

$$\dot{x}(t) \le -(\pi/4)(\pi/4)[1 - (5/8)(t-1)].$$

Hence, integrating, we obtain

$$\begin{aligned} x(2) &\le 1 - (\pi^2/16) - (\pi^2/16)[1 - (5/8)(1/2)] = 1 - (\pi^2/16) - (\pi^2/16) \cdot (11/16) \\ &= 1 - (27\pi^2/256) \le 1 - 27 \cdot 9.5/256 = 1 - 256.5/256 < 0, \end{aligned}$$

and consequently x has a first zero z_1 in $[1, 2]$. □

Set $m_i := |x(\mu_i)|$ for $i \in \mathbb{N}$; then $m_i = \max_{t \in [z_i, z_{i+1}]} |x(t)|$. We first focus attention on the time interval $(0, \pi/16\varepsilon]$. Let $J \in \mathbb{N}$ be such that the extrema of x in this interval occur at the times $\mu_1, \dots, \mu_J$. The following estimate exploits the fact that for t in $[0, \pi/16\varepsilon]$ one has $|g(t, y)| \le q|y|$ $(y \in \mathbb{R})$ with some $q \in (0, 1)$.

PROPOSITION 4.2. *For $t \in [0, z_{J+1}]$ one has $|x(t)| \le 1$. Further, with $q := 5\pi/16$, one has*

$$m_{i+1} \le q m_i \quad \text{if } i \in \{1, \dots, J-1\}.$$

Proof. Note that $|\arctan(y)| \le |y|$ for $y \in \mathbb{R}$. As long as $t \le \pi/16\varepsilon$, we thus have

$$|g(t, y)| \le (\pi/4 + \pi/16)|y| = 5\pi/16|y| = q|y|.$$

Since $|x(\cdot)| \le 1$ on $[0, z_1]$, we have $m_1 \le q < 1$. Further, if $i \in \{1, \dots, J-1\}$, we obtain (using $z_{i+1} - z_i > 1$) that

$$m_{i+1} = \left| \int_{z_{i+1}}^{z_{i+1}+1} g(s, x(s-1))\, ds \right| \le \int_{z_{i+1}-1}^{z_{i+1}} q|x(s)|ds \le q m_i.$$

Together with $|x(\cdot)| \le 1$ on $[0, \mu_1]$ and the fact that $|x(\cdot)|$ decreases on $[\mu_J, z_{J+1}]$, it follows that $|x(\cdot)| \le 1$ on $[0, z_{J+1}]$. ∎

Next, we give a decay estimate for the case that $\mu_i - \mu_{i-1}$ is "large."

PROPOSITION 4.3. *Assume $i \in \{2, \dots, J+1\}$ and $\mu_{i-1} + 1 \le z_i$. Then for all $j \in \mathbb{N}_0$ with $\mu_{i-1} + j \le z_i$ one has*

$$|x(\mu_{i-1} + j)| \le q^{j-1} m_{i-1}.$$

Proof. The estimate is trivial for $j = 0$. Since $|x(\cdot)|$ decreases on $[\mu_{i-1}, z_i]$ and $m_{i-1} \le 1$, we have $|x(\cdot)| \le 1$ on $[\mu_{i-1}, z_i]$, and $|x(\mu_{i-1}+1)| \le m_{i-1}$. Hence the assertion holds for $j = 1$. Now if $j \in \mathbb{N}$ and $[\mu_{i-1} + j, \mu_{i-1} + j + 1] \subset [\mu_{i-1}, z_i]$, we obtain (using $|\arctan(y)| \ge (\pi/4)|y|$ if $|y| \le 1$, and the monotonicity of $|x(\cdot)|$ on $[\mu_{i-1} + j - 1, \mu_{i-1} + j]$) that

$$\begin{aligned}
|x(\mu_{i-1} + j + 1)| &= \left| x(\mu_{i-1} + j) + \int_{\mu_{i-1}+j}^{\mu_{i-1}+j+1} g(s, x(s-1))\, ds \right| \\
&\le |x(\mu_{i-1} + j)| - \min_{s \in [\mu_{i-1}+j-1, \mu_{i-1}+j]} |g(s+1, x(s))| \\
&\le |x(\mu_{i-1} + j)| - (\pi/4)|\arctan(x(\mu_{i-1} + j))| \\
&\le |x(\mu_{i-1} + j)| - (\pi^2/16)|x(\mu_{i-1} + j)| \\
&= [(16 - \pi^2)/16]|x(\mu_{i-1} + j)| \le q|x(\mu_{i-1} + j)|.
\end{aligned}$$

For $j \in \mathbb{N}$ with $\mu_{i-1} + j \le z_i$, it follows inductively that

$$|x(\mu_{i-1} + j)| \le q^{j-1}|x(\mu_{i-1} + 1)| \le q^{j-1} m_{i-1}. \qquad ∎$$

Proposition 4.2 above relates the value m_i to the index i, but not to the time μ_i at which it occurs. This is achieved in the next result.

PROPOSITION 4.4. *With the negative number $\lambda := \log(q)/4$, one has*

$$m_j \le q^{-1} \exp(\lambda \mu_j) \quad (j = 1, \dots, J).$$

Proof. Let $i \in \{2, \dots, J\}$. If $z_i - \mu_{i-1} \le 2$, then $\mu_i - \mu_{i-1} \le 3 < 4$ and

$$m_i/m_{i-1} \le q = \exp(4\lambda) \le \exp(\lambda(\mu_i - \mu_{i-1})). \tag{4.1}$$

Consider now the case $z_i - \mu_{i-1} > 2$. Then, setting

$$j_1 := \max\{j \in \mathbb{N} \mid \mu_{i-1} + j + 1 \le z_i\},$$

we obtain from Proposition 4.3 that

$$|x(\mu_{i-1} + j_1)| \le q^{j_1 - 1} m_{i-1}.$$

Note that $[z_i - 1, z_i] \subset [\mu_{i-1} + j_1, z_i]$, and hence $|x(t)| \le |x(\mu_{i-1} + j_1)|$ for $t \in [z_i - 1, z_i]$. We infer from the differential equation that

$$m_i \le q|x(\mu_{i-1} + j_1)| \le q q^{j_1 - 1} m_{i-1} = q^{j_1} m_{i-1}.$$

Now, from the definition of j_1,

$$\mu_i - \mu_{i-1} = z_i + 1 - \mu_{i-1} \le \mu_{i-1} + j_1 + 3 - \mu_{i-1} = j_1 + 3 \le 4j_1,$$

and thus

$$m_i/m_{i-1} \le q^{j_1} = \exp(\lambda \cdot 4j_1) \le \exp(\lambda(\mu_i - \mu_{i-1})),$$

and we see that (4.1) also holds in the second case.

We conclude that for $j \in \{1, \dots, J\}$ one has

$$\begin{aligned} m_j &= m_1 \prod_{i=2}^{j} (m_i/m_{i-1}) \le m_1 \prod_{i=2}^{j} \exp(\lambda(\mu_i - \mu_{i-1})) \\ &= m_1 \exp[\lambda(\mu_j - \mu_1)] = m_1 \exp(-\lambda\mu_1) \exp(\lambda\mu_j), \end{aligned}$$

where the product is to be read as 1 if $j = 1$. Since $m_1 \le 1$ (Proposition 4.2) and $\mu_1 = z_1 + 1 \le 3 < 4$ (Proposition 4.1), it follows that $m_j \le \exp(-4\lambda)\exp(\lambda\mu_j) = q^{-1}\exp(\lambda\mu_j)$. ∎

We can now obtain an exponential decay estimate for x (which is not based on linearization at zero) for the time interval $[0, \pi/16\varepsilon]$.

COROLLARY 4.5. *For $t \in [0, \pi/16\varepsilon]$, one has $|x_t| \le 2q^{-3}\exp(\lambda t)$.*

Proof. 1. For $t \in [0, \mu_1]$, one has $|x_t| \le 1$, and $\mu_1 \le 3 < 4$ implies

$$2q^{-3}\exp(\lambda t) \ge 2q^{-3}\exp(4\lambda) = 2q^{-2} > 1,$$

so the assertion is true for these t.

2. Let $t \in [\mu_1, \pi/16\varepsilon]$. There exists $i \in \{2, \dots, J+1\}$ with $t \in [\mu_{i-1}, \mu_i]$.

Case 1: $t \le \mu_{i-1} + 2$. We have $|x(s)| \le m_{i-1}$ for $s \in [z_{i-1}, z_i]$, and

$$|\dot{x}(s)| \le |(-\pi/4 - \pi/16)| m_{i-1} = q m_{i-1} \le m_{i-1}$$

for $s \in [z_i, t]$ if $t \ge z_i$. In this case, $t - z_i \le \mu_{i-1} + 2 - z_i \le 2$, so $|x(s)| \le 2m_{i-1}$ for $s \in [z_i, t]$. With Proposition 4.4, it follows that

$$\begin{aligned} |x_t| &\le 2m_{i-1} \le 2q^{-1}\exp(\lambda\mu_{i-1}) = 2q^{-1}\exp(\lambda(\mu_{i-1} - t))\exp(\lambda t) \\ &\le 2q^{-1}\exp(-2\lambda)\exp(\lambda t) \le 2q^{-2}\exp(\lambda t). \end{aligned}$$

Case 2: $t > \mu_{i-1} + 2$. Then $z_i = \mu_i - 1 \ge t - 1 \ge \mu_{i-1} + 1$. Setting $j_1 := \max\{j \in \mathbb{N} \mid \mu_{i-1} + j \le \min\{t, z_i\}\}$, we obtain from Proposition 4.3 that

$$|x(\mu_{i-1} + j_1)| \le q^{j_1 - 1} m_{i-1}.$$

Subcase 2a: $t \le z_i$. Then $j_1 \ge 2$, and $t - (\mu_{i-1} + j_1) \le 1$, and it follows from Propositions 4.3 and 4.4 that

$$\begin{aligned} |x_t| &\le |x_{\mu_{i-1}+j_1}| = |x(\mu_{i-1} + j_1 - 1)| \le q^{j_1 - 2} m_{i-1} \\ &\le q^{j_1 - 2} q^{-1} \exp(\lambda\mu_{i-1}) = q^{-1}\exp[4\lambda(j_1 - 2)]\exp(\lambda\mu_{i-1}) \\ &\le q^{-1}\exp[\lambda(j_1 - 2)]\exp(\lambda\mu_{i-1}) = q^{-1}\exp(\lambda t)\exp[\lambda(j_1 - 2 + \mu_{i-1} - t)] \\ &\le q^{-1}\exp(-3\lambda)\exp(\lambda t) \le q^{-2}\exp(\lambda t). \end{aligned}$$

Subcase 2b: $t > z_i$. Then $t \in [z_i, \mu_i] = [z_i, z_i + 1]$. From Subcase 2a, applied to z_i, we obtain $|x_{z_i}| \le q^{-2}\exp(\lambda z_i)$. For $s \in [z_i, t]$, one has $\dot{x}(s) \le |(-\pi/4 - \pi/16)|\,|x_{z_i}| \le q q^{-2}\exp(\lambda z_i)$. It follows that

$$\begin{aligned} |x_t| &\le q^{-2}\exp(\lambda z_i) = q^{-2}\exp(\lambda t)\exp(\lambda(z_i - t)) \\ &\le q^{-2}\exp(-\lambda)\exp(\lambda t) \le q^{-3}\exp(\lambda t). \end{aligned}$$

From part 1 and the different cases of part 2, the asserted estimate is obtained. ∎

Combining the above estimate with the ones which were obtained from linearization at zero in Theorem 3.2, we can now provide a lower estimate on the length of the time interval on which $|x(t)| \le 1$. Since our solution x is slowly oscillating, exponential growth is to be expected (and certainly seen in numerical simulation) after the coefficient crosses $-\pi/2$, and the solution will also reach the amplitude 1 of the starting segment again (at some time bounded below by our estimate). We do not pursue a formal proof of this; compare the second remark following Theorem 3.2.

With c_- and c_+ from Theorem 3.2, we set $c_1 := \lambda\pi/16 - c_+\pi^2/16 - 5c_-\pi^2/512$ and $c_2 := c_+\pi/4 + c_-\pi/32$. Note that $c_1 < 0 < c_2$.

THEOREM 4.6. *There exists $\varepsilon_0 \in (0, 0.01]$ such that for $\varepsilon \in (0, \varepsilon_0]$ the function $t \mapsto |x_t|$ decreases to values below ε on the interval $[0, \pi/4\varepsilon]$, and then reaches the value $\sqrt{\varepsilon}$ again not before the time $|c_1/2c_2\varepsilon|$. (In particular, $|x_t| \le 1$ on the interval $[0, c_1/2c_2\varepsilon]$.)*

Proof. With K from Proposition 2.3, we set $L := Ke^2$. There exists $\varepsilon_0 \in (0, 0.01]$ such that for $\varepsilon \in (0, \varepsilon_0]$ the following estimates hold:

$$2q^{-3}\exp(\lambda\pi/16\varepsilon) \le \varepsilon, \tag{4.2}$$

$$5L\sqrt{\varepsilon} \le c_-\pi/32, \tag{4.3}$$

$$L2q^{-3}\exp[(\lambda\pi/16 - c_-\pi^2/512)/\varepsilon)] \le \varepsilon, \tag{4.4}$$

$$|\log(q^3\sqrt{\varepsilon}/2L)| \le |c_1|/2\varepsilon. \tag{4.5}$$

Let now $\varepsilon \in (0, \varepsilon_0]$. We set $T := T(\varepsilon) := 1/\sqrt{\varepsilon}$. We then have for all $s \in \mathbb{R}$

$$V_a(s, s+T) = \varepsilon T = \sqrt{\varepsilon} \le 1.$$

It follows that with $V := \sqrt{\varepsilon}$ and with K_V, L_V as in Lemma 3.1, one has $K_V = K$ and $L_V = Ke^2 = L$. Further, we have $\log(L_V)/T = \sqrt{\varepsilon}\log(L)$.

We set $\xi := \sqrt{\varepsilon}$; then the constant $C = C(V, T, \xi)$ from Theorem 3.2 satisfies

$$C \le L\sqrt{\varepsilon} + 5L\sqrt{\varepsilon}/2 + \log(L)\sqrt{\varepsilon} \le 5L\sqrt{\varepsilon}.$$

From Corollary 4.5 and (4.2), we obtain that

$$|x_{\pi/16\varepsilon}| \le 2q^{-3}\exp(\lambda\pi/16\varepsilon) \le \varepsilon < \xi.$$

Now we set $t_- := \pi/16\varepsilon$, and $t_+ := \min\{\inf\{t > t_- \mid |x_t| > \xi\}, \pi/2\varepsilon\}$, and we apply Theorem 3.2. It follows that with $u(t, t_-)$ defined as in that theorem, one has

$$\forall t \in [t_-, t_+] : |x_t| \le L2q^{-3}\exp(\lambda\pi/16\varepsilon)u(t, t_-).$$

Next, we estimate $u(t, s)$ for t, s in different time intervals. Note that for $t \in \mathbb{R}$, one has $|a(t) + \pi/2| = |-\pi/4 + \pi/2 - \varepsilon t| = |\pi/4 - \varepsilon t|$. Thus, for $t \in [\pi/16\varepsilon, 3\pi/16\varepsilon]$, we have $|a(t) + \pi/2| \ge \pi/16$. It follows from Theorem 3.2(b) and (4.3) that, for these t,

$$\begin{aligned} u(t, \pi/16\varepsilon) &\le \exp\left[\int_{\pi/16\varepsilon}^{t} (-c_-\pi/16 + C)\, ds\right] \\ &\le \exp\left[\int_{\pi/16\varepsilon}^{t} (-c_-\pi/16 + 5L\sqrt{\varepsilon})\, ds\right] \\ &\le \exp\left[\int_{\pi/16\varepsilon}^{t} (-c_-\pi/32)\, ds\right] = \exp[(-c_-\pi/32)(t - \pi/16\varepsilon)]. \end{aligned} \tag{4.6}$$

With $t_0 := \pi/4\varepsilon$, we have $a(t_0) = -\pi/2$. For $t \in [3\pi/16\varepsilon, t_0]$, one has

$$\rho_0(a(t)) + C \le C \le 5L\sqrt{\varepsilon},$$

and for these t one has from the definition of $u(\cdot,\cdot)$ and from (4.3) that

$$\begin{aligned} u(t, 3\pi/16\varepsilon) &\le \exp[5L\sqrt{\varepsilon}(t - 3\pi/16\varepsilon)] \le \exp[5L\sqrt{\varepsilon}\pi/16\varepsilon] \\ &\le \exp[c_-\pi^2/512\varepsilon]. \end{aligned} \tag{4.7}$$

Combining (4.6) and (4.7), we see that

$$\begin{aligned} |x_{t_0}| &\le L2q^{-3}\exp[\lambda\pi/16\varepsilon]u(t_0, 3\pi/16\varepsilon)u(3\pi/16\varepsilon, \pi/16\varepsilon) \\ &\le L2q^{-3}\exp[\lambda\pi/16\varepsilon + c_-\pi^2/512\varepsilon - c_-(\pi/32)(\pi/8\varepsilon)] \\ &= L2q^{-3}\exp[(\lambda\pi/16 - c_-\pi^2/512)/\varepsilon]. \end{aligned}$$

Now (4.4) shows that $|x_{t_0}| \le \varepsilon < \xi$, in particular, $t_+ > t_0$.

Finally, for $t \in [t_0, t_+]$ we have $|a(t) + \pi/2| \le \pi/4$. Using part (b) of Theorem 3.2, together with the inequalities $C \le 5L\sqrt{\varepsilon}$ and (4.3), one sees that

$$u(t, t_0) \le \exp[(c_+\pi/4 + C)(t - t_0)] \le \exp[(c_+\pi/4 + c_-\pi/32)(t - \pi/4\varepsilon)]. \tag{4.8}$$

Combining the estimates (4.6), (4.7), and (4.8), we conclude that for $t \in [t_0, t_+]$ one has

$$\begin{aligned} |x_t| &\le L2q^{-3}\exp[(\lambda\pi/16 - c_-\pi^2/512)/\varepsilon \\ &\qquad - (c_+\pi/4 + c_-\pi/32)(\pi/4\varepsilon) + (c_+\pi/4 + c_-\pi/32)t] \\ &= L2q^{-3}\exp[(\lambda\pi/16 - c_+\pi^2/16 - 5c_-\pi^2/512)/\varepsilon + (c_+\pi/4 + c_-\pi/32)t] \\ &= L2q^{-3}\exp[c_1/\varepsilon + c_2 t]. \end{aligned}$$

First case: $t_+ < \pi/2\varepsilon$. Then

$$L2q^{-3}\exp[c_1/\varepsilon + c_2 t_+] \ge \xi = \sqrt{\varepsilon}, \quad \text{so } t_+ \ge [\log(q^3\sqrt{\varepsilon}/2L) - c_1/\varepsilon]/c_2.$$

Using (4.5), we infer $t_+ \ge -c_1/2c_2\varepsilon$. Thus, $|x_t|$ reaches the value $\xi = \sqrt{\varepsilon}$ again not earlier than this time.

Second case: $t_+ = \pi/2\varepsilon$. Then the function $t \mapsto |x_t|$ is bounded by $\sqrt{\varepsilon}$ on the interval $[t_0, \pi/2\varepsilon]$. From the expressions for c_1 and c_2, it is not difficult to see that $|c_1/2c_2| < \pi/2$. Hence, the assertion also holds in the second case. □

Remark. The estimate in Theorem 4.6 is, of course, quantitatively correct only in the sense that it predicts a "growth" time of order $1/\varepsilon$. Further, the upper bound 0.01 for ε, which we used above, is only of a technical nature.

REFERENCES

[1] R. Aris, *Mathematical Modeling. A Chemical Engineer's Perspective*, Process Systems Engineering 1, Academic Press, San Diego, 1999.

[2] E. Benoît, ed., *Dynamic Bifurcation*, Lecture Notes in Math. 1493, Springer-Verlag, New York, 1991.

[3] J. L. Callot, F. Diener, and M. Diener, *Le probleme de la "chasse au canard,"* C. R. Acad. Sci. Paris Sér. A-B, 286 (1978), pp. A1059–A1061.

[4] E. Cumberbatch and A. Fitt, eds., *Mathematical Modelling. Case Studies from Industry*, Cambridge University Press, Cambridge, UK, 2001.

[5] O. Diekmann, S. A. van Gils, S. N. Verduyn-Lunel, and H.-O. Walther, *Delay Equations*, Appl. Math. Sci. 110, Springer-Verlag, New York, 1995.
[6] J. M. Ferreira and I. Györi, *Oscillatory behavior in linear retarded functional-differential equations*, J. Math. Anal. Appl., 128 (1987), pp. 332–346.
[7] G. N. Gorelov and V. A. Sobolev, *Mathematical modeling of critical phenomena in thermal explosion theory*, Combust. Flame, 87 (1991), pp. 203–210.
[8] J. K. Hale, *Functional Differential Equations*, Appl. Math. Sci. 3, Springer-Verlag, New York, 1971.
[9] J. K. Hale and S. M. Verduyn-Lunel, *Introduction to Functional Differential Equations*, Appl. Math. Sci. 99, Springer-Verlag, New York, 1993.
[10] B. Lani-Wayda, *Wandering Solutions of Sine-Like Delay Equations*, Mem. Am. Math. Soc. 151, AMS, Providence, RI, 2001.
[11] J. Mallet-Paret and G. R. Sell, *Systems of differential delay equations: Floquet multipliers and discrete Lyapunov functions*, J. Differential Equations, 125 (1996), pp. 385–440.
[12] Ph. K. Maini and H. G. Othmer, eds., *Mathematical Models for Biological Pattern Formation*, IMA Vol. Math. Appl. 121, Springer-Verlag, New York, 2001.
[13] A. I. Neishtadt, *On delayed stability loss under dynamic bifurcations* I, Differ. Uravn., 23 (1987), pp. 2060–2067 (in Russian).
[14] V. F. Butuzov, N. N. Nefedov, and K. R. Schneider, *Singularly perturbed problems in case of exchange of stabilities*, J. Math. Sci., 21 (2004), pp. 1973–2079.
[15] E. M. Wright, *A non-linear difference-differential equation*, J. Reine Angew. Math., 194 (1955), pp. 66–87.

SIAM J. MATH. ANAL.
Vol. 36, No. 5, pp. 1540–1579

INTEGRAL FUNCTIONALS AND THE GAP PROBLEM: SHARP BOUNDS FOR RELAXATION AND ENERGY CONCENTRATION*

GIUSEPPE MINGIONE† AND DOMENICO MUCCI†

Abstract. We consider integral functionals of the type $F(u) := \int_\Omega f(x,u,Du)\ dx$ exhibiting a gap between the coercivity and the growth exponent

$$L^{-1}|Du|^p \le f(x,u,Du) \le L(1+|Du|^q), \qquad 1<p<q, \qquad 1 \le L < +\infty\,.$$

We give lower semicontinuity results and conditions ensuring that the relaxed functional $\overline{F}$ is equal to $\int_\Omega Qf(x,u,Du)\ dx$, where Qf denotes the usual quasi-convex envelope; our conditions are sharp. Indeed, we also provide counterexamples where such an integral representation fails, showing that energy concentrations appear in the relaxation procedure leading to a measure representation of $\overline{F}$ with a nonzero singular part, which is explicitly computed. The main point in our analysis is that such relaxation results depend in a subtle way on the interaction between the ratio q/p and the degree of regularity of the integrand f with respect to the variable x. Our results extend theorems for nonconvex integrals due to Fonseca and Malý and Kristensen; the energies we treat are related to strongly anisotropic settings.

Key words. relaxation, gap phenomenon, quasi-convexity, nonstandard growth conditions

AMS subject classifications. 49J45, 49Q10

DOI. 10.1137/S0036141003424113

1. Introduction.

In recent years there has been increasing interest in variational integrals defined on Sobolev spaces and exhibiting a gap between the growth and coercivity exponents

$$\int_\Omega f(x,u,Du)\ dx\,; \qquad L^{-1}|Du|^p \le f(x,u,Du) \le L(1+|Du|^q), \qquad L \ge 1, \tag{1.1}$$

where $1<p<q<+\infty$, $u:\Omega \to \mathbb{R}^N$ and Ω is a domain in $\mathbb{R}^n$. The main issues treated in this setting are concerned with the lower semicontinuity, relaxation, and regularity of minimizers of such functionals. Therefore a great many analytical techniques have been developed; examples of papers devoted to such an issue are [1], [8], [24], [25], [33], [34], and [40]. In particular, in the paper [24] Fonseca and Malý addressed the issue of the relaxation and the lower semicontinuity of quasi-convex functionals satisfying (1.1) with $f \equiv f(Du)$. They succeeded in proving that

$$\int_\Omega f(Du)\ dx \le \liminf_k \int_\Omega f(Du_k)\ dx \tag{1.2}$$

for any sequence of functions $u_k \in W^{1,q}(\Omega;\mathbb{R}^N)$ weakly converging to u, $u_k \rightharpoonup u$, in $W^{1,p}(\Omega;\mathbb{R}^N)$; moreover they proved that the relaxed functional (when considered with respect to the weak topology of $W^{1,p}(\Omega;\mathbb{R}^N)$) is a Radon measure, say μ_u. Also see the work of Kristensen [33], [34], and [36] concerning this type of result. The

*Received by the editors March 6, 2003; accepted for publication (in revised form) March 5, 2004; published electronically April 29, 2005. This research was partially supported by MIUR via the project "Calcolo delle Variazioni" (Cofin 2000 and 2002).
http://www.siam.org/journals/sima/36-5/42411.html
†Dipartimento di Matematica dell'Università di Parma, Via D'Azeglio 85/A, I-43100 Parma, Italy (giuseppe.mingione@unipr.it, domenico.mucci@unipr.it).

previous theorems are valid, provided the gap between p and q, measured in terms of the ratio q/p, is not too large, depending on the dimension n, i.e.,

$$\frac{q}{p} < \frac{n}{n-1}\ ; \tag{1.3}$$

see [24] and [37] for discussion on the optimality of (1.3); see also [40], [30]. Subsequently, in [8], Bouchitté, Fonseca, and Malý also proved that the density of the absolutely continuous part (with respect to the Lebesgue measure) of μ_u coincides with the quantity $Qf(Du)$, where $Qf(\cdot)$ denotes the quasi-convex envelope of f (see [14]). A main problem of the issue is, at this stage, saying something about the *singular part* of μ_u. In a more recent paper [1], Acerbi, Bouchitté, and Fonseca examined the nonautonomous case $f \equiv f(x, Du)$, analyzing the relaxed functional and proving, under the main assumption of convexity of the function $z \mapsto f(x, z)$, that the existence of the singular part of the measure μ_u is related to the presence of the Lavrentiev phenomenon that such functionals typically present, i.e., the impossibility to approximate in energy a given function $u \in W^{1,p}$ with $W^{1,q}$-functions. In particular, they prove that, if there is no Lavrentiev phenomenon at u, then there is no singular part of the measure μ_u. Note that the significance of the situation of the paper [1] (even if f is considered to be convex with respect to the gradient variable) lies in the combination of the facts that f both depends on x and exhibits a gap. Needless to say there is no Lavrentiev gap when one of the two previous conditions fails (by a well-known convolution argument based on the convexity of f and Jensen inequality). This suggests that, when dealing with functionals as in (1.1), the presence of the x and, even worse, of both x and u determines a critical situation. In any case not much is known about the relaxed functional and the singular part of μ_u in the general case (1.1); compare [9, Chap. 21] for a partial result. It is important to note that all the analysis in [1] is based on the convexity of f. Let us explicitly remark that the techniques of the previous works do not apply to quasi-convex energy densities of the type $f(x, Du)$ without imposing severe restrictions on the way the function f depends on x.

The aim of this paper is to investigate such an issue, concentrating on some classes of nonconvex functionals as in (1.1) that will have to satisfy certain structure assumptions but that, nevertheless, will allow the consideration of a large class of functionals not covered in the available literature. For ease of exposition we assume that

$$F(u) \equiv F(u, \Omega) := \int_\Omega f(x, Du)\ dx \tag{1.4}$$

and consider the relaxed functional (see also Remark 2.1 below)

$$\overline{F}(u, \Omega) := \inf \left\{ \liminf_{k \to +\infty} F(u_k, \Omega) \quad | \quad \{u_k\} \subset W^{1,q}_{\mathrm{loc}}(\Omega; \mathbb{R}^N), \ u_k \rightharpoonup u \ \text{ in } \ W^{1,p}(\Omega; \mathbb{R}^N) \right\}. \tag{1.5}$$

The problem we address is: proving measure representation properties of the relaxed functional, representing its absolute continuous part and finally discovering whether or not in the relaxation procedure a singular part emerges. Moreover, the problem of finding explicit examples of singular parts of μ_u, when the Lavrentiev phenomenon

does occur is also relevant. In this direction very few results are available in the literature; see [12], [25], [26], [40], [42].

Due to the lack of a general theory, our analysis starts, and largely proceeds, by considering some model examples. Let us consider the following relevant ones:

$$F_1(u) := \int_\Omega |Du|^{p(x)}\,dx \qquad\qquad F_2(u) := \int_\Omega (|Du|^p + a(x)|Du|^q)\,dx, \tag{1.6}$$

where $p \le p(x) \le q$ and $0 \le a(x) \le L < +\infty$ are continuous functions.

What we are going to discover in the following is that, in such a situation, the form of the relaxed functional is linked to a subtle interplay between the gap of the functional and the regularity of the energy density $f(x, Du)$ with respect to the variable x. Roughly speaking, and, for the sake of clarity, referring to F_2, we are going to show that the larger the gap between p and q, the higher the regularity required on the function $f(x, \cdot)$. Indeed, we shall see that for any functional of the type in (1.1) controlled by F_2 in the sense

$$L^{-1}(|z|^p + a(x)|z|^q) \le f(x, z) \le L(|z|^p + a(x)|z|^q + 1), \qquad L \ge 1,$$

the relaxed functional described in (1.5) is exactly

$$\int_\Omega Qf(x, Du)\,dx,$$

provided the function $a(x)$ is α-Hölder continuous and the following bound is satisfied:

$$\frac{q}{p} \le \frac{n+\alpha}{n}. \tag{1.7}$$

Therefore, no energy concentration appears in the relaxation procedure. This condition must clearly be compared to the one appearing in (1.3): the difference is that the regularity of f with respect to the variable x comes into play via the exponent α. Now, though this bound may appear of a technical nature (at least looking at the proof) the interesting thing is that it actually turns out to be sharp: indeed, we build a functional, which is exactly F_2 for a particular choice of the function $a(x)$, for which the relaxation process does not lead to Radon measure, but rather to a Borel measure, in the form of an infinite Dirac mass concentrated in one point. This can be done as soon as the bound in (1.7) is violated; note that this counterexample can be obtained already in the scalar case $N = 1$ and in the case of convex integrals. A similar situation occurs when considering the relaxation problem for functional F_1, where another condition, in some sense similar to (1.7), involving the oscillations and the regularity of the exponent function $p(x)$ must be considered; see (5.6) below and section 8.

But let us give an outlook on the content of this paper. To be general, we shall treat functionals like the one in (1.1) and satisfying the following additional structure assumption:

$$L^{-1}\psi(x, |z|) \le f(x, z) \le L(\psi(x, |z|) + 1), \tag{1.8}$$

where $\psi(x, |z|)$ is a suitable convex function with (p, q) growth (with respect to z), typical examples being the energy densities of the functionals F_1 and F_2; see Remark 3.1 below. Therefore, we shall not deal with typical examples of quasi-convex energy

densities such as $|z|^p + |\det z|$ as considered, for example, in [40], [25], [24]. In order to prove the integral representation, a key point will be certain continuity estimates on the maximal function with respect to the function $\psi(x, |z|)$ and the density of smooth maps in energy; see sections 4 and 5. This is the point where bounds as in (1.7) come into the play. Then we proceed building in sections 7 and 8 the counterexamples proving the sharpness of our assumptions. It is worth pointing out that all the counterexamples we work out are developed in the scalar case ($N = 1$).

Finally, let us say that for the sake of brevity we confine our analysis to integral functionals of the type in (1.1), which already incorporate all the technical and applicative significance of the present issues; the same results can be extended without serious additional efforts to integrands of the type $f \equiv f(x, u, Du)$.

2. Notation and preliminary results. In what follows, Ω is always a fixed open subset of $\mathbb{R}^n$ and $\mathcal{A}$ is the family of its open subsets; if $A, B \in \mathcal{A}$, by $A \subset\subset B$ we mean that the closure $\overline{A}$ of A is a compact set contained in B, and by $\mathcal{A}_0$ we denote the class of all $A \in \mathcal{A}$ such that $A \subset\subset \Omega$. Also, $B_r(x)$ denotes the ball of radius $r > 0$ centered at $x \in \mathbb{R}^n$ and $B_r := B_r(0)$. We will denote $L^p(\Omega; \mathbb{R}^N)$ and $W^{1,p}(\Omega; \mathbb{R}^N)$, $p \geq 1$, the standard Lebesgue and Sobolev spaces of functions $u : \Omega \to \mathbb{R}^N$; for the sake of brevity these spaces will be also denoted omitting the dependence on the target space, e.g., $W^{1,p}(\Omega)$, $L^p(\Omega)$, and so on. As customary, in the rest of the paper c will denote an unspecified positive constant, possibly varying from line to line; the relevant connections will be emphasized when needed while more peculiar occurrences will be stressed by $c_1, c_2, \widetilde{c}$, etc. We will consider nonnegative variational functionals $F : L^1(\Omega; \mathbb{R}^N) \to [0, +\infty]$ of the type

$$F(u) = \begin{cases} \displaystyle\int_\Omega f(x, Du(x))\, dx & \text{if } u \in C^1(\Omega; \mathbb{R}^N), \\ +\infty & \text{elsewhere on } L^1(\Omega; \mathbb{R}^N), \end{cases}$$

where $f : \Omega \times \mathbb{R}^{N \times n} \to [0, +\infty)$ is a Borel measurable function satisfying a nonstandard growth condition, see (3.1) and (3.2). We are interested in the study of the relaxed functional of F with respect to the strong $L^1(\Omega; \mathbb{R}^N)$ convergence, i.e., the lower semicontinuous envelope of F with respect to the $L^1(\Omega; \mathbb{R}^N)$ topology. To show measure property and integral representation of the relaxed functional we make use of the localization method, which considers at the same time the dependence on the function and on the open set. To this aim, we will work with nonnegative variational functionals $F : L^1(\Omega; \mathbb{R}^N) \times \mathcal{A} \to [0, +\infty]$ of the form

$$F(u, A) := \begin{cases} \displaystyle\int_A f(x, Du(x))\, dx & \text{if } u \in C^1(A; \mathbb{R}^N), \\ +\infty & \text{elsewhere on } L^1(\Omega; \mathbb{R}^N) \end{cases} \tag{2.1}$$

for any open set $A \in \mathcal{A}$. Also, for every $A \in \mathcal{A}$, we denote by $\overline{F}(\cdot, A)$ the relaxed functional of $F(\cdot, A)$ with respect to the strong $L^1(\Omega; \mathbb{R}^N)$ convergence, given for all $u \in L^1(\Omega; \mathbb{R}^N)$ by

$$\overline{F}(u, A) := \inf \Big\{ \liminf_{k \to +\infty} F(u_k, A) \quad | \quad \{u_k\} \subset L^1(\Omega; \mathbb{R}^N), \ u_k \to u \text{ in } L^1(\Omega; \mathbb{R}^N) \Big\}. \tag{2.2}$$

Remark 2.1. Since each sequence $\{u_k\} \subset L^1(A; \mathbb{R}^N)$ converging to u strongly in $L^1(A; \mathbb{R}^N)$ can be extended to a sequence $L^1(\Omega; \mathbb{R}^N)$-converging to u, if $\overline{F}(u, A) <$

$+\infty$ by (2.1) we have

$$\overline{F}(u,A)=\inf\Big\{\liminf_{k\to+\infty}\int_A f(x,Du_k(x))\,dx \quad | \quad \{u_k\}\subset C^1(A;\mathbb{R}^N), \\ u_k\to u \ \text{ in } \ L^1(A;\mathbb{R}^N)\Big\}.$$

We explicitly remark that whenever the function f satisfies the following (p,q)-growth condition:

$$L^{-1}|z|^p\le f(x,z)\le L(1+|z|^q),\qquad 1<p<q<+\infty,\qquad 1\le L,$$

then the previous relaxed functional coincides with the one following, analyzed in [1] [8] [24]:

$$\overline{F}(u,A)=\inf\Big\{\liminf_{k\to+\infty}\int_A f(x,Du_k(x))\,dx \quad | \quad \{u_k\}\subset W^{1,q}_{\rm loc}(A;\mathbb{R}^N), \\ u_k\rightharpoonup u \ \text{ in } \ W^{1,p}(A;\mathbb{R}^N)\Big\}.$$

To show the measure property we recall some well-known facts about set functions.

DEFINITION 2.2. *A function $\alpha:\mathcal{A}\to[0,+\infty]$ is called an increasing set function if $\alpha(\emptyset)=0$ and $\alpha(A)\le\alpha(B)$ if $A\subseteq B$. An increasing set function α is said to be subadditive if*

$$\alpha(A\cup B)\le\alpha(A)+\alpha(B)$$

for all $A,B\in\mathcal{A}$, and it is said to be superadditive if

$$\alpha(A\cup B)\ge\alpha(A)+\alpha(B)$$

for all $A,B\in\mathcal{A}$ with $A\cap B=\emptyset$; finally α is said to be inner regular if for all $A\in\mathcal{A}$

$$\alpha(A)=\sup\{\alpha(B)\mid B\in\mathcal{A}\,,\quad B\subset\subset A\}.$$

Remark 2.3. Since $f\ge 0$, then $\overline{F}(u,\cdot)$ is an increasing set function for every $u\in L^1(\Omega;\mathbb{R}^N)$. Moreover, by definition of relaxation one directly obtains that $\overline{F}(u,\cdot)$ is superadditive. Finally, we denote by $\overline{F}_-(u,\cdot)$ the inner regular envelope of $\overline{F}(u,\cdot)$, given by

$$\overline{F}_-(u,C):=\sup\{\overline{F}(u,B)\mid B\in\mathcal{A}\,,\ B\subset\subset C\} \tag{2.3}$$

for every $C\in\mathcal{A}$, so that $\overline{F}(u,\cdot)$ is inner regular if $\overline{F}(u,\cdot)\equiv\overline{F}_-(u,\cdot)$ on $\mathcal{A}$. We will apply the following criterion due to De Giorgi–Letta [18]; compare also [9, sect. 10.2].

THEOREM 2.4 (measure property criterion). *Let $\alpha:\mathcal{A}\to[0,+\infty]$ be an increasing set function. Then the following statements are equivalent:*

(i) *α is the trace on $\mathcal{A}$ of a Borel measure on Ω;*
(ii) *α is subadditive, superadditive, and inner regular;*
(iii) *the set function $\widetilde{\alpha}(E):=\inf\{\alpha(A)\mid A\in\mathcal{A},\ E\subset A\}$ defines a Borel measure on Ω.*

We recall a celebrated lower semicontinuity result first obtained by De Giorgi [17] and due to Ioffe [32] in the following general form.

THEOREM 2.5 (L^1-semicontinuity). *Let A be a bounded open set of $\mathbb{R}^n$ and let $g:A\times\mathbb{R}^N\times\mathbb{R}^{N\times n}\to[0,+\infty)$ be a Carathéodory function such that $g(x,u,\cdot)$ is convex for every $u\in\mathbb{R}^N$ and for a.e. $x\in A$. Then the functional*

$$G(u):=\int_A g(x,u(x),Du(x))\,dx$$

is lower semicontinuous on $W^{1,1}(A;\mathbb{R}^N)$ *with respect to the weak convergence in* $W^{1,1}(A;\mathbb{R}^N)$.

We end this section by stating an elementary lemma which is a version of De Giorgi's slicing argument.

LEMMA 2.6 (slicing lemma revisited). *Let* $\{f_k\}$ *be a sequence of nonnegative functions in* $L^1(B_1)$ *with*

$$\sup_k \int_{B_1} f_k \, dx \le M < +\infty.$$

Then, fixed $0 < s < t < 1$, *for every* $\epsilon > 0$ *there exist* $N \equiv N(\epsilon, M)$, *an integer* $1 \le h \le N$, *and a (not relabelled) subsequence* $\{f_k\}$ *such that*

$$\sup_k \int_{A_h} f_k \, dx \le \epsilon,$$

where, for $i \in \{0, 1, 2, \dots, N-1\}$,

$$A_i := B_{s_{i+1}} \setminus B_{s_i} \qquad \text{and} \qquad s_i := s + \frac{t-s}{N} \, i.$$

Proof. Choose N in such a way that $N\epsilon > M$. It follows that for each $k \in \mathbb{N}$ there exists $i \equiv i(k)$ such that

$$\int_{A_{i(k)}} f_k \, dx \le \frac{M}{N},$$

and the assertion follows via a standard compactness argument. □

3. Measure property of the relaxed functional. In this section we consider nonnegative variational functionals $F : L^1(\Omega;\mathbb{R}^N) \times \mathcal{A} \to [0, +\infty]$ of the form (2.1) for any open set $A \in \mathcal{A}$, where $f : \Omega \times \mathbb{R}^{N\times n} \to [0, +\infty)$ is a Borel measurable function satisfying a nonstandard growth condition of the form

$$\alpha \, \psi(x, |z|) \le f(x, z) \le b(x) + \beta \, \psi(x, |z|) \tag{3.1}$$

for a.e. $x \in \Omega$ and all $z \in \mathbb{R}^{N\times n}$. Also, for every $A \in \mathcal{A}$, we denote by $\overline{F}(\cdot, A)$ the relaxed functional of $F(\cdot, A)$ with respect to the strong $L^1(\Omega;\mathbb{R}^N)$ convergence, given for all $u \in L^1(\Omega;\mathbb{R}^N)$ by (2.2).

Here $0 < \alpha \le \beta < +\infty$, $b(x)$ is a nonnegative function in $L^1(\Omega)$, and $\psi : \Omega \times [0, +\infty) \to [0, +\infty)$ is a suitable Borel function satisfying the following properties:

(i) $t \mapsto \psi(x, t)$ is nondecreasing and convex for a.e. $x \in \Omega$, with $\psi(x, 0) \equiv 0$;

(ii) for every open set $A \in \mathcal{A}_0$ there exist $1 < c = c(A) < +\infty$ and $1 < p = p(A) \le q = q(A) < +\infty$ such that for a.e. $x \in A$ we have

$$\begin{aligned} c^{-1} t^p \le \psi(x, t) &\le c\,(t^q + 1) \qquad \forall\, t \ge 0, \\ \psi(x, \lambda t) &\le c \max\{\lambda^q, \lambda^p\} \psi(x, t) \qquad \forall\, t \ge 0\,,\ \lambda \ge 0\,, \\ \psi(x, t_1 + t_2) &\le c\, 2^{q-1} (\psi(x, t_1) + \psi(x, t_2)) \qquad \forall\, t_1\,,\, t_2 \ge 0. \end{aligned} \tag{3.2}$$

Remark 3.1. Note that the third property in (3.2) follows from the second one and from the convexity of $\psi(x, \cdot)$. Moreover, by monotonicity and convexity of $\psi(x, \cdot)$, it follows that $z \mapsto \psi(x, |z|)$ is convex for a.e. $x \in \Omega$. Therefore our analysis of functionals with a gap in the sense of (1.1) is confined to those special functionals

with an energy density satisfying (3.1); these also satisfy (1.1) in view of (i), for a suitable choice of (p, q). Observe that the second property in (3.2) is a sort of $\triangle_2$ condition for the function $t \mapsto \psi(x,t)$, uniform with respect to x.

We introduce the following classes of measurable functions in $L^1(A;\mathbb{R}^N)$ and $W^{1,1}(A;\mathbb{R}^N)$, for every $A \in \mathcal{A}$:

$$\begin{aligned} L^{\psi}(A;\mathbb{R}^N) &:= \{u \in L^1(A;\mathbb{R}^N) \mid \ \psi(x,|u(x)|) \in L^1(A)\},\\ W^{\psi}(A;\mathbb{R}^N) &:= \{u \in W^{1,1}(A;\mathbb{R}^N) \mid u \in L^{\psi}(A;\mathbb{R}^N), \quad Du \in L^{\psi}(A;\mathbb{R}^{N\times n})\},\\ W^{\psi}_{\mathrm{loc}}(A;\mathbb{R}^N) &:= \{u \in L^1(A;\mathbb{R}^N) \mid u_{|B} \in W^{\psi}(B;\mathbb{R}^N) \quad \forall\, B \in \mathcal{A},\, B \subset\subset A\}. \end{aligned}$$

Note that, by definition of ψ, these are all convex sets; by (3.2) one infers that $W^{\psi}_{\mathrm{loc}}(A;\mathbb{R}^N)$ is a vector space. We remark that if $A \in \mathcal{A}_0$, these spaces, when equipped with a suitable norm via a suitable Jague function and under certain assumptions, become Banach spaces known as Orlicz–Musielak spaces; these are currently the object of intensive investigation (see, for instance, [43], [22], [19], [20], [31]).

DEFINITION 3.2. *We say that $W^{\psi}(\Omega;\mathbb{R}^N)$ satisfies a Sobolev type property if, for any function $u \in L^1(\Omega;\mathbb{R}^N)$ and every open set $A \in \mathcal{A}_0$ with Lipschitz boundary such that $\int_A \psi(x,|Du|)\,dx < +\infty$, we have*

$$\int_B \psi(x,|u(x)|)\,dx \le C\left(\int_B \psi(x,|Du(x)|)\,dx + \int_B |u(x)|\,dx\right)^{\beta} \quad \forall\, B \in \mathcal{A}_0,\ B \subset\subset A,$$

where $C, \beta \in [1,+\infty)$ are constants, possibly depending on A. Moreover, we say that $W^{\psi}(\Omega;\mathbb{R}^N)$ satisfies a Rellich's-type property if, for every function $u \in L^1_{\mathrm{loc}}(\Omega;\mathbb{R}^N)$, every open set $A \in \mathcal{A}_0$ with Lipschitz boundary, and every $\{u_j\} \subset W^{1,1}(A;\mathbb{R}^N)$ with $u_j \to u$ strongly in $L^1(A;\mathbb{R}^N)$ and $\sup_j \int_A \psi(x,|Du_j|)\,dx < +\infty$, we have

$$\lim_{j\to+\infty} \int_A \psi(x,|u_j - u|)\,dx = 0.$$

In particular, if $W^{\psi}(\Omega;\mathbb{R}^N)$ satisfies a Sobolev-type property, we easily obtain for every $A \in \mathcal{A}$

$$W^{\psi}_{\mathrm{loc}}(A;\mathbb{R}^N) = \{u \in L^1(A;\mathbb{R}^N) \mid Du \in L^{\psi}(B;\mathbb{R}^{N\times n}) \quad \forall\, B \in \mathcal{A},\, B \subset\subset A\}. \tag{3.3}$$

In this section we prove the following.

THEOREM 3.3 (measure property). *Let $F : L^1(\Omega;\mathbb{R}^N) \times \mathcal{A} \to [0,+\infty]$ be as in* (2.1)*, with f as in* (3.1)*, and $\psi : \Omega \times [0,+\infty) \to [0,+\infty)$ satisfying* (i) *and* (ii) *above. Suppose that $W^{\psi}(\Omega;\mathbb{R}^N)$ satisfies a Sobolev and a Rellich-type property. Then, for every function $u \in L^1(\Omega;\mathbb{R}^N)$, the functional $\overline{F}(u,\cdot)$ is the trace on $\mathcal{A}$ of a Borel measure on Ω.*

Example 3.4. Of course, $\psi(x,|z|) := |z|^p$, $p > 1$, verifies Theorem 3.3. Here we outline two important classes of convex functions satisfying the hypotheses of Theorem 3.3. The first one is the case of dependence on x on the growth exponent, i.e.,

$$\psi(x,|z|) := |z|^{p(x)}, \tag{3.4}$$

where $p : \Omega \to (1,+\infty)$ is any fixed continuous function with $p(x) > 1$ for every $x \in \Omega$. It is easy to show that $|z|^{p(x)}$ satisfies (3.2), since for every $A \in \mathcal{A}_0$ we have

$1 < p(A) \equiv \inf_A p(x) \le \sup_A p(x) \equiv q(A) < +\infty$. Moreover, in [13] it is shown that $|z|^{p(x)}$ satisfies both a Sobolev and a Rellich-type property. The second example is

$$\psi(x, |z|) := |z|^p + a(x)\,|z|^q, \tag{3.5}$$

where $1 < p \le q < +\infty$ and $a(x) \in L^\infty(\Omega)$, with $a(x) \ge 0$. Of course, the Sobolev and Rellich-type property hold if $q < p^*$, where p^* is the Sobolev conjugate of p, i.e., $p^* = np/(n-p)$ if $p < n$, $p^* = +\infty$ if $p \ge n$.

Before proving Theorem 3.3, we give some preliminary results. The following lemma is a straightforward consequence of the previous definitions and Theorem 2.5.

LEMMA 3.5. *Under the hypotheses of Theorem* 3.3*, let $A \in \mathcal{A}_0$ and u be a function in $L^1(\Omega;\mathbb{R}^N)$ such that $\overline{F}(u, A) < +\infty$. Then $u \in W^{\psi}_{\mathrm{loc}}(A;\mathbb{R}^N)$ and*

$$\int_A \psi(x, |Du|)\,dx \le \frac{1}{\alpha}\,\overline{F}(u, A) < +\infty.$$

Let us now recall that if A', A are open sets in $\mathcal{A}$, with $A' \subset\subset A$, a cut-off function between A' and A is a smooth function $\phi \in C_0^\infty(\Omega)$ with $\operatorname{spt}\phi \subset A$, $0 \le \phi \le 1$ and $\phi \equiv 1$ on A'.

Due to growth condition (3.1), we now obtain the following fundamental L^ψ estimate. The proof is a readaptation of [9, sect. 12.2], taking into account the new growth conditions dictated by (3.2). We omit it for the sake of brevity.

LEMMA 3.6 (fundamental estimate). *Under the hypotheses of Theorem* 3.3*, for all open sets $A, A', B \in \mathcal{A}$, with $A' \subset\subset A$, and for every $\sigma > 0$, there exists a constant $M_\sigma > 0$ such that for every $u, v \in L^1(\Omega;\mathbb{R}^N)$ there exists a cut-off function ϕ between A' and A such that*

$$\begin{aligned} F(\phi u + (1-\phi)v, A' \cup B) \;\le\; & (1+\sigma)(F(u, A) + F(v, B)) \\ & + M_\sigma \int_{A \cap B} \psi(x, |u - v|)\,dx + \sigma. \end{aligned} \tag{3.6}$$

By using the Rellich-type property and the fundamental estimate above, and following arguments from [42], it is possible to prove a weak subadditivity property for the set function $\overline{F}(w, \cdot)$.

LEMMA 3.7 (weak subadditivity). *Under the hypotheses of Theorem* 3.3*, for every $w \in L^1(\Omega;\mathbb{R}^N)$ we have*

$$\overline{F}(w, A' \cup B) \le \overline{F}(w, A) + \overline{F}(w, B) \tag{3.7}$$

for every $A', A \in \mathcal{A}$, with $A' \subset\subset A$, and every $B \in \mathcal{A}$ such that B has a Lipschitz boundary.

We are now going to give the following proof.

Proof of Theorem 3.3. *Step* 1*: the case $f(x, Du) := \psi(x, |Du|)$.* Define $\Psi : L^1(\Omega;\mathbb{R}^N) \times \mathcal{A} \to [0, +\infty]$ by

$$\Psi(u, A) := \begin{cases} \displaystyle\int_A \psi(x, |Du(x)|)\,dx & \text{if } u \in C^1(A;\mathbb{R}^N), \\ +\infty & \text{elsewhere on } L^1(\Omega;\mathbb{R}^N), \end{cases} \tag{3.8}$$

and let $\overline{\Psi}(\cdot, A)$ be the $L^1(\Omega)$-lower semicontinuous envelope of $\Psi(\cdot, A)$ for every $A \in \mathcal{A}$. Finally, let $\overline{\Psi}_-(u, \cdot)$ be the inner regular envelope of $\overline{\Psi}(u, \cdot)$ (see (2.3)), i.e., for every $u \in L^1(\Omega;\mathbb{R}^N)$,

$$\overline{\Psi}_-(u, C) := \sup\{\overline{\Psi}(u, B) \mid B \in \mathcal{A},\ B \subset\subset C\}, \qquad C \in \mathcal{A}. \tag{3.9}$$

Making use of a convexity argument, we are able to prove inner regularity. We omit the details of the proof of (3.17) and (3.18) and refer to [42, Prop. 3.1] for a similar computation (see also Remark 2.3).

PROPOSITION 3.8 (inner regularity). *Let* $f(x,z) := \psi(x,|z|)$ *and* $\Psi : L^1(\Omega;\mathbb{R}^N) \times \mathcal{A} \to [0,+\infty]$ *be given by* (3.8). *Then for every* $u \in L^1(\Omega;\mathbb{R}^N)$ *the increasing set function* $\overline{\Psi}(u,\cdot)$ *is inner regular, i.e., for every* $C \in \mathcal{A}$

$$\overline{\Psi}(u,C) = \overline{\Psi}_-(u,C), \tag{3.10}$$

where $\overline{\Psi}_-(u,C)$ *is given by* (3.9).

Proof. By the monotonicity of $\overline{\Psi}(u,\cdot)$, it suffices to show that "$\le$" holds in (3.10), in case $\overline{\Psi}_-(u,C) < +\infty$. To this aim, for every $\epsilon > 0$ and $j \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$, let $A^j \in \mathcal{A}_0$ be such that $A^j \subset\subset A^{j+1} \subset\subset C$, A^j has a Lipschitz boundary so that $|\partial A^j| = 0$, $C = \cup_j A_j$ and

$$\overline{\Psi}_-(u,C) - \epsilon\, 2^{-j} \le \overline{\Psi}(u,A^j) \le \overline{\Psi}_-(u,C) \qquad \forall\, j \in \mathbb{N}_0. \tag{3.11}$$

For every $j \in \mathbb{N}_0$, let $\{u_h^j\}_h \subset L^1(\Omega)$, obviously depending also on ϵ, be such that

$$\lim_{h\to+\infty} \|u_h^j - u\|_{L^1(\Omega)} = 0 \quad \text{ and } \quad \overline{\Psi}(u,A^j) = \liminf_{h\to+\infty} \Psi(u_h^j, A^j) < +\infty\,. \tag{3.12}$$

Possibly passing to a subsequence, we can suppose that $u_h^j \to u$ a.e. on Ω,

$$\sup_h \int_{A^j} \psi(x,|Du_h^j|)\,dx < +\infty,$$

$\{u_{h|A^j}^j\}_h \subset C^1(A^j)$, and that the lower limit in (3.12) is a limit. Then, by the Rellich-type property (Definition 3.2) and (3.12)

$$\lim_{h\to+\infty} \int_{A^j} \psi(x,|u_h^j - u|)\,dx = 0 \qquad \forall\, j \in \mathbb{N}_0. \tag{3.13}$$

Set $A^{-1} := \emptyset$ and let us consider a partition of unity $\{\phi_j\}_{j\in\mathbb{N}_0}$ relative to the open covering of C given by $\{A^{j+1} \setminus \overline{A}^{j-1}\}_{j\in\mathbb{N}_0}$. More precisely, for every $j \in \mathbb{N}_0$ we have that $\phi_j \in C_0^1(A^{j+1} \setminus \overline{A}^{j-1})$ and

$$0 \le \phi_j(x) \le 1\,, \quad \sum_{j=0}^{+\infty} \phi_j(x) = 1 \qquad \forall\, x \in C\,. \tag{3.14}$$

For every $j \in \mathbb{N}$, let $h(j) \in \mathbb{N}$ be chosen later, set $v_j := u_{h(j)}^j$, and

$$w_\epsilon(x) := \sum_{j=1}^{+\infty} \phi_{j-1}(x)\, v_j(x), \qquad x \in C. \tag{3.15}$$

Note that, since $v_{j|A^j} \in C^1(A^j)$, we have that $\phi_{j-1}(x)\, v_j(x) \in C_0^1(C)$ for every $j \in \mathbb{N}$. Moreover, since every x in C has a neighborhood contained at most in the union of three sets of the type $A^{j+1} \setminus \overline{A}^{j-1}$, for every $x \in C$ the infinite sum in the right-hand side of (3.15) reduces to a finite one, hence $w_\epsilon \in C^1(C)$ for every $\epsilon > 0$.

Taking $w_\epsilon \equiv u$ in $\Omega \setminus C$, for every $t \in]0,1[$ the function tw_ϵ belongs to $L^1(\Omega)$ and by (3.14)

$$\|tw_\epsilon - u\|_{L^1(\Omega)} \le t \sum_{j=1}^{+\infty} \int_{A^j} |u^j_{h(j)} - u|\, dx + (1-t)\, \|u\|_{L^1(\Omega)}. \tag{3.16}$$

Now, it is possible to choose the sequence $\{h(j)\}$ so that by (3.16)

$$tw_\epsilon \to u \quad \text{in } L^1(\Omega) \text{ as } \epsilon \to 0^+ \text{ and } t \to 1^-. \tag{3.17}$$

Moreover, taking account of the convexity of $z \to \psi(x,|z|)$, since $0 \le \phi_{j-1} \le 1$ and the sum in (3.15) is locally finite, arguing as in [42, Prop. 3.1], by (3.11), (3.12), and (3.13) we can choose $\{h(j)\}$ so that for any $t \in]0,1[$ we also have

$$\int_C \psi(x,|tDw_\epsilon|)\, dx \le \overline{\Psi}_-(u,C) + 5\epsilon + (1-t)\, \epsilon < +\infty. \tag{3.18}$$

In particular, since $tw_\epsilon \in C^1(C)$, by (3.8) we have $\overline{\Psi}(tw_\epsilon, C) = \int_C \psi(x, t|Dw_\epsilon|)\, dx$. Finally, as $\epsilon \to 0^+$ and $t \to 1^-$, by (3.18) and (3.17) we obtain that $\overline{\Psi}(u,C) \le \overline{\Psi}_-(u,C)$ and hence the assertion. □

Now, since the increasing set function $\overline{\Psi}(u,\cdot)$ is inner regular, and $\overline{\Psi}(u,\cdot)$ is superadditive, thanks to Theorem 2.4 we obtain measure property of $\overline{\Psi}(u,\cdot)$, for every $u \in L^1(\Omega;\mathbb{R}^N)$, if we show that $\overline{\Psi}(u,\cdot)$ is subadditive.

PROPOSITION 3.9 (subadditivity). *For every $w \in L^1(\Omega;\mathbb{R}^N)$ we have*

$$\overline{\Psi}(w, A\cup B) \le \overline{\Psi}(w,A) + \overline{\Psi}(w,B) \qquad \forall\, A, B \in \mathcal{A}. \tag{3.19}$$

Proof. By inner regularity (Proposition 3.8), it is well known that weak subadditivity (Lemma 3.7 with $F = \Psi$) yields (3.19) for any $A, B \in \mathcal{A}$, provided B has a Lipschitz boundary. In fact, for any $C \in \mathcal{A}$ with $C \subset\subset A \cup B$, by enlarging the subset $C \setminus \overline{B}$ a bit, we can find $A' \subset\subset A$ such that $C \subset A' \cup B$, which yields, by (3.7), $\overline{\Psi}(w,C) \le \overline{\Psi}(w, A'\cup B) \le \overline{\Psi}(w,A) + \overline{\Psi}(w,B)$, and hence (3.19), by inner regularity, letting $C \nearrow A \cup B$. Finally, to prove (3.19) for any $B \in \mathcal{A}$, in case $\overline{\Psi}(w, A\cup B) < +\infty$, for each small $\epsilon > 0$ take $C \subset\subset A\cup B$ such that by inner regularity $\overline{\Psi}(w,C) \ge \overline{\Psi}(w, A\cup B) - \epsilon$. We can find an open set $\widetilde{B} \in \mathcal{A}$ with $C \setminus \overline{A} \subset\subset \widetilde{B} \subset B$ and such that $\widetilde{B}$ has a Lipschitz boundary. Then since $C \subset A \cup \widetilde{B}$,

$$\begin{aligned} \overline{\Psi}(w, A\cup B) &\le \overline{\Psi}(w,C) + \epsilon \\ &\le \overline{\Psi}(w, A\cup \widetilde{B}) + \epsilon \\ &\le \overline{\Psi}(w,A) + \overline{\Psi}(w,\widetilde{B}) + \epsilon \\ &\le \overline{\Psi}(w,A) + \overline{\Psi}(w,B) + \epsilon, \end{aligned}$$

and hence we obtain (3.19), letting $\epsilon \to 0^+$. In case $\overline{\Psi}(w, A\cup B) = +\infty$, take $C \subset\subset A \cup B$ with $\overline{\Psi}(w,C) > 1/\epsilon$, so that arguing as before

$$\epsilon^{-1} \le \overline{\Psi}(w,C) \le \overline{\Psi}(w, A\cup\widetilde{B}) \le \overline{\Psi}(w,A) + \overline{\Psi}(w,\widetilde{B}) \le \overline{\Psi}(w,A) + \overline{\Psi}(w,B)$$

and hence (3.19) follows by again letting $\epsilon \to 0^+$. □

Step 2: Measure property of $\overline{F}(u,\cdot)$. Consider now any Borel function f as in Theorem 3.3. We first prove the following.

PROPOSITION 3.10 (inner regularity). *For every $w \in L^1(\Omega;\mathbb{R}^N)$, $\overline{F}(w,\cdot)$ is an inner regular set function.*

Proof. Since $\overline{F}(w,\cdot)$ is an increasing set function, if $\overline{F}_-(w,\cdot)$ is defined by (2.3), it suffices to prove that

$$\overline{F}(w,C) \le \overline{F}_-(w,C) \tag{3.20}$$

for every fixed open set $C \in \mathcal{A}$ and every function $w \in L^1(\Omega)$ such that $\overline{F}_-(w,C) < +\infty$. To this aim note that growth condition (3.1) yields the estimate

$$\alpha\, \overline{\Psi}(w,A) \le \overline{F}(w,A) \le \int_A b(x)\,dx + \beta\, \overline{\Psi}(w,A) \tag{3.21}$$

for every $w \in L^1(\Omega)$ and $A \in \mathcal{A}$, where Ψ is given by (3.8), and the same estimate with $\overline{\Psi}_-$ and $\overline{F}_-$, respectively, instead of $\overline{\Psi}$ and $\overline{F}$ in (3.21). In particular, by the monotonicity and the inner regularity of $\overline{\Psi}(w,\cdot)$ (see Proposition 3.8)

$$\overline{\Psi}(w,A) \le \overline{\Psi}(w,C) = \overline{\Psi}_-(w,C) \le \frac{1}{\alpha}\overline{F}_-(w,C) < +\infty \tag{3.22}$$

for every $A \in \mathcal{A}$ with $A \subset C$. For every $\epsilon > 0$, we can choose an open set $A_\epsilon \in \mathcal{A}$ with a Lipschitz boundary and such that $A_\epsilon \subset\subset C$ so that, by inner regularity of $\overline{\Psi}(w,\cdot)$ and absolute continuity of $b(x) \in L^1(\Omega)$,

$$\overline{\Psi}(w,C) \le \overline{\Psi}(w,A_\epsilon) + \epsilon \qquad \text{and} \qquad 0 \le \int_{C\setminus A_\epsilon} b(x)\,dx \le \epsilon. \tag{3.23}$$

Also, let $B_\epsilon := C \setminus \overline{A}_\epsilon \in \mathcal{A}$, so that if $\widetilde{\Psi}(w,\cdot)$ is the Borel measure given by the extension of $\overline{\Psi}(w,\cdot)$ to Ω (see (iii) in Theorem 2.4), by (3.23) we have

$$\overline{\Psi}(w,B_\epsilon) = \overline{\Psi}(w,C) - \widetilde{\Psi}(w,\overline{A}_\epsilon) \le \overline{\Psi}(w,C) - \overline{\Psi}(w,A_\epsilon) \le \epsilon. \tag{3.24}$$

Moreover, there exists a sequence $\{v_j\} \subset L^1(\Omega)$, converging to w in $L^1(\Omega)$, such that $v_{j|B_\epsilon} \in C^1(B_\epsilon)$ for every j and

$$\overline{\Psi}(w,B_\epsilon) = \lim_{j\to+\infty} \int_{B_\epsilon} \psi(x,|Dv_j|)\,dx < +\infty. \tag{3.25}$$

In particular, by (3.1), (3.23), and (3.25)

$$\liminf_{j\to+\infty} F(v_j,B_\epsilon) \le \int_{C\setminus A_\epsilon} b(x)\,dx + \beta \lim_{j\to+\infty} \Psi(v_j,B_\epsilon) \le \epsilon + \beta\, \overline{\Psi}(w,B_\epsilon). \tag{3.26}$$

Choose now $A', A \in \mathcal{A}_0$ such that A has a Lipschitz boundary and $A_\epsilon \subset\subset A' \subset\subset A \subset \subset C$. Since $\overline{F}(w,A) < +\infty$, there exists a sequence $\{u_j\} \subset L^1(\Omega)$, converging to w in $L^1(\Omega)$, such that $u_{j|A} \in C^1(A)$ for every j and

$$\overline{F}(w,A) = \lim_{j\to+\infty} \int_A f(x,Du_j)\,dx < +\infty. \tag{3.27}$$

By the fundamental estimate (Lemma 3.6) applied with u_j on A and v_j on B_ϵ, for any $\sigma > 0$, we can find $M_\sigma > 0$ and a sequence $\{\phi_j\}$ of smooth cut-off functions between A' and A such that

$$\begin{aligned} F(w_j, A' \cup B_\epsilon) \;\le\; & (1+\sigma)(F(u_j,A) + F(v_j,B_\epsilon)) \\ & + M_\sigma \int_{A\cap B_\epsilon} \psi(x,|u_j - v_j|)\,dx + \sigma, \end{aligned} \tag{3.28}$$

where $w_j := \phi_j u_j + (1-\phi_j)v_j$. By (3.25), (3.27), and (3.1) we have

$$\sup_j \int_{A\cap B_\epsilon} (\psi(x,|Du_j|) + \psi(x,|Dv_j|))\,dx < +\infty.$$

Moreover, $A\cap B_\epsilon = A\setminus \overline{A}_\epsilon \in \mathcal{A}_0$, whereas $A_\epsilon \subset\subset A$, and hence $A\cap B_\epsilon$ has a Lipschitz boundary given by the disjoint union $\partial A \cup \partial A_\epsilon$. Then, by the Rellich-type property (Definition 3.2) and (3.2), we conclude that

$$\lim_{j\to+\infty} \int_{A\cap B_\epsilon} \psi(x,|u_j - v_j|)\,dx = 0.$$

Then, since $w_j = \phi_j u_j + (1-\phi_j)v_j \to w$ in $L^1(\Omega)$, by (3.28), (3.27), and (3.26) we obtain

$$\begin{aligned} \overline{F}(w, A'\cup B_\epsilon) &\le \liminf_{j\to+\infty} F(w_j, A'\cup B_\epsilon) \\ &\le (1+\sigma)(\overline{F}(w,A) + \epsilon + \beta\,\overline{\Psi}(w,B_\epsilon)) + \sigma. \end{aligned} \tag{3.29}$$

Finally, since $B_\epsilon = C\setminus\overline{A}_\epsilon$ yields $A'\cup B_\epsilon = C$, taking $\epsilon > 0$ small so that $\epsilon(1+\beta)\le\sigma$, by (3.24) and (3.29)

$$\overline{F}(w,C) \le (1+\sigma)(\overline{F}(w,A)+\sigma) + \sigma \le (1+\sigma)(\overline{F}_-(w,C)+\sigma) + \sigma$$

and hence (3.20) holds by the arbitrariness of $\sigma > 0$. □

Since we have just proved that $\overline{F}(w,\cdot)$ is inner regular for every $w \in L^1(\Omega;\mathbb{R}^N)$, arguing as in Proposition 3.9, by weak subadditivity (3.7) we obtain that $\overline{F}(w,\cdot)$ is subadditive. Since $\overline{F}(w,\cdot)$ is trivially superadditive, by Theorem 2.4 the proof of Theorem 3.3 is complete. □

4. Integral representation of the relaxed functional. In this section we show that, under suitable hypotheses on the function $\psi(x,t)$ defined in the previous section, the relaxed functional $\overline{F}(u,A)$ obtained in Theorem 3.3 is of variational type.

DEFINITION 4.1. *We say that a sequence* $\{u_j\} \subset W^\psi_{\rm loc}(\Omega;\mathbb{R}^N)$ *converges to* $u \in W^\psi_{\rm loc}(\Omega;\mathbb{R}^N)$ *strongly in* $W^\psi_{\rm loc}(\Omega;\mathbb{R}^N)$ *if for every* $A \in \mathcal{A}_0$

$$\lim_{j\to+\infty} \int_A (\psi(x,|u_j-u|) + \psi(x,|Du_j - Du|))\,dx = 0. \tag{4.1}$$

Remark 4.2. If $u_j \to u$ in $W^\psi_{\rm loc}(\Omega;\mathbb{R}^N)$, then by (3.2), $u_j \to u$ in $L^1_{\rm loc}(\Omega;\mathbb{R}^N)$. Moreover, $\psi(x,|Du_j|) \to \psi(x,|Du|)$ in $L^1_{\rm loc}(\Omega)$ too. In fact, by the monotonicity of $\psi(x,\cdot)$, for every $A \in \mathcal{A}_0$ we estimate

$$\psi(x,|Du_j|) \le c\,2^{q-1}(\psi(x,|Du_j - Du|) + \psi(x,|Du|))$$

for a.e. $x \in A$, where $c = c(A)$ and $q = q(A)$ are given by (3.2); hence it suffices to apply the dominated convergence theorem.

DEFINITION 4.3. *If* $f \in L^1_{\rm loc}(\Omega)$, *define the maximal function* $M(f)$ *by*

$$(Mf)(x) := \sup_{r>0}(M_{(r)}f)(x), \qquad \textit{where} \qquad (M_{(r)}f)(x) := \frac{1}{|B_r(x)|}\int_{B_r(x)\cap\Omega} |f(y)|\,dy.$$

DEFINITION 4.4. *We say that the function ψ enjoys the maximal property if, for every bounded open set $A \in \mathcal{A}_0$ and every function $f \in L^1(A)$ with $\psi(x,|f(x)|) \in L^1(A)$, we have*

$$\int_A \psi(x,|(Mf)(x)|)\,dx \le C\left(\int_A \psi(x,|f(x)|)\,dx + 1\right)^{\beta}, \tag{4.2}$$

where $C, \beta \in (1,+\infty)$ are positive constants possibly depending on n, A, and ψ.

DEFINITION 4.5. *We say that the function $\psi(x,|z|)$ satisfies the density property if for every $u \in W^{\psi}_{\mathrm{loc}}(\Omega;\mathbb{R}^N)$ there exists a sequence of smooth functions $\{u_j\} \subset C^{\infty}_0(\Omega;\mathbb{R}^N)$ such that $u_j \to u$ in $W^{\psi}_{\mathrm{loc}}(\Omega;\mathbb{R}^N)$. If in addition $u \in W^{\psi}(\Omega;\mathbb{R}^N)$, then we also require that $u_j \to u$ in $L^1(\Omega;\mathbb{R}^N)$.*

Let us observe the following relation between the definitions given above.

PROPOSITION 4.6. *The maximal property implies the density property.*

Proof. It is a consequence of the Lebesgue dominated convergence theorem; we shall keep the notation introduced for Definition 4.3. Let $\{\varphi_\epsilon\}_{\epsilon\in(0,1)}$, be a family of standard mollifiers and let $w \in W^{\psi}_{\mathrm{loc}}(\Omega;\mathbb{R}^N)$ with $w_\epsilon(x) := w * \varphi_\epsilon(x)$ for every $x \in \Omega$ such that $dist(x,\partial\Omega) \ge 2\epsilon$. Observe that by the very definition of convolution and maximal function it follows that $|w_\epsilon(x)| \le (Mw)(x)$ for every $x \in A$ such that $\operatorname{dist}(x,\partial A) \ge 2\epsilon$. Now take an increasing sequence of open subsets $A_j \nearrow \Omega$ such that $A_j \subset\subset A_{j+1} \subset \Omega$ and define a related sequence of cut-off functions $\eta_j \in C^{\infty}_0(A_{j+1})$ such that $\eta_j \equiv 1$ on A_j, and finally we define $w_j := \eta_j w_{1/j}$; clearly $w_{1/j} \in C^{\infty}_0(\Omega)$. Now let $A \subset\subset \Omega$, being an open subset; there exists $j_0 \in \mathbb{N}$ such that $A \subset A_j$ whenever $j \ge j_0$. For such values of j we observe that by the fact that ψ is nondecreasing with respect to the last variable, we find

$$\int_A \psi(x,|w_{1/j}(x)|)\,dx \le \int_A \psi(x,|(Mw)(x)|)\,dx \le C\int_A (\psi(x,|w(x)|)+1)\,dx,$$

$$\int_A \psi(x,|Dw_{1/j}(x)|)\,dx \le \int_A \psi(x,|(MDw)(x)|)\,dx \le C\int_A (\psi(x,|Dw(x)|)+1)\,dx,$$

where C depends also on $\int_A \psi(x,|w(x)|)\,dx$ and $\int_A \psi(x,|Dw(x)|)\,dx$; see Definition 4.4. Therefore, since

$$\psi(x,|w_{1/j}(x)|) \to \psi(x,|w(x)|) \qquad \text{and} \qquad \psi(x,|Dw_{1/j}(x)|) \to \psi(x,|Dw(x)|)$$

a.e., by Remark 4.2 such a convergence also holds in $L^1(A)$. Now we conclude, using the third property in (3.2), as follows:

$$\int_A \psi(x,|w_{1/j}(x)-w(x)|)\,dx \le c\int_A \psi(x,|w_{1/j}(x)|)\,dx + c\int_A \psi(x,|w(x)|)\,dx,$$

$$\int_A \psi(x,|Dw_{1/j}(x)-Dw(x)|)\,dx \le c\int_A \psi(x,|Dw_{1/j}(x)|)\,dx + c\int_A \psi(x,|Dw(x)|)\,dx,$$

and the conclusion follows from a well-known variant of Lebesgue's dominated convergence theorem. Finally, it easy to see that if $w \in L^1(\Omega;\mathbb{R}^N)$, then $w_{1/j} \to w$ in $L^1(\Omega;\mathbb{R}^N)$. □

Before stating the representation results, we recall that a Borel function $\varphi : \Omega \times \mathbb{R}^n \times \mathbb{R}^{N\times n} \to \mathbb{R}$ is called quasi-convex in the sense of Morrey [41] if, for a.e. $x_0 \in \Omega$, every $u_0 \in \mathbb{R}^n$, $z_0 \in \mathbb{R}^{N\times n}$, every bounded open set A of $\mathbb{R}^n$, and every function $\phi \in C_0^1(A;\mathbb{R}^N)$, we have

$$|A|\,\varphi(x_0,u_0,z_0) \le \int_A \varphi(x_0,u_0,z_0+D\phi(x))\,dx.$$

Moreover, the quasi-convex envelope Qf of a function $f(x,u,z)$ is the greatest function $\varphi(x,u,z)$, which is quasi-convex, being less than or equal to f (see [14], [15]).

THEOREM 4.7. *Under the hypotheses of Theorem* 3.3*, suppose, in particular, that* $\psi(x,t)$ *satisfies the density property; see Definition* 4.5*. Then for every* $A \in \mathcal{A}$ *we have*

$$\overline{F}(u,A) = \begin{cases} \displaystyle\int_A \varphi(x,Du(x))\,dx & \text{if } u \in W^{\psi}_{\mathrm{loc}}(A;\mathbb{R}^N), \\ +\infty & \text{elsewhere on } L^1(\Omega;\mathbb{R}^N), \end{cases} \tag{4.3}$$

where $\varphi : \Omega\times\mathbb{R}^{N\times n} \to [0,+\infty)$ *is a quasi-convex function satisfying growth condition* (3.1) *for a.e.* $x \in \Omega$ *and all* $z \in \mathbb{R}^N$.

Example 4.8. In case $\psi(x,|z|) := |z|^{p(x)}$, let $p : \Omega \to]1,+\infty)$ be a continuous function satisfying the following local estimate about the modulus of continuity: for all $A \in \mathcal{A}_0$,

$$\exists\,\gamma_A > 0 : \quad |p(x)-p(y)| \le \frac{\gamma_A}{|\log|x-y||} \quad \forall\, x,y \in A\,, \quad 0<|x-y|<\frac{1}{2}. \tag{4.4}$$

Then in Proposition 5.2 we show that $\psi(x,|z|)$ satisfies the maximal property and therefore the density property (this result is actually contained in [19] and extended here to a more general class of functions). As a consequence, Theorem 4.7 holds. Similarly, in case $\psi(x,|z|) := |z|^p + a(x)\,|z|^q$, suppose in particular that $a(x)$ is a bounded nonnegative Hölder continuous function in $C^{0,\alpha}(\Omega)$, for some $0<\alpha\le 1$, and

$$1 < p \le q \le \frac{n+\alpha}{n}\,p. \tag{4.5}$$

Then from Proposition 5.1 it follows that the function $|z|^p + a(x)\,|z|^q$ satisfies the maximal property and Theorem 4.7 holds also in this case.

In order to prove Theorem 4.7, we make use of the following readaptation of the classical integral representation theorem [11, Thm. 1.1] in the setting of W^{ψ}-spaces.

PROPOSITION 4.9. *Suppose* $\psi(x,t)$ *is as in Theorem* 4.7*. Let* $\mathcal{F} : L^1(\Omega;\mathbb{R}^N)\times\mathcal{A} \to [0,+\infty]$ *satisfy the following conditions:*

(i) (locality) $\mathcal{F}$ *is local, i.e.,* $\mathcal{F}(u,A) = \mathcal{F}(v,A)$ *for every* $A \in \mathcal{A}$ *and* $u,v \in L^1(\Omega;\mathbb{R}^N)$ *with* $u = v$ *a.e. on* A*;*

(ii) (measure property) *for all* $u \in L^1(\Omega;\mathbb{R}^N)$ *the set function* $\mathcal{F}(u,\cdot)$ *is increasing, and is the trace on* $\mathcal{A}$ *of a Borel measure;*

(iii) (growth conditions) *there exist* $\beta > 0$ *and* $b(x) \in L^1_{\mathrm{loc}}(\Omega)$ *such that*

$$0 \le \mathcal{F}(u,A) \le \int_A (b(x) + \beta\,\psi(x,|Du(x)|))\,dx$$

for all $u \in W^{\psi}(\Omega;\mathbb{R}^N)$ *and* $A \in \mathcal{A}$*;*

(iv) (translation invariance in u) $\mathcal{F}(u+c,A)=\mathcal{F}(u,A)$ *for all* $u\in L^1(\Omega;\mathbb{R}^N)$, $A\in\mathcal{A}$, $c\in\mathbb{R}^N$;

(v) (lower semicontinuity) $\mathcal{F}(\cdot,A)$ *is sequentially lower semicontinuous with respect to the strong convergence in* $L^1(\Omega;\mathbb{R}^N)$ *for all* $A\in\mathcal{A}$.

Then there exists a Carathéodory function $\varphi:\Omega\times\mathbb{R}^{N\times n}\to[0,+\infty)$ *such that*

$$\mathcal{F}(u,A)=\int_A\varphi(x,Du(x))\,dx \tag{4.6}$$

for every $A\in\mathcal{A}$ *and for every* $u\in L^1(\Omega;\mathbb{R}^N)$ *such that* $u_{|A}\in W^{\psi}_{\rm loc}(A;\mathbb{R}^N)$*; in addition, the function* $\varphi(x,\cdot)$ *is quasi-convex in* $\mathbb{R}^{N\times n}$ *for a.e.* $x\in\Omega$ *and satisfies the growth condition*

$$0\le\varphi(x,z)\le b(x)+\beta\,\psi(x,|z|) \tag{4.7}$$

for a.e. $x\in\Omega$ *and all* $z\in\mathbb{R}^{N\times n}$.

Proof. We recall that a function $u\in L^1(\Omega;\mathbb{R}^N)$ is piecewise affine in Ω if there exists a countable family $\{\Omega_i\}_{i\in I}$ of disjoint open subsets of Ω and a Borel subset N of Ω with $|N|=0$ such that $\Omega=(\bigcup_{i\in I}\Omega_i)\bigcup N$ and $u_{|\Omega_i}$ is affine on each Ω_i.

Step 1: Following [16, Thm. 20.1] or [9, Thm. 9.1], we find a Carathéodory function φ satisfying (4.7), such that (4.6) holds for all $A\in\mathcal{A}$ and all piecewise affine on $u\in W^{\psi}(\Omega)$.

Step 2: $\mathcal{F}(u,A)\le\int_A\varphi(x,Du(x))\,dx$ *for* $u\in W^{\psi}(\Omega;\mathbb{R}^N)$ *and* $A\in\mathcal{A}$. By Step 1 and in particular by (4.7), we have that for every $A'\in\mathcal{A}_0$ the functional

$$u\mapsto\int_{A'}\varphi(x,Du(x))\,dx \tag{4.8}$$

is continuous with respect to the $W^{\psi}_{\rm loc}(\Omega)$ convergence (Definition 4.1). Moreover, by the density property of ψ, the following density result can be achieved via the approximation argument in [21, Chap. X, Prop. 2.1].

LEMMA 4.10. *If* ψ *satisfies the hypotheses of Theorem* 4.7*, then for every function* u *in* $W^{\psi}(\Omega;\mathbb{R}^N)$ *there exists a sequence* $\{u_j\}\subset W^{\psi}(\Omega;\mathbb{R}^N)$ *of functions that are piecewise affine on* Ω *and such that* $u_j\to u$ *both in* $L^1(\Omega;\mathbb{R}^N)$ *and in* $W^{\psi}_{\rm loc}(\Omega;\mathbb{R}^N)$.

Now, let $u\in W^{\psi}(\Omega)$ and $A\in\mathcal{A}$. By Lemma 4.10 there exists a sequence $\{u_j\}$ of functions in $W^{\psi}(\Omega)$ that are piecewise affine on Ω and such that $u_j\to u$ in $L^1(\Omega)$ and in $W^{\psi}_{\rm loc}(\Omega)$. Then by lower semicontinuity v) of $\mathcal{F}$, Step 1 and the continuity of the functional (4.8) in $W^{\psi}_{\rm loc}(\Omega)$, we obtain for every $A'\in\mathcal{A}_0$, $A'\subset\subset A$,

$$\mathcal{F}(u,A')\le\liminf_{j\to+\infty}\mathcal{F}(u_j,A')=\lim_{j\to+\infty}\int_{A'}\varphi(x,Du_j(x))\,dx=\int_{A'}\varphi(x,Du(x))\,dx.$$

Since $\mathcal{F}(u,\cdot)$ is a measure, taking the limit as $A'\nearrow A$ we get by the monotone convergence theorem

$$\mathcal{F}(u,A)\le\int_A\varphi(x,Du(x))\,dx \tag{4.9}$$

for every $u\in W^{\psi}(\Omega)$ and $A\in\mathcal{A}$.

Step 3: $\mathcal{F}(u,A)=\int_A\varphi(x,Du(x))\,dx$ *for* $u\in W^{\psi}(\Omega;\mathbb{R}^N)$ *and* $A\in\mathcal{A}$. Fix $u\in W^{\psi}(\Omega)$ and let $A,A'\in\mathcal{A}$ with $A'\subset\subset A$. We modify the function u in the

following way: take $A'' \in \mathcal{A}_0$ such that $A' \subset\subset A'' \subset\subset \Omega$, let ϕ be a cut-off function between A' and A'', and set $\widetilde{u} := \phi u$. Since $\widetilde{u}$ has compact support, by (3.2) we obtain that $\widetilde{u} \in W^\psi(\Omega)$ and that $\widetilde{u} + v \in W^\psi(\Omega)$ for every $v \in W^\psi(\Omega)$. Consider the functional $G : L^1(\Omega) \times \mathcal{A} \to [0, +\infty]$ defined by

$$G(v, B) := \mathcal{F}(v + \widetilde{u}, B).$$

Then G satisfies all hypotheses of Proposition 4.9. Indeed, (i), (ii), (iv), and (v) are trivially satisfied, whereas for all $v \in W^\psi(\Omega)$ and all $B \in \mathcal{A}$ we have

$$\begin{aligned} 0 \le G(v, B) = \mathcal{F}(v + \widetilde{u}, B) &\le \int_B (b(x) + \beta\, \psi(x, |D\widetilde{u} + Dv|))\, dx \\ &\le \int_B (g(x) + \gamma\, \psi(x, |Dv|))\, dx, \end{aligned}$$

where $\gamma = 2^{q-1} c\, \beta$ and $g(x) = b(x) + 2^{q-1}\, c\, \psi(x, |D\widetilde{u}(x)|) \in L^1_{\mathrm{loc}}(\Omega)$, with $c = c\,(A'')$ and $q = q(A'')$ given by (3.2). Therefore, from Steps 1 and 2 above, it follows that there exists a Carathéodory function $g : \Omega \times \mathbb{R}^{N \times n} \to [0, +\infty)$, satisfying (4.7) with γ and $g(x)$ instead of β and $b(x)$, such that

$$G(v, B) \le \int_B g(x, Dv(x))\, dx \qquad \forall\, v \in W^\psi(\Omega)\,, \quad \forall\, B \in \mathcal{A}\,, \tag{4.10}$$

with equality for v piecewise affine in Ω. In addition, arguing as for (4.8), we can prove that for every $B' \in \mathcal{A}_0$ the functional

$$v \mapsto \int_{B'} g(x, Dv(x))\, dx \tag{4.11}$$

is continuous in $W^\psi_{\mathrm{loc}}(\Omega)$. We now prove that

$$\mathcal{F}(u, A') = \int_{A'} \varphi(x, Du(x))\, dx\,; \tag{4.12}$$

since $\mathcal{F}(u, \cdot)$ is a measure, taking $A' \nearrow A$ we will obtain (4.6) for all $A \in \mathcal{A}$ and $u \in W^\psi(\Omega)$. By Lemma 4.10 there exists a sequence $\{u_j\}$ of functions in $W^\psi(\Omega)$, piecewise affine in Ω, such that $u_j \to \widetilde{u}$ in $L^1(\Omega)$ and in $W^\psi_{\mathrm{loc}}(\Omega)$. Then, using the locality (i) of $\mathcal{F}$, Steps 1 and 2, (4.10), and the continuity of the functionals (4.8) and (4.11), we obtain

$$\begin{aligned} \int_{A'} g(x, 0)\, dx &= G(0, A') = \mathcal{F}(\widetilde{u}, A') = \mathcal{F}(u, A') \le \int_{A'} \varphi(x, Du)\, dx \\ &= \int_{A'} \varphi(x, D\widetilde{u})\, dx = \lim_{j \to +\infty} \int_{A'} \varphi(x, Du_j)\, dx = \lim_{j \to +\infty} \mathcal{F}(u_j, A') \\ &= \lim_{j \to +\infty} G(u_j - \widetilde{u}, A') \le \lim_{j \to +\infty} \int_{A'} g(x, D(u_j - \widetilde{u}))\, dx \\ &= \int_{A'} g(x, 0)\, dx \end{aligned}$$

and (4.12) is proved.

Step 4: $\mathcal{F}(u, A) = \int_A \varphi(x, Du(x))\, dx$ for $u_{|A} \in W^\psi_{\mathrm{loc}}(A; \mathbb{R}^N)$ and $A \in \mathcal{A}$. If $u \in L^1(\Omega)$, $A \in \mathcal{A}$, and $u_{|A} \in W^\psi_{\mathrm{loc}}(A)$, then for every $A' \in \mathcal{A}_0$, $A' \subset\subset A$, we can

find a function $v \in W^{\psi}(\Omega)$ such that $v_{|A'} = u_{|A'}$ (it suffices to take $v = \phi u$, where $\phi \in C_0^{\infty}(\Omega)$ is a cut-off function between A' and A'', with $A' \subset\subset A'' \subset\subset A$). Then, by the locality of $\mathcal{F}$ and Step 3, we have

$$\mathcal{F}(u, A') = \mathcal{F}(v, A') = \int_{A'} \varphi(x, Dv(x))\, dx = \int_{A'} \varphi(x, Du(x))\, dx,$$

and we obtain the assertion as $A' \nearrow A$, by the measure property of $\mathcal{F}$.

Step 5*: Quasi-convexity of* φ. It is enough to prove that, for every $A \in \mathcal{A}_0$ with a Lipschitz boundary, $\varphi(x, \cdot)$ is quasi-convex on $\mathbb{R}^{N\times n}$ for a.e. $x \in A$. If $A \in \mathcal{A}_0$ is fixed and $c = c(A)$, $q = q(A)$, by (3.2) the restriction $\varphi : A \times \mathbb{R}^{N\times n} \to [0, +\infty)$ is a Carathéodory integrand with

$$0 \le \varphi(x, z) \le b(x) + \beta\, \psi(x, |z|) \le b(x) + \beta\, (c\, |z|^q + 1),$$

where $b(x) \in L^1(A)$. In addition, by lower semicontinuity (v), the functional (4.8) is sequentially weakly lower semicontinuous on $W^{1,q}(A)$; hence by [10, Thm. 4.1.5] we obtain that $f(x, \cdot)$ is quasi-convex in $\mathbb{R}^{N\times n}$ for a.e. $x \in A$. □

Proof of Theorem 4.7. We apply Proposition 4.9 to the relaxed functional $\overline{F}(u, A)$. Indeed, the locality property (i) is well known (see e.g., [16, Prop. 16.15]), the measure property (ii) is proved in Theorem 3.3, and growth condition (iii) follows from (3.21), whereas (iv) and (v) are trivially satisfied. Therefore, Proposition 4.9 implies that (4.6) holds for all $u \in W^{\psi}_{\mathrm{loc}}(A)$ and $A \in \mathcal{A}$, with φ quasi-convex in z. Finally (3.21) and (3.8) yield the growth estimate (3.1) for φ and the integral representation (4.3) on all of $L^1(\Omega)$. □

We are now able to state the following theorem.

THEOREM 4.11. *Under the hypotheses of Theorem* 4.7*, suppose that* ψ *enjoys the maximal property. Then, if* f *is a Carathéodory integrand, the function* φ *in* (4.3) *is equal to the quasi-convex envelope of* $z \mapsto f(x, z)$.

This is an easy consequence of Theorem 4.7 and of the following lower semicontinuity result, which we state in full generality.

THEOREM 4.12 (lower semicontinuity). *Under the hypotheses of Theorem* 4.7*, suppose that* ψ *enjoys the maximal property. Let* $\varphi : \Omega \times \mathbb{R}^N \times \mathbb{R}^{N\times n} \to [0, +\infty)$ *be a quasi-convex Carathéodory function satisfying*

$$0 \le \varphi(x, u, z) \le b(x) + C\, (\psi(x, |u|) + \psi(x, |z|)), \tag{4.13}$$

where $C > 1$ *and* $b(x) \in L^1_{\mathrm{loc}}(\Omega)$ *with* $b(x) \ge 0$. *Then for every sequence* $\{u_k\} \subset W^{1,1}(\Omega; \mathbb{R}^N)$ *with* $u_k \to u$ *in* $L^1(\Omega; \mathbb{R}^N)$ *and*

$$\sup_k \int_{\Omega} \psi(x, |Du_k(x)|)\, dx < +\infty \tag{4.14}$$

we have that $u \in W^{\psi}_{\mathrm{loc}}(\Omega; \mathbb{R}^N)$, $\psi(x, |Du|) \in L^1(\Omega)$, *and*

$$\int_{\Omega} \varphi(x, u(x), Du(x))\, dx \le \liminf_{k\to+\infty} \int_{\Omega} \varphi(x, u_k(x), Du_k(x))\, dx. \tag{4.15}$$

Proof. Our starting point is the classical lower semicontinuity proof of Acerbi and Fusco for quasi-convex integrals with p-growth. Hence we refer to the proof of [2, Thm. II.4], where we will point out the differences. The main ingredients are the

density property in $W^\psi(\Omega;\mathbb{R}^N)$ and the fact that ψ enjoys the maximal property; see Definitions 4.5 and 4.4. Set now for every Borel set $A \subset \Omega$

$$\mathcal{F}(u,A) := \int_A \varphi(x,u(x),Du(x))\,dx.$$

We divide the rest of the proof into four steps.

Step 1: $u \in W^\psi_{\mathrm{loc}}(\Omega;\mathbb{R}^N)$ and $\int_\Omega \psi(x,|Du|)\,dx < +\infty$. Setting $\Omega_j := \{x \in \Omega \mid |x| < j \text{ and } \operatorname{dist}(x,\partial\Omega) > 1/j\} \in \mathcal{A}_0$, if $p = p(\Omega_j) > 1$ is given by (3.2), passing to a subsequence we have that $u_k \rightharpoonup u$ weakly in $W^{1,p}(\Omega_j)$, and hence weakly in $W^{1,1}(\Omega_j)$. Then we can apply Theorem 2.5 with $A = \Omega_j$ and $g(x,u,z) = \psi(x,|z|)$ to obtain

$$\int_{\Omega_j} \psi(x,|Du|)\,dx \le \liminf_{k\to+\infty} \int_{\Omega_j} \psi(x,|Du_k|)\,dx$$

for every j. Hence (4.14) gives $\psi(x,|Du|) \in L^1(\Omega)$ and finally (3.3) yields $u \in W^\psi_{\mathrm{loc}}(\Omega)$.

Step 2: Preliminary reductions. Since the supremum of lower semicontinuous functions is lower semicontinuous, we can restrict to prove (4.15) on a ball (or a hypercube) compactly contained in Ω. Hence, relabelling by Ω such ball, which we shall take for the sake of simplicity as B_1, and possibly passing to a subsequence, which we relabel $\{u_k\}$, we can suppose that the lower limit in (4.15) is a finite limit. Moreover, we can suppose that (3.2) holds on the whole of Ω. Then, setting $z_k := u_k - u$, by (3.2) and Step 1 we have that (4.14) holds for $\{z_k\}$. Hence, the Sobolev-type property (Definition 3.2) yields that $\{z_k\} \subset W^\psi(\Omega)$ and there exists $M < +\infty$ such that

$$\sup_k \int_\Omega (\psi(x,|z_k|) + \psi(x,|Dz_k|))\,dx < M. \tag{4.16}$$

By applying the density property (Definition 4.5) to z_k, since φ is a Carathéodory function satisfying (4.13), by the Dominated convergence theorem, we can find for each k a sequence $\{w_j\} \subset C^\infty(\Omega)$ such that $w_j \to z_k$ in $L^1(\Omega)$ as $j \to +\infty$ and

$$\lim_{j\to+\infty} \mathcal{F}(u+w_j,A) = \mathcal{F}(u+z_k,A) \qquad \forall\, A \subset\subset \Omega.$$

Again, using the fact that the supremum of a family of lower semicontinuous integrals is semicontinuous, taking a smaller ball Ω, we can finally assume the sequence $\{z_k\}$ to be in $C^\infty(\Omega)$.

Step 3: The case $\operatorname{supp} z_k \subset \Omega$. First, we extend each z_k to the whole $\mathbb{R}^n$ by letting $z_k \equiv 0$ outside Ω. We define (according to [2])

$$(M^* z_k)(x) := (Mz_k)(x) + \sum_{i=1}^n (MD_i z_k)(x). \tag{4.17}$$

We observe that if the support of u is contained in Ω, then $(Mv)(x)$ as defined in Definition 4.4 coincides with the standard maximal function as employed in [2].

By (3.2), (4.16), (4.17), and the fact that ψ enjoys the maximal property (Definition 4.4), we have

$$\sup_k \int_\Omega \psi(x,(M^* z_k^{(i)})(x))\,dx < +\infty \qquad \forall\, i = 1,\dots,N,$$

where $z_k = (z_k^{(1)}, \dots, z_k^{(N)})$, hence we can apply the Biting lemma [2, Lemma I.7] to obtain for each $\epsilon > 0$ a (not relabelled) subsequence $\{z_k\}$, a set $A_\epsilon \subset \Omega$, with $|A_\epsilon| < \epsilon$, and a real number $\delta > 0$ such that

$$\sup_k \int_B \psi(x, (M^* z_k^{(i)})(x))\, dx < \epsilon \qquad \forall\, i = 1, \dots, N \tag{4.18}$$

for every Borel set $B \subset \Omega \setminus A_\epsilon$ with $|B| < \delta$. Also, let $\eta : \mathbb{R}^+ \to \mathbb{R}^+$ be a continuous increasing function, with $\eta(0) = 0$, such that for every measurable set $B \subset \Omega$

$$\int_B [b(x) + C(\psi(x, |u(x)|) + \psi(x, |Du(x)|))]\, dx < \eta(|B|). \tag{4.19}$$

From this point on we shall closely follow the proof of Theorem II.4 from [2]. Once λ, $H^\lambda_{i,k}$, H^λ_k, $g_k^{(i)}$, $v^{(i)}$, g_k, and v are defined as in [2, Thm. II.4], by (4.19), (4.18), and growth condition (4.13), since $|(\Omega \setminus A_\epsilon) \setminus H^\lambda_{i,k}| < \delta$, if q and c are given by (3.2) with $A = \Omega$, we obtain

$$\begin{aligned}
&\mathcal{F}(u + g_k, (\Omega \setminus A_\epsilon) \setminus H^\lambda_k)\\
&\le c\, 2^{q-1} \left\{ \eta(N\epsilon) + c\,(n, \Omega) \int\limits_{(\Omega \setminus A_\epsilon) \setminus H^\lambda_k} \psi(x, \lambda)\, dx \right\}\\
&\le c\, 2^{q-1} \left\{ \eta(N\epsilon) + c\,(n, \Omega) \sum_{i=1}^N \int\limits_{(\Omega \setminus A_\epsilon) \setminus H^\lambda_{i,k}} \psi(x, (M^* z_k^{(i)})(x))\, dx \right\}\\
&\le c\, 2^{q-1} \{ \eta(N\epsilon) + N \cdot c\,(n, \Omega)\, \epsilon \} = o_\epsilon,
\end{aligned}$$

where $o_\epsilon \to 0$ when $\epsilon \to 0$; this last estimate replaces the one at the top of p. 131 in [2]. The rest of the proof in this case follows [2, Thm. II.4].

Step 4: The general case $\{z_k\} \subset C^\infty(\Omega)$. In the following we adopt the notation of Lemma 2.6. We fix $0 < s < t < 1$ and take $\epsilon \in (0, 1)$; according to Lemma 2.6 (applied to $f_k := \psi(x, |Dz_k|)$) we select $N \equiv N(\epsilon, M)$ and M is from (4.16) (recall that we already reduced to the case $\Omega \equiv B_1$). Therefore we find a thin layer A_h and a not-relabelled subsequence $\{z_k\}$ such that

$$\sup_k \int_{A_h} \psi(x, |Dz_k|) + \psi(x, |Du|))\ dx \le \epsilon. \tag{4.20}$$

Now we take a cut-off function η between B_{s_h} and $B_{s_{h+1}}$ such that $\|D\eta\| \le 2N/(t-s)$ and define $\tilde z_k := \eta z_k$. Since $D\tilde z_k = D\eta \otimes z_k + \eta\, Dz_k$ and by (3.2)

$$\int_{B_1} \psi(x, |D\tilde z_k|)\, dx \le c \int_{B_1} (\psi(x, |Dz_k|)\, dx + \psi(x, |z_k|))\, dx$$

for some absolute constant $c > 0$ possibly depending on B_1 and η, by (4.16) we obtain that (4.14) holds for $\{\tilde z_k\}$. Therefore, by Step 3 and condition $\operatorname{supp} \tilde z_k \subset B_{s_{h+1}}$,

$$\int_{B_s} f(x, Du)\, dx \le \liminf_{k \to +\infty} \int_{B_{s_{h+1}}} f(x, Du + D\tilde z_k)\, dx. \tag{4.21}$$

As a consequence, again by (3.2)

$$\begin{aligned}\int_{B_{s_{h+1}}} f(x,Du+D\tilde{z}_k)\,dx &= \int_{B_{s_h}} f(x,Du+Dz_k)\,dx+\int_{A_h} f(x,Du+D\tilde{z}_k)\,dx\\ &\le \int_{B_1} f(x,Du+Dz_k)\,dx\\ &\quad +c\int_{A_h}(\psi(x,|Du|)+\psi(x,|Dz_k|))\,dx\\ &\quad +c(\|D\eta\|_\infty)\int_{B_1}\psi(x,|z_k|)\,dx.\end{aligned}$$

Therefore, using (4.20) and combining with (4.21), since by the Rellich-type property $\int_{B_1}\psi(x,|z_k|)\,dx\to 0$ as $k\to+\infty$, we obtain

$$\int_{B_s} f(x,Du)\,dx \le \lim_{k\to+\infty}\int_{B_1} f(x,Du+Dz_k)\,dx+o_\epsilon,$$

with $o_\epsilon\to 0^+$ as $\epsilon\to 0^+$. Finally the full statement follows by first letting $\epsilon\to 0$ and then $s\to 1^-$. □

5. Continuity estimates for the maximal function. Throughout this section we always assume that Ω is a bounded open set. We will prove that in case $\psi(x,t)$ is equal to $t^{p(x)A(t)}$ or to $t^p+a(x)\,t^q$, under suitable hypotheses in both cases ψ enjoys the maximal property, as described in Definition 4.4.

PROPOSITION 5.1. *Let $0\le a(x)\le L$ be such that $a(x)\in C^{0,\alpha}(\Omega)$, where $0<\alpha\le 1$ and (4.5) holds. Then for every function $f\in L^1(\Omega)$ with*

$$\int_\Omega (|f(x)|^p+a(x)\,|f(x)|^q)\,dx<+\infty$$

we have

$$\begin{aligned}&\int_\Omega (|(Mf)(x)|^p+a(x)\,|(Mf)(x)|^q)\,dx\\ &\qquad\le \widetilde{C}\left(\int_\Omega (|f(x)|^p+a(x)\,|f(x)|^q)\,dx+1\right)^{q/p},\end{aligned}\tag{5.1}$$

where $\widetilde{C}$ is a positive constant depending on n, Ω, p, q, L, $[a]_{0,\alpha}$.

Proof. Let us first prove that for every $x\in\Omega$ and $r>0$,

$$a(x)\,|(M_{(r)}f)(x)|^q\le |(M_{(r)}(a(\cdot)^{1/q}f))(x)|^q+C\,|(M_{(r)}f)(x)|^p\cdot\|f\|^{q-p}_{L^p(\Omega)}.\tag{5.2}$$

Denoting $B=B_r(x)$ for simplicity and

$$a_r(x):=\inf\{a(y)\mid y\in\Omega,\,|y-x|<r\},$$

trivially if $|B\cap\Omega|>0$, then

$$a_r(x)\,|(M_{(r)}f)(x)|^q\le\left|\frac{1}{|B|}\int_{B\cap\Omega}a(y)^{1/q}|f(y)|\,dy\right|^q=|(M_{(r)}(a(\cdot)^{1/q}f))(x)|^q.\tag{5.3}$$

Moreover

$$(a(x)-a_r(x))\,|(M_{(r)}f)(x)|^q\le(a(x)-a_r(x))\,|(M_{(r)}f)(x)|^{q-p}\cdot|(M_{(r)}f)(x)|^p,\tag{5.4}$$

whereas by (4.5) we have $\alpha/(q-p) \geq n/p$, and hence, for $0 < r \leq 1$, we estimate $r^{\alpha/(q-p)} \leq r^{n/p}$. Then, by Hölder inequality, since $f \in L^p(\Omega)$,

$$(a(x) - a_r(x))\,|(M_{(r)}f)(x)|^{q-p} \leq [a]_{0,\alpha} \cdot |r^{n/p} \cdot |B_r|^{-1/p}\,\|f\|_{L^p(\Omega)}|^{q-p} \leq C\,\|f\|^{q-p}_{L^p(\Omega)}. \tag{5.5}$$

Also, (5.5) trivially holds if $r > 1$, since $a(x)$ is bounded. Now, (5.3), (5.4), and (5.5) yield (5.2) so that, taking the supremum on r and passing to the integrals on Ω, since $f \in L^p(\Omega)$ and $a(\cdot)^{1/q}\, f \in L^q(\Omega)$, by the standard Hardy–Littlewood maximal theorem [45, Thm. 1, Sec. I.1] we obtain

$$\int_\Omega (|(Mf)(x)|^p + a(x)\,|(Mf)(x)|^q)\,dx \leq C\,\left(\int_\Omega |f(x)|^p\,dx\right)^{q/p} + C\int_\Omega a(x)\,|f(x)|^q\,dx$$

and finally (5.1). □

The next proposition extends a result due to Diening [19].

PROPOSITION 5.2. *Let $A : \mathbb{R}^+ \to \mathbb{R}^+$ be a continuous function with $1 < s_1 \leq A(t) \leq s_2 < +\infty$ and such that $t \to t^{\bar p A(t)}$ is nondecreasing and convex in $\mathbb{R}^+$ for every $\bar p > 1$. Moreover, let $p : \Omega \to (1, +\infty)$ be a uniformly continuous function such that $\inf_\Omega p(x) > 1$ and for some $C_0 > 1$*

$$|p(x) - p(y)| \leq \frac{C_0}{|\log|x-y||} \quad \forall\, x, y \in \Omega\,, \quad 0 < |x-y| < \frac{1}{2}. \tag{5.6}$$

Set $1 < p := s_1 \inf_\Omega p(x) \leq s_2 \sup_\Omega p(x) =: q < +\infty$. Then, for every function $f \in L^1(\Omega)$ such that $\int_\Omega |f(x)|^{p(x)A(|f(x)|)}\,dx < +\infty$, we have

$$\int_\Omega |(Mf)(x)|^{p(x)A(|(Mf)(x)|)}\,dx \leq \widetilde{C}\,\left(\int_\Omega |f(x)|^{p(x)A(|f(x)|)}\,dx + 1\right)^{q/p}, \tag{5.7}$$

where $\widetilde{C}$ is a positive constant depending on n, Ω, p, q, s_1, s_2.

Remark 5.3. With a slightly different proof it is possible to obtain the following inequality:

$$\int_\Omega A(|(Mf)(x)|)\cdot|(Mf)(x)|^{p(x)}\,dx \leq \widetilde{C}\,\left(\int_\Omega A(|f(x)|)\cdot|f(x)|^{p(x)}\,dx + 1\right)^{q/p}$$

in the case $t \to t^{\bar p}\,A(t)$ is convex for any $\bar p > 1$ and $t^{s_1} \leq A(t) \leq L(1 + t^{s_2})$ where, this time, $s_2 \geq s_1 > 1$, $q := \sup p(x) + s_2$ and $p := \inf p(x) + s_1$.

Proof. First of all we can assume f is defined on the whole $\mathbb{R}^n$ by letting $f \equiv 0$ outside Ω so that

$$(M_{(r)}f)(x) = \frac{1}{|B_r(x)|}\int_{B_r(x)} |f(y)|\,dy$$

in any case; this will allow us to apply the Jensen inequality in the second inequality from (5.10). Let us first prove that for every $x \in \Omega$ and $r > 0$

$$|(M_{(r)}f)(x)|^{p(x)A(|(M_{(r)}f)(x)|)} \leq \gamma\,[(M_{(r)}|f(\cdot)|^{p(\cdot)A(|f(\cdot)|)})(x) + 1], \tag{5.8}$$

where

$$\gamma := c\,\left(\int_\Omega |f(x)|\,dx + 1\right)^{q-p} \tag{5.9}$$

and c depends on n, p, and q. Set, for every ball $B \subset \mathbb{R}^n$ with $|B \cap \Omega| > 0$,

$$p_B^- := \inf_{x \in B \cap \Omega} p(x), \qquad p_B^+ := \sup_{x \in B \cap \Omega} p(x),$$

and we have

$$|B|^{p_B^- - p_B^+} \le c(n, p, q).$$

Indeed, the previous inequality is trivial when $r > 1/4$ and is a consequence of (5.6) in the other case. Since $z \mapsto |z|^{p_B^- A(|z|)}$ is convex, and $f(\cdot)^{p_B^- A(|f(\cdot)|)} \in L^1(B)$, by Jensen inequality, denoting $B = B_r(x)$ for simplicity, we have

$$\begin{aligned} |(M_{(r)}f)(x)|^{p(x)A(|(M_{(r)}f)(x)|)} &= |(M_{(r)}f)(x)|^{(p(x)-p_B^-)A(|(M_{(r)}f)(x)|)} \\ &\times |(M_{(r)}f)(x)|^{(p_B^-)A(|(M_{(r)}f)(x)|)} \\ &\le \gamma\, |B|^{(p_B^- - p(x))s_2} |(M_{(r)}f)(x)|^{(p_B^-)A(|(M_{(r)}f)(x)|)} \qquad (5.10) \\ &\le c\,\gamma\, M_{(r)}[f(\cdot)^{p_B^- A(|f(\cdot)|)}] \\ &\le c\,\gamma\, M_{(r)}[f(\cdot)^{p(\cdot)A(|f(\cdot)|)}] + c. \end{aligned}$$

Now setting $q(x) := p(x)\, s_1/p > 1$, since $\inf_\Omega p(x) > 1$, then

$$p/s_1 > 1 \quad \text{and} \quad \int_\Omega \Big(|f(x)|^{q(x)A(|f(x)|)}\Big)^{p/s_1}\, dx < +\infty,$$

and we can use the boundedness of the maximal operator as follows:

$$\begin{aligned} \int_\Omega |(M|f(\cdot)|^{q(\cdot)A(|f(\cdot)|)})(x)|^{p/s_1}\, dx &\le c \int_\Omega (|f(x)|^{q(x)A(|f(x)|)})^{p/s_1}\, dx \qquad (5.11) \\ &= c \int_\Omega |f(x)|^{p(x)A(|f(x)|)}\, dx, \end{aligned}$$

where the above constant c depends on n, Ω, and p/s_1. Applying (5.8) (which is a pointwise inequality) with $q(x)$ instead of $p(x)$, we finally obtain, also using (5.9), (5.11), and Hölder inequality,

$$\begin{aligned} &\int_\Omega |(Mf)(x)|^{p(x)A(|(Mf)(x)|)}\, dx \\ &= \int_\Omega (|(Mf)(x)|^{q(x)A(|(Mf)(x)|)})^{p/s_1}\, dx \\ &\le c \left(\int_\Omega |f(x)|^p\, dx + 1\right)^{(q-p)/p} \int_\Omega (\,[(M|f(\cdot)|^{q(\cdot)A(|f(\cdot)|)})(x) + 1])^{p/s_1}\, dx \\ &\le c \left(\int_\Omega |f(x)|^p\, dx + 1\right)^{(q-p)/p} \cdot \int_\Omega (|f(x)|^{p(x)A(|f(x)|)} + 1)\, dx. \end{aligned}$$

Therefore (5.7) immediately follows as $p \le \inf_\Omega p(x)\, A(|f(x)|)$. ∎

Remark 5.4. It is interesting to note that conditions (4.5) and (5.6) are sharp in order to guarantee the validity of the maximal property, that is, (5.1) and (5.7), respectively. This is again a consequence of the counterexamples in sections 7 and 8 below. Indeed, suppose (4.5) and (5.6) fail to hold but (5.1) and (5.7) are satisfied; then by Proposition 4.6 the density property also holds true and in turn Theorem 4.7 would imply that $\overline{F}(u, A)$ is an absolutely continuous Radon measure as soon as $Du \in L^{\psi}(A; \mathbb{R}^N)$, in the case $F \equiv F_2$ and $F \equiv F_1$, respectively (see (1.6)). This is in contrast to the counterexamples presented in sections 7 and 8, where it is shown that, in general, the failure of (4.5) and (5.6) causes the rising of a singular Borel measure in the relaxation procedure (see, in particular, Theorem 7.4 for (4.5) and Theorem 8.3 for (5.6)). This observation, together with the forthcoming examples in sections 7 and 8, clarifies the unifying role of the continuity assumptions of the type (4.5) and (5.6).

6. Models and applications. In this section we want to outline how to apply the previous results to general classes of functionals, including many model examples available in the literature to which standard relaxation techniques do not apply. If $\{\psi_i\}_i$ is a finite collection of functions satisfying (i) and (ii) from section 3 together with the maximal property (and hence also satisfying the density property by Proposition 4.6), then the new function defined by

$$\overline{\psi}(x,t) := \sum_i a_i(x)\, \psi_i(x,t), \qquad L^{-1} \le a_i(x) \le L < +\infty,$$

also enjoys the same properties. Using this simple observation it immediately follows that the maximal estimates of the previous section allow the use of the model examples introduced there as building blocks to construct new functionals to which our theory applies. The main point we would like to stress here is that the model functionals presented in section 5 describe the way the presence of the variable x in the energy density f modifies the growth with respect to the gradient variable z. Using the previous observation, Theorem 4.12 may be applied, via the maximal estimates of section 5 and Proposition 4.6, in the cases when

$$\begin{aligned}
\psi_1(x,|z|) &:= a(x)\,|z|^{p(x)} \log(1+|z|)\,, \qquad L^{-1} \le a(x) \le L,\\
\psi_2(x,|z|) &:= A(|z|)^{p(x)}\,, \qquad |z| \le A(|z|) \le L(1+|z|),\\
\psi_3(x,|z|) &:= (e+|z|^2)^{p(x)(\theta_1+\theta_2 \sin\log\log(e+|z|^2))} \qquad \text{for suitable } \theta_1,\ \theta_2,\\
\psi_4(x,|z|) &:= f_p(x,|z|) + a(x) f_q(x,|z|)\,,\\
&\quad\ |z|^s \le f_s(x,|z|) \le L(1+|z|^s), \quad s = p, q, \quad 0 \le a(x) \le L.
\end{aligned}$$

In turn, any finite combination of ψ_i works, and so on. Let us observe that energies related to ψ_1 appear in the context of Prandtl–Eyring fluids (see [27]), while ψ_2 is related to electrorheological fluids (see [44] and [4]). The function ψ_3 has been studied in the setting of functionals with nonstandard growth conditions in [29] while $|z|^{p(x)}$ and ψ_4 have been introduced, in the context of homogenization theory, by Zhikov [46]. Finally we want to briefly mention that the results of the previous sections could be extended to the case of the so-called anisotropic functionals, i.e., functionals in which each direction is penalized with a different exponent. Functionals of this type come

up when studying reinforced materials. In this case (3.1) is replaced by

$$L^{-1}\sum_{i=1}^{n} a_i(x)|D_i u|^{p_i(x)} \le f(x,Du) \le L\left(1+\sum_{i=1}^{n} a_i(x)|D_i u|^{p_i(x)}\right), \qquad 1 \le L < +\infty,$$

where, in the models for reinforced materials, the exponents are constants: $p_i(x) \equiv p_i \equiv$ constant.

7. A sharp example with energy concentration. Let $\Omega = B_1$, the unit ball of $\mathbb{R}^n$, and $f(x,z) := |z|^p + a(x)\,|z|^q$; see Example 3.4, where $a(x)$ is a suitable bounded nonnegative function in $C^{0,\alpha}(B_1)$ for some $0 < \alpha \le 1$.

In this section we will first show (Theorem 7.4) that energy concentration does occur in the process of relaxation in the case (4.5) is violated, more precisely, when

$$1 < p < n < n+\alpha < q < p^*, \tag{7.1}$$

where, as usual, $p^* := np/(n-p)$. Second, if in particular

$$q > n\,(1+\alpha) \qquad \text{and} \qquad n\,\frac{1+\alpha}{2+\alpha} < p < n, \tag{7.2}$$

we are then able to give a complete representation of the relaxed functional (Theorem 7.6). We emphasize here that it is a significant feature of our analysis that the examples proposed in this section and the next already work in the scalar case $N = 1$, in which we specialize henceforth. For every $0 < \alpha \le 1$, we define

$$a(x) := \max\left\{\left(x_n^2 - \sum_{i=1}^{n-1} x_i^2\right), 0\right\}^{\alpha} |x|^{-\alpha}, \qquad x = (x_1,\dots,x_n) \in \mathbb{R}^n, \tag{7.3}$$

so that $a(x) \in C^{0,\alpha}(\mathbb{R}^n)$ and $a(x) > 0$ in the open cone

$$C^+ := \left\{ x \in \mathbb{R}^n \mid x_n^2 - \sum_{i=1}^{n-1} x_i^2 > 0 \right\}.$$

By (7.1) the assumptions of Theorem 3.3 are satisfied (see Example 3.4). Then in this section we denote by $F(u,A)$ and $\overline{F}(u,A)$ the functionals given by (2.1) and (2.2), respectively, with $\Omega = B_1$ and, when not specified differently, $f(x,z) := |z|^p + a(x)\,|z|^q$, where $a(x)$ is given by (7.3) and (7.1) holds, so that $\overline{F}(u,A)$ satisfies the measure property.

Remark 7.1. For every $u \in L^1(B_1)$ and $A \in \mathcal{A}$, it is possible to find a sequence $\{u_k\} \in L^1(B_1)$ with $u_k \to u$ in $L^1(B_1)$ and $u_{k|A} \in W^{1,q}(A)$ for every $k \in \mathbb{N}$. Moreover, since

$$f(x,z) \le |z|^p + \|a\|_{L^\infty(B_1)}\,|z|^q \le c(1+|z|^q),$$

by Remark 2.1, for every $A \in \mathcal{A}$ and $u \in L^1(B_1)$,

$$\overline{F}(u,A) = \inf\left\{ \liminf_{k\to+\infty} \int_A (|Du_k|^p + a(x)\,|Du_k|^q)\,dx \quad | \quad \{u_k\} \subset W^{1,q}_{\mathrm{loc}}(A), \; u_k \to u \text{ in } L^1(A) \right\}. \tag{7.4}$$

Let us now introduce some notation. If $n = 2$, we are going to use the following polar coordinates:

$$x_1 = \rho \sin \phi \,, \quad x_2 = \rho \cos \phi \,, \qquad \rho \geq 0 \,,\ 0 \leq \phi \leq 2\pi.$$

If $n \geq 3$, we use the spherical coordinate transformation $x = F(\rho, \phi, \Theta)$, $\Theta := (\theta_1, \ldots, \theta_{n-2})$, where $(\phi, \Theta) \in I(\phi, \Theta) := [0, \pi] \times \left(\prod_{i=1}^{n-3} [0, \pi] \right) \times [0, 2\pi]$ and

$$x_1 = \rho \sin \phi \cdot \prod_{j=1}^{n-2} \cos \theta_j \,, \qquad x_{n-1} = \rho \sin \phi \sin \theta_1 \,, \qquad x_n = \rho \cos \phi,$$
$$x_i = \rho \sin \phi \cdot \prod_{j=1}^{n-1-i} \cos \theta_j \cdot \sin \theta_{n-i} \,, \qquad i = 2, \ldots, n-2.$$

Moreover, for any function u on $\mathbb{R}^n$, in what follows we will always denote

$$\widetilde{u}(\rho, \phi, \Theta) := u(F(\rho, \phi, \Theta))$$

the corresponding function written in spherical coordinates. For example, if $a(x)$ is given by (7.3), we have

$$\widetilde{a}(\rho, \phi, \Theta) = (\rho \, (\cos(2\phi))^+)^\alpha, \qquad C^+ = \{x = F(\rho, \phi, \Theta) \mid \cos(2\phi) > 0\},$$

where y^+ denotes the positive part of real number y, i.e., $y^+ := \max\{y, 0\}$. Finally,

$$\partial C^+ := \left\{ x \in \mathbb{R}^n \mid x_n^2 = \sum_{i=1}^{n-1} x_i^2 \right\} = \{x = F(\rho, \phi, \Theta) \mid \cos(2\phi) = 0\}$$

is the boundary of C^+, for every $0 < \beta < \pi/4$

$$C^\beta := \{x = F(\rho, \phi, \Theta) \mid \cos(2\phi) > c_\beta\}, \qquad c_\beta := \cos\left(\frac{\pi}{2} - 2\beta \right)$$

is the subset of C^+ given by a cone of smaller angle and, for $0 < r < 1$,

$$C_r^+ := C^+ \cap B_r \,, \qquad\qquad C_r^\beta := C^\beta \cap B_r$$

is the intersection with the open ball B_r of radius r; moreover, we introduce the following "half cones":

$$\begin{aligned} {+C_r^\beta} &:= \{x \equiv F(\rho, \phi, \Theta) \in C_r^\beta \mid 0 \leq \phi < \pi/4 - \beta\}, \\ {-C_r^\beta} &:= \{x \equiv F(\rho, \phi, \Theta) \in C_r^\beta \mid 3\pi/4 + \beta < \phi \leq \pi\}, \end{aligned}$$

being the upper and the lower part, respectively, of C_r^β; accordingly, we define

$$\begin{aligned} {+C_r^+} &:= \{x \equiv F(\rho, \phi, \Theta) \in C_r^+ \mid 0 \leq \phi < \pi/4\}, \\ {-C_r^+} &:= \{x \equiv F(\rho, \phi, \Theta) \in C_r^+ \mid 3\pi/4 < \phi \leq \pi\}. \end{aligned}$$

The following result will allow us to consider the traces in the origin of a function with finite energy in the cone C_r^β (see (7.8)) in the case (7.1) holds.

LEMMA 7.2. *Let $u \in L^1(B_r)$, $0 < r < 1$, be such that $\int_{B_r} a(x) \, |Du|^q \, dx < +\infty$, where $a(x)$ is given by (7.3) and $q > n + \alpha$. Then for every $n < s < q$, with $q/s > (n + \alpha)/n$, we have $\int_{C_r^\beta} |Du|^s \, dx < +\infty$ for every $0 < \beta < \pi/4$. In particular,*

$u \in C^{0,1-n/s}(+\overline{C}_r^\beta)$ and $u \in C^{0,1-n/s}(-\overline{C}_r^\beta)$, i.e., there exists a constant c, depending only on $\int_{B_r} a(x)\,|Du|^q\,dx$ such that

$$\begin{aligned} &|u(x_1)-u(x_2)| \le c|x_1-x_2|^{1-n/s} && \forall\, x_1, x_2 \in +\overline{C}_r^\beta,\\ &|u(y_1)-u(y_2)| \le c|y_1-y_2|^{1-n/s} && \forall\, y_1, y_2 \in -\overline{C}_r^\beta. \end{aligned} \tag{7.5}$$

Proof. First note that $a(x)^{s/(s-q)} \in L^1(C_r^\beta)$. In fact, by the area formula [23, 3.2.3] we have

$$\begin{aligned} &\int_{C_r^\beta} a(x)^{s/(s-q)}\,dx\\ &\le c(n) \int_{[0,\pi/4-\beta]\cup[3\pi/4+\beta,\pi]} \frac{(\sin\phi)^{n-2}}{(\cos(2\phi))^{\alpha s/(q-s)}}\,d\phi \int_{C_r^\beta} |x|^{\alpha s/(s-q)}\,dx\\ &\le c(n)\, c_\beta^{\alpha s/(s-q)} \int_0^r \rho^{n-1+\alpha s/(s-q)}\,d\rho, \end{aligned}$$

which is finite since $n+\alpha s/(s-q) > 0$ if $q/s > (n+\alpha)/n$. Then by Hölder inequality we have

$$\int_{C_r^\beta} |Du|^s\,dx \le \left(\int_{C_r^\beta} a(x)|Du|^q\,dx\right)^{s/q} \cdot \left(\int_{C_r^\beta} a(x)^{s/(s-q)}\,dx\right)^{(q-s)/q} < +\infty. \tag{7.6}$$

The assertions concerning Hölder continuity follow via Sobolev embedding theorem and Morrey's theorem, since $s > n$. ∎

With a stronger assumption on the exponent q—that is, replacing (7.1) by (7.2)—we can similarly prove the following lemma.

LEMMA 7.3. *Under the hypotheses of Lemma 7.2, suppose, in particular, that $q > n\,(1+\alpha)$. Then for every $n < s < q/(1+\alpha)$ with $q/s > 1+\alpha$, we have $\int_{C_r^+} |Du|^s\,dx < +\infty$ and hence $u \in C^{0,1-n/s}(+\overline{C}_r^+)$ and $u \in C^{0,1-n/s}(-\overline{C}_r^+)$, with estimates analogous to (7.5) with $\beta = 0$.*

Proof. Now we have $a(x)^{s/(s-q)} \in L^1(C_r^+)$. In fact, for $n \ge 3$ ($n = 2$ is similar)

$$\begin{aligned} &\int_{C_r^+} a(x)^{s/(s-q)}\,dx\\ &= c(n) \int_{[0,\pi/4]\cup[3\pi/4,\pi]} \frac{(\sin\phi)^{n-2}}{(\cos(2\phi))^{\alpha s/(q-s)}}\,d\phi \int_0^r \rho^{n-1+\alpha s/(s-q)}\,d\rho, \end{aligned} \tag{7.7}$$

which is finite since $1/(\cos(2\phi))^+ \in L^{\alpha s/(q-s)}(0,2\pi)$ as $q/s > 1+\alpha$. Then (7.6) holds again, with C_r^+ instead of C_r^β. The rest follows as for Lemma 7.2. ∎

Traces at $0_{\mathbb{R}^n}$. Let $u \in L^1(B_1)$ be such that $a(x)\,|Du|^q \in L^1_{\mathrm{loc}}(A)$ for some open set $A \in \mathcal{A} = \mathcal{A}(B_1)$ with $0_{\mathbb{R}^n} \in A$. Since $B_r \subset\subset A$ for r sufficiently small, if $q > n+\alpha$ by Lemma 7.2, we can therefore define for every $0 < \beta < \pi/4$

$$\lambda_1 := \lim_{\substack{\rho\to 0^+\\ \phi\in[0,\pi/4-\beta]}} \widetilde{u}(\rho,\phi,\Theta)\,, \qquad \lambda_2 := \lim_{\substack{\rho\to 0^+\\ \phi\in[3\pi/4+\beta,\pi]}} \widetilde{u}(\rho,\phi,\Theta) \tag{7.8}$$

($\phi \in [0,\pi/4-\beta]\cup[7\pi/4+\beta,2\pi)$ and $\phi \in [3\pi/4+\beta, 5\pi/4-\beta]$, respectively, if $n = 2$), where the finite limits exist uniformly in Θ since u is Hölder continuous up to the closure of both $+C_r^\beta$ and $-C_r^\beta$. Moreover, if $q > n\,(1+\alpha)$ by Lemma 7.3, we obtain

(7.8) with $\beta = 0$, i.e., the traces in the origin do exist in both the upper and lower half cones of C^+.

We will now prove the following result, which actually shows that (4.5) is a sharp condition to prevent energy concentration in the process of relaxation.

THEOREM 7.4. *Let $F(u, A)$ and $\overline{F}(u, A)$ be given by* (2.1) *and* (2.2) *with $\Omega = B_1$ and $f(x, z) := |z|^p + a(x)\,|z|^q$, where $a(x)$ is given by* (7.3) *and* (7.1) *holds. Let $0_{\mathbb{R}^n} \in A \in \mathcal{A}$ and $|Du|^p + a(\cdot)\,|Du|^q \in L^1_{\rm loc}(A)$, so that* (7.8) *holds. Then, if $\lambda_1 \neq \lambda_2$, we have $\overline{F}(u, A) = +\infty$; hence an infinite singular measure is concentrated in the origin.*

Example 7.5. In particular, for $n \geq 3$, if $u_0 : B_1 \to \mathbb{R}$ is given in spherical coordinates by

$$\widetilde{u}_0(\rho, \phi, \Theta) := \begin{cases} 1 & \text{if} \quad 0 \leq \phi \leq \pi/4, \\ \sin(2\phi) & \text{if} \quad \pi/4 \leq \phi \leq 3\pi/4, \\ -1 & \text{if} \quad 3\pi/4 \leq \phi \leq \pi, \end{cases} \tag{7.9}$$

and similarly for $n = 2$, then, since $\lambda_1 = 1$ and $\lambda_2 = -1$, there is energy concentration in the origin, i.e.,

$$\overline{F}(u_0, A) = +\infty \qquad \forall\, A \in \mathcal{A} \quad \text{such that} \quad 0_{\mathbb{R}^n} \in A. \tag{7.10}$$

Proof of Theorem 7.4. We argue by contradiction supposing that $\overline{F}(u, A) < +\infty$. Then we pick a radius $r > 0$ such that $B_r \subset A$ and a sequence $\{u_k\} \subset C^1(B_r)$ such that $u_k \to u$ in $L^1(B_r)$ and a.e. and

$$\lim_{k \to +\infty} \int_{B_r} (|Du_k|^p + a(x)\,|Du_k|^q)\, dx = \overline{F}(u, B_r) < +\infty.$$

By the Hölder estimates in Lemma 7.2 we obtain that the sequence $\{u_k\} \subset C^1(B_r)$ is equi-uniformly continuous, on both $+\overline{C}^{\beta}_r$ and $-\overline{C}^{\beta}_r$; since each u_k is continuous, this yields that the sequence $\{u_k\} \subset C^1(B_r)$ is equi-uniformly continuous on the whole $\overline{C}^{\beta}_r$. Then by the Ascoli–Arzelà theorem, up to a not-relabelled subsequence, $u_k \to u$ uniformly on $\overline{C}^{\beta}_r$, which, in turn, yields to the continuity of u at $0_{\mathbb{R}^n}$. This is a contradiction since $\lambda_1 \neq \lambda_2$ makes the function u discontinuous at $0_{\mathbb{R}^n}$. □

With a bit more effort, if (7.2) holds we are able to prove the following complete representation result.

THEOREM 7.6. *Let $F(u, A)$ and $\overline{F}(u, A)$ be given by* (2.1) *and* (2.2) *with $\Omega = B_1$ and let $f(x, z)$ be a Carathéodory function such that*

$$c_1\,(|z|^p + a(x)\,|z|^q) \leq f(x, z) \leq c_2\,(|z|^p + a(x)\,|z|^q + 1) \tag{7.11}$$

for a.e. $x \in \Omega$ and all $z \in \mathbb{R}^n$, where $c_2 > c_1 > 0$. If $a(x)$ is given by (7.3) *and* (7.2) *holds, then we have*

$$\overline{F}(u, A) = \begin{cases} \displaystyle\int_A (Cf)(x, Du)\, dx + \mu(u, A) & \text{if } |Du|^p + a(\cdot)\,|Du|^q \in L^1(A), \\ +\infty & \text{elsewhere on } L^1(\Omega), \end{cases} \tag{7.12}$$

where Cf denotes the usual convexification of f and $\mu(u, \cdot)$ is an infinite singular measure concentrated in the origin. More precisely, we have

$$\mu(u, A) = \begin{cases} 0 & \text{if} \quad 0_{\mathbb{R}^n} \notin A, \\ \chi^{\lambda_1}_{\lambda_2} & \text{if} \quad 0_{\mathbb{R}^n} \in A, \end{cases} \tag{7.13}$$

where λ_1 *and* λ_2 *are defined by* (7.8) *and*

$$\chi_{\lambda_2}^{\lambda_1} := \begin{cases} 0 & \text{if} \quad \lambda_1 = \lambda_2, \\ +\infty & \text{if} \quad \lambda_1 \neq \lambda_2. \end{cases}$$

Proof. We will first give the proof in the case

$$f(x,z) := |z|^p + a(x)\,|z|^q.$$

The first part of the statement is trivial. In fact, following Lemma 3.5, Theorem 2.5 yields that if $\overline{F}(u,A) < +\infty$ for some $A \in \mathcal{A}$, then $|Du|^p + a(\cdot)\,|Du|^q \in L^1(A)$. Moreover, we note that for every $A \in \mathcal{A}$ with $A \subset\subset (B_1 \setminus \partial C^+)$ we have that $L^{-1} \leq a(x) \leq L$ on $A^+ := A \cap C^+$, for some positive constant L depending on A, whereas $a(x) = 0$ on $A \setminus A^+$. Hence, by convexity of $z \mapsto |z|^p + L\,|z|^q$ and by the dominated convergence theorem, if $\overline{F}(u,A) < +\infty$, we can easily find a sequence of smooth maps $\{u_k\} \in C^1(A)$ with $u_k \to u$ in $L^1(A)$ and

$$\lim_{k\to+\infty} \int_A (|Du_k(x)|^p + a(x)\,|Du_k(x)|^q)\,dx = \int_A (|Du(x)|^p + a(x)\,|Du(x)|^q)\,dx.$$

Then, by (7.4) and by inner regularity of $\overline{F}(u,\cdot)$, for every $A \in \mathcal{A}$ with $A \cap \partial C^+ = \emptyset$ we have that

$$\overline{F}(u,A) = \int_A (|Du(x)|^p + a(x)\,|Du(x)|^q)\,dx \tag{7.14}$$

if $u \in L^1(B_1)$ is such that $|Du|^p + a(\cdot)\,|Du|^q \in L^1(A)$. As a consequence, we infer that the absolute continuous part of the measure $\overline{F}(u,\cdot)$ is the integral given in (7.12), and that its singular part $\mu(u,\cdot)$ is concentrated in the $(n-1)$-dimensional cone ∂C^+.

We now show that there is no energy concentration on open sets which do not contain the origin.

PROPOSITION 7.7. *Under the hypotheses of Theorem* 7.6, *if* $A \in \mathcal{A}$, $0_{\mathbb{R}^n} \notin A$, *and* $|Du|^p + a(\cdot)\,|Du|^q \in L^1(A)$, *then* $\mu(u,A) = 0$ *in* (7.12) *and hence*

$$\overline{F}(u,A) = \int_A (|Du|^p + a(x)\,|Du|^q)\,dx.$$

Proof. We adapt the approximation and reflection arguments of Lemmas 3.4 and 3.5 in [22]. Indeed, following this paper it is possible to show that for every $A' \in \mathcal{A}_0$, with $A' \subset\subset A$, there exists a sequence of functions $\{u_k\} \subset W^{1,q}(A')$ such that $u_k \to u$ in $L^1(A')$ and

$$\lim_{k\to+\infty} \int_{A'} (|Du_k|^p + a(x)\,|Du_k|^q)\,dx = \int_{A'} (|Du|^p + a(x)\,|Du|^q)\,dx.$$

Then, by (7.4), this yields

$$\overline{F}(u,A') \leq \int_{A'} (|Du|^p + a(x)\,|Du|^q)\,dx,$$

and hence, by inner regularity, letting $A' \nearrow A$ one obtains the assertion by the fact that $\mu(u,A) \geq 0$. We explicitly remark that in [22] the proof is given for the case in which the function $a(x)$ is replaced (in polar coordinates) by

$$\widetilde{a}(\rho,\phi,\Theta) = \rho^\alpha \cos(2\phi)^+.$$

Then the proof is achieved, taking advantage of the fact that the function $\cos(2\phi)$ satisfies the so-called Muckenhoupt condition A_q; this gives the possibility to build an approximation procedure based on a reflection argument (where the Muckenhoupt property enters). It is easy to see that the same argumentation works here for the function $(\cos(2\phi))^\alpha$, which comes from the study of our case. □

Now let $0_{\mathbb{R}^n} \in A \in \mathcal{A}$ and $|Du|^p + a(\cdot)\,|Du|^q \in L^1_{\mathrm{loc}}(A)$. By Theorem 7.4, it follows that $\mu(u, A) = +\infty$ if $\lambda_1 \neq \lambda_2$ in (7.8). To conclude with (7.13), it then remains to show that $\mu(u, A) = 0$ if $\lambda_1 = \lambda_2$. To this aim, by (7.4) it suffices to prove the following proposition.

PROPOSITION 7.8. *Let $0_{\mathbb{R}^n} \in A \in \mathcal{A}$ and $u \in L^1(B_1)$ be such that $|Du|^p + a(\cdot)\,|Du|^q \in L^1_{\mathrm{loc}}(A)$, with $\lambda_1 = \lambda_2$ in (7.8). Then for each $\epsilon > 0$ there exists a sequence $\{w_k\} \subset W^{1,q}(A)$ such that $w_k \to u$ in $L^1(A)$ and*

$$\liminf_{k\to+\infty} \int_A (|Dw_k|^p + a(x)\,|Dw_k|^q)\,dx \le \int_A (|Du|^p + a(x)\,|Du|^q)\,dx + \epsilon. \tag{7.15}$$

Proof. Observe that we may and do assume that the right-hand side of (7.15) is finite. We will denote by ν the outward unit normal to ∂B_R and by $\tau := (\tau_1, \dots, \tau_{n-1})$ an orthonormal basis to the tangent $(n-1)$-space to ∂B_R. Then, setting $D_\tau u := (D_{\tau_1} u, \dots, D_{\tau_{n-1}} u)$, we have that $|Du|^2 = |D_\nu u|^2 + |D_\tau u|^2$. Also, if $u \in W^{1,p}(B_1)$ and $0 < R < 1$, we will denote by $\mathbf{T}_R u := \mathbf{T}[\partial B_R] u$ the usual trace operator: that is, $T_R u \in W^{1-\frac{1}{p},p}(\partial B_R)$ is the trace of u on ∂B_R.

Now fix $0 < \delta < \operatorname{dist}(0_{\mathbb{R}^n}, \partial A)$ and let $r \in (0, \delta/2)$. Then, by Remark 7.1 and Proposition 7.7, we select a sequence $\{u_k\} \subset W^{1,q}(A \setminus \overline{B}_r)$ such that $u_k \to u$ in $L^1(A \setminus \overline{B}_r)$ and

$$\begin{aligned} &\lim_{k\to+\infty} \int_{A\setminus B_r} (|Du_k|^p + a(x)\,|Du_k|^q)\,dx \\ &\qquad = \overline{F}(u, A \setminus \overline{B}_r) \\ &\qquad = \int_{A\setminus B_r} (|Du|^p + a(x)\,|Du|^q)\,dx < +\infty. \end{aligned} \tag{7.16}$$

Up to passing to a not-relabelled subsequence, by uniform convexity, (7.16) yields

$$\lim_{k\to+\infty} \int_{A\setminus B_r} (|Du_k - Du|^p + a(x)\,|Du_k - Du|^q)\,dx = 0.$$

In particular, by an estimate similar to (7.6), with $\beta = 0$, which is allowed since now (7.2) is in force (see also (7.7)), we have

$$\lim_{k\to+\infty} \int_{(B_{2r}\setminus B_r)\cap C^+} |Du_k - Du|^s\,dx = 0$$

for some $s > n$. As a consequence, by Sobolev, Morrey, and Rellich's theorems, passing again to a not-relabelled subsequence, we can select $R \in (r, 2r)$ such that $\mathbf{T}_R u \in W^{1,p}(\partial B_R) \cap W^{1,q}(\partial B_R \cap C^+)$, $\mathbf{T}_R u_k \in W^{1,q}(\partial B_R)$ for every k,

$$\begin{aligned} &\int_{\partial B_R} (|D_\tau u_k|^p + a(x)\,|D_\tau u_k|^q)\,d\mathcal{H}^{n-1} \\ &\qquad \le \int_{\partial B_R} (|D_\tau u|^p + a(x)\,|D_\tau u|^q)\,d\mathcal{H}^{n-1} + \frac{\epsilon}{3} R, \end{aligned} \tag{7.17}$$

$$\int_{\partial B_R} |u_k-\lambda|^p\,d\mathcal{H}^{n-1} \le \int_{\partial B_R} |u-\lambda|^p\,d\mathcal{H}^{n-1} + \frac{\epsilon}{3}\,R^{p-1}, \tag{7.18}$$

$$\int_{\partial B_R\cap C^+} |u_k-\lambda|^q\,d\mathcal{H}^{n-1} \le \int_{\partial B_R\cap C^+} |u-\lambda|^q\,d\mathcal{H}^{n-1} + \frac{\epsilon}{3}\,R^{q-\alpha-1}, \tag{7.19}$$

where $\lambda := \lambda_1 = \lambda_2$ is given by (7.8). Now define

$$v_k(x) := \begin{cases} u_k(x) & \text{if} \quad x \in A\setminus \overline{B}_R, \\ \dfrac{|x|}{R}\left(u_k\Big(R\dfrac{x}{|x|}\Big)-\lambda\right)+\lambda & \text{if} \quad x\in B_R. \end{cases} \tag{7.20}$$

Trivially $\{v_k\}\subset L^q(A)$ and $v_k\to u$ in $L^1(A\setminus B_R)$, whereas, since for a.e. $x\in B_R$

$$|Dv_k(x)|^2 = R^{-2}\left|u_k\Big(R\frac{x}{|x|}\Big)-\lambda\right|^2 + \left|D_\tau u_k\Big(R\frac{x}{|x|}\Big)\right|^2,$$

we infer

$$\int_{B_R}|Dv_k|^q\,dx \le c(q)\int_{\partial B_R}(R^{1-q}\cdot|u_k-\lambda|^q + R\cdot|D_\tau u_k|^q)\,d\mathcal{H}^{n-1}$$

and hence $\{v_k\}\subset W^{1,q}(A)$. We now show that, using the aforementioned information, for any $r\in(0,\delta/2)$ we can find $R\in(r/2,r)$ such that

$$\begin{aligned} &\liminf_{k\to+\infty}\int_A (|Dv_k|^p + a(x)\,|Dv_k|^q)\,dx \\ &\qquad \le \int_{A\setminus B_R}(|Du|^p + a(x)\,|Du|^q)\,dx + O(R) + \epsilon, \end{aligned} \tag{7.21}$$

where $O(R)\to 0^+$ as $R\to 0^+$. To this end, since $|a(x)|\le R^\alpha$ for $x\in B_R$, we first estimate

$$\begin{aligned} &\int_{B_R}(|Dv_k|^p + a(x)\,|Dv_k|^q)\,dx \\ &\le c\,(p,q)\Big\{R^{1-p}\int_{\partial B_R}|u_k-\lambda|^p\,d\mathcal{H}^{n-1} \\ &\quad + R^{1+\alpha-q}\int_{\partial B_R\cap C^+}|u_k-\lambda|^q\,d\mathcal{H}^{n-1} \\ &\quad + R\int_{\partial B_R}(|D_\tau u_k|^p + a(x)\,|D_\tau u_k|^q)\,d\mathcal{H}^{n-1}\Big\}. \end{aligned} \tag{7.22}$$

We now make use of the following embedding result (see [42, Lem. 5.8] for a proof).

LEMMA 7.9. *If $u\in W^{1,p}(B_\delta)$ with $1\le p<n$, $B_\delta\subset\mathbb{R}^n$ being the n-ball of radius δ, and $\lambda\in\mathbb{R}$, then for a.e. $0<R<\delta$ we have*

$$\begin{aligned} &R^{1-p}\int_{\partial B_R}|u-\lambda|^p\,d\mathcal{H}^{n-1} \\ &\le c\,(n,p)\left\{\int_{B_R}|Du|^p\,dx + \left(\int_{B_R}|u-\lambda|^{p^*}\,dx\right)^{p/p^*}\right\}, \end{aligned} \tag{7.23}$$

where $p^* := np/(n-p)$ *is the Sobolev conjugate of* p.

Now, condition $B_\delta \subset\subset A$ yields that $u(\cdot) - \lambda \in W^{1,p}(B_\delta)$. Then, by (7.18), (7.23), the Sobolev embedding theorem, and absolute continuity, we obtain

$$R^{1-p} \int_{\partial B_R} |u_k - \lambda|^p \, d\mathcal{H}^{n-1} \leq O(R) + \frac{\epsilon}{3}. \tag{7.24}$$

Recall now that since $a(\cdot)|Du|^q \in L^1(A)$, by Lemma 7.3 and Morrey's theorem [5, Thm. 5.4], since both $\pm C_R^+$ have Lipschitz boundaries, we have

$$|u(x) - u(y)| \leq c \, \|Du\|_{L^s(\pm C_R^+)} \, |x-y|^{1-n/s} \qquad \forall \, x, y \in \pm C_R^+,$$

where $c > 0$ is an absolute constant and $s > n$. In particular, by (7.8), with $\lambda = \lambda_1 = \lambda_2$, for every $x \in \partial B_R \cap C^+$ we then infer

$$|u(x) - \lambda| \leq c_0 \, \|Du\|_{L^s(C_R^+)} \, R^{1-n/s}. \tag{7.25}$$

Now, since by (7.7) (with $r = 1$)

$$c_2 := \|a(\cdot)^{-1}\|_{L^{s/(q-s)}(C_1^+)} = \left(\int_{C_1^+} a(x)^{s/(s-q)} \, dx \right)^{(q-s)/s} < +\infty,$$

by homogeneity of $a(x)$ we compute

$$\|a(\cdot)^{-1}\|_{L^{s/(q-s)}(C_R^+)} = c_2 \, R^{(n(q-s)-\alpha s)/s}. \tag{7.26}$$

Moreover, by (7.6) (with $\beta = 0$) and (7.26) we estimate

$$\begin{aligned} \|Du\|^q_{L^s(C_R^+)} &\leq \|a(\cdot)\,|Du|^q\|_{L^1(C_R^+)} \cdot \|a(\cdot)^{-1}\|_{L^{s/(q-s)}(C_R^+)} \\ &= \|a(\cdot)\,|Du|^q\|_{L^1(C_R^+)} \cdot c_2 \, R^{(n(q-s)-\alpha s)/s}. \end{aligned}$$

As a consequence, by (7.25) we have

$$|u(x) - \lambda|^q \leq {c_0}^q \, R^{q-nq/s} \, \|a(\cdot)\,|Du|^q\|_{L^1(C_R^+)} \cdot c_2 \, R^{(n(q-s)-\alpha s)/s}$$

for every $x \in \partial B_R \cap C^+$ and hence

$$\begin{aligned} &R^{1+\alpha-q} \int_{\partial B_R \cap C^+} |u - \lambda|^q \, d\mathcal{H}^{n-1} \\ &\leq c(n) \, R^{n+\alpha-q} \, {c_0}^q \, c_2 \, R^{q-nq/s} \, R^{(n(q-s)-\alpha s)/s} \, \|a(\cdot)\,|Du|^q\|_{L^1(C_R^+)} \\ &= C \, \|a(\cdot)\,|Du|^q\|_{L^1(C_R^+)}. \end{aligned}$$

Then, by absolute continuity and (7.19) we obtain

$$\begin{aligned} R^{1+\alpha-q} \int_{\partial B_R \cap C^+} |u_k - \lambda|^q \, d\mathcal{H}^{n-1} &\leq C \, \|a(\cdot)\,|Du|^q\|_{L^1(C_R^+)} + \frac{\epsilon}{3} \\ &\leq O(R) + \frac{\epsilon}{3}. \end{aligned} \tag{7.27}$$

Finally, since $|Du|^p + a(\cdot)\,|Du|^q \in L^1(B_\delta)$, setting

$$f(\rho) := \int_{\partial B_\rho} (|D_\tau u|^p + a(x)\,|D_\tau u|^q) \, d\mathcal{H}^{n-1}, \qquad 0 < \rho < \delta,$$

by the coarea formula, one has $f(\rho) \in L^1(0,\delta)$. Therefore, since $f(\rho) \geq 0$, we have $\liminf_{\rho\to 0^+} \rho \cdot f(\rho) = 0$. As a consequence, without loss of generality we can choose R so that $R \cdot f(R) = O(R)$ and hence, by (7.17),

$$R \int_{\partial B_R} (|D_\tau u_k|^p + a(x)\,|D_\tau u_k|^q)\, d\mathcal{H}^{n-1} \leq O(R) + \frac{\epsilon}{3}. \tag{7.28}$$

Then, by (7.24), (7.27), and (7.28), the right-hand side of (7.22) is smaller than $O(R) + \epsilon$ and, finally, by lower semicontinuity and (7.16), we obtain (7.21).

We finally make use of a diagonal argument, as follows. We first select $r_j \searrow 0$ and $R_j \in (r_j, 2r_j)$ as above; then for any fixed j via (7.20) we define $\{u_k^{(j)}\} \subset W^{1,q}(A \backslash B_{r_j})$ so that $u_k^{(j)} \to u$ in $L^1(A \setminus B_{r_j})$ and (7.16) holds with $r = r_j$; we then construct $\{v_k^{(j)}\} \subset W^{1,q}(A)$ such that $v_k^{(j)} \to u$ in $L^1(A \setminus B_{R_j})$ and (7.21) holds with $R = R_j$. Finally, we set $w_k := w_k^{(k)}$, so that $\{w_k\} \subset W^{1,q}(A)$, $w_k \to u$ in $L^1(A)$, and by (7.21)

$$\begin{aligned}
&\liminf_{k\to+\infty} \int_A (|Dw_k|^p + a(x)\,|Dw_k|^q)\, dx \\
&\leq \liminf_{k\to+\infty} \left\{ \int_{A\setminus B_{R_k}} (|Du|^p + a(x)\,|Du|^q)\, dx + O(R_k) + \epsilon \right\}
\end{aligned}$$

so that (7.15) holds, as required. □

End of the Proof of Theorem 7.6. In order to prove Theorem 7.6 for general integrands f, since we have shown that the density property (Definition 4.5) holds out of the origin, arguing as in Proposition 4.9, and taking into account that $Qf \equiv Cf$ in the scalar case $N = 1$, we obtain (7.12) where the singular measure $\mu(u,\cdot)$ is concentrated in the origin. Finally, (7.13) follows from growth condition (7.11). □

8. Another sharp example with energy concentration. In this section we describe another counterexample, involving probably the finest analysis of the paper; we show that if $\Omega = B_1$, the unit ball of $\mathbb{R}^2$, and $f(x,z) := |z|^{p(x)}$, where $p : \Omega \to (1,+\infty)$ is a suitable continuous exponent, energy concentration does occur in the process of relaxation in case (4.4) is violated: more precisely, following Zhikov [46] we set

$$p(x) := 2 + \frac{x_1 x_2}{|x|} \left(\log\left(\frac{2}{|x|} \right) \right)^{-t}, \qquad x = (x_1, x_2) \in B_1, \tag{8.1}$$

where $0 < t < 1$ is fixed. In this case the assumptions of Theorem 3.3 (see Example 3.4) hold, while those of Theorems 4.11 and 4.12 are not satisfied (see Example 4.8). Zhikov considered in [46] the homogeneous extension $\overline{u}(x) := \varphi(x/|x|)$ on B_1 of the function φ defined in standard polar coordinates $x = (\cos\theta, \sin\theta)$ on ∂B_1 by

$$\varphi(\theta) := \begin{cases} 1 & \text{if} \quad -\alpha \leq \theta \leq \pi/2 + \alpha, \\ 2(\theta + \alpha - \pi)/(4\alpha - \pi) & \text{if} \quad \pi/2 + \alpha \leq \theta \leq \pi - \alpha, \\ 0 & \text{if} \quad \pi - \alpha \leq \theta \leq 3\pi/2 + \alpha, \\ 2(\theta - \alpha - 3\pi/2)(\pi - 4\alpha) & \text{if} \quad 3\pi/2 + \alpha \leq \theta \leq 2\pi - \alpha, \end{cases} \tag{8.2}$$

where $0 < \alpha \ll \pi/4$ is a fixed small angle. Of course $|Du|^{p(x)} \in L^1(\Omega)$, but Zhikov showed that the Dirichlet problem for the $p(x)$-energy with boundary condition $\overline{u}$

on ∂B_1 is not regular, i.e., the infimum over $W^{1,p(x)}$-maps is strictly less that the infimum over smooth maps. For future purposes, we remark that the key point in Zhikov's argument is the summability near 0 of the function $\rho \mapsto \rho^{-1+c(\log(2\rho^{-1}))^{-t}}$ for any $0 < t < 1$ and $c > 0$, since for some $c_1, c_2 \equiv c_1, c_2(c) > 0$ we have

$$\begin{aligned} \int_0^1 \rho^{-1+c(\log(2\rho^{-1}))^{-t}}\, d\rho &\leq e^{c(\log 2)^{1-t}} \int_0^1 \rho^{-1}\, e^{-c\,(\log(2\rho^{-1}))^{1-t}}\, d\rho \\ &= c_1 \int_{\log 2}^{+\infty} e^{-c_2\, x^{1-t}}\, dx < +\infty. \end{aligned} \tag{8.3}$$

Moreover, for convenience, we also note that

$$\int_{\log 2}^{+\infty} x^t e^{-c_2\, x^{1-t}}\, dx < +\infty \qquad \forall\, t \in]0,1[\,, \quad c_2 > 0. \tag{8.4}$$

In this section we denote by $F(u, A)$ and $\overline{F}(u, A)$ the functionals given by (2.1) and (2.2), respectively, where $\Omega = B_1$ and $f : B_1 \times \mathbb{R}^2 \to [0, +\infty)$ is a Carathéodory function satisfying (8.7), where $p(x)$ is given by (8.1), so that $\overline{F}(u, A)$ satisfies the measure property (see Example 3.4). Since $p(x)$ satisfies (4.4) out of the origin (it is actually Lipschitz continuous far from the origin), we infer that energy concentration can occur only in $x = 0_{\mathbb{R}^2}$. More precisely, by Theorems 4.7 and 4.11 (see Example 4.8) we immediately obtain the following proposition.

PROPOSITION 8.1. *Let $u \in L^1(B_1)$ be such that $|Du|^{p(x)} \in L^1_{\mathrm{loc}}(A)$ for some open set $A \subset B_1$ with $0_{\mathbb{R}^2} \notin A$, where $p(x)$ is given by* (8.1). *Then*

$$\overline{F}(u, A) = \int_A |Du(x)|^{p(x)}\, dx. \tag{8.5}$$

Now we define, for every $0 < \beta < \pi/4$, the open cones

$$\begin{gathered} C^\beta \equiv +C^\beta := \{x = \rho\, e^{i\theta} \mid \beta < \theta < \pi/2 - \beta\}, \\ -C^\beta := \{-x \mid x \in C^\beta\}\,, \qquad \pm C^\beta_r := \pm C^\beta \cap B_r. \end{gathered}$$

Since inside $\pm C^\beta$ we have $p(x) > 2$ and $p(x) \to 2^+$ very rapidly as $x \to 0_{\mathbb{R}^2}$, we are able to define the traces in the origin of a function with finite energy in the cones $\pm C^\beta_r$, see (8.6). Of course we do not have at our disposal a standard estimate of the type in Lemma 7.2, since in our case $p(x) \to 2$ (the borderline case of Sobolev embedding) as $x \to 0$; anyway, we are able to prove the following theorem.

THEOREM 8.2 (trace theorem). *Let $u \in L^1(B_1)$ be such that $|Du|^{p(x)} \in L^1_{\mathrm{loc}}(A)$ for some open set $A \subset B_1$ with $0_{\mathbb{R}^2} \in A$, where $p(x)$ is given by* (8.1). *Then for every $0 < \beta < \pi/4$ the following finite limits exist:*

$$\lambda_1 := \lim_{\substack{x \to 0_{\mathbb{R}^2} \\ x \in \overline{C}^\beta_r}} u(x) \qquad \text{and} \qquad \lambda_2 := \lim_{\substack{x \to 0_{\mathbb{R}^2} \\ x \in -\overline{C}^\beta_r}} u(x). \tag{8.6}$$

In particular, if $r > 0$ is such that $B_r \subset\subset A$, we have that u is a continuous function up to the boundary of both the cones $\pm C^\beta_r$.

Thanks to Theorem 8.2, as in Theorem 7.4 we show that there is energy concentration in the origin if the traces in (8.6) take different values, for example, when

$$u(x) \equiv \overline{u}(x) := \varphi(x/|x|),$$

with φ given by (8.2).

THEOREM 8.3. *Let $F(u,A)$ and $\overline{F}(u,A)$ be given by* (2.1) *and* (2.2)*, with $\Omega = B_1$ and $f(x,z)$ being a Carathéodory function such that*

$$c_1\,|z|^{p(x)} \le f(x,z) \le c_2\,(|z|^{p(x)}+1), \tag{8.7}$$

where $p(x)$ is given by (8.1) *and $c_2 > c_1 > 0$. Moreover, let $u \in L^1(B_1)$ be such that $|Du|^{p(x)} \in L^1_{\rm loc}(A)$ for some open set $A \subset B_1$ with $0_{\mathbb{R}^2} \in A$. Then, if $\lambda_1 \neq \lambda_2$ in* (8.6)*, it follows that $\overline{F}(u,A) = +\infty$.*

Remark 8.4. In contrast to the previous section, this time we do not give the complete representation of the relaxed functional (that is, an analogue of Theorem 7.6), confining ourselves to emphasizing the main concentration phenomenon in the origin. This, for the sake of brevity: indeed, severe technical complications intervene in the upper bound estimate for the energy in the case $\lambda_1 = \lambda_2$. Regardless, it should be possible to obtain the same complete representation of the type in Theorem 7.6 also in this case.

Proof of Theorem 8.2. It is not restrictive to suppose

$$A = B_1 \qquad \text{and} \qquad \int_{B_1} |Du|^{p(x)}\,dx < +\infty.$$

Moreover, we will show the existence of the first limit in (8.6), the second limit being treated the same way. We remind the reader that in the following $c > 1$ continues to denote a constant possibly varying from line to line; we shall emphasize the relevant connections.

Step 1*: Dyadic type sequences.* We consider a sequence $\{y_k\} \subset \overline{C}^{\beta}_r$ of the type

$$y_k := r_k(\cos\theta_k, \sin\theta_k), \tag{8.8}$$

where $\theta_k \in [\beta, \pi/2-\beta]$ and $r_k \to 0^+$ is a decreasing sequence such that

$$L^{-1}/2^k \le r_k \le L/2^k \tag{8.9}$$

with $L \in [1,+\infty)$. We have

$$|y_k - y_{k+1}| \le |y_k|\cdot|\theta_k - \theta_{k+1}| + |r_k - r_{k+1}| \le (\pi+1)\,r_k. \tag{8.10}$$

By applying Morrey's theorem to the closure of the smooth set

$$S_k := C^{\beta} \cap B^{L2^{-k}}_{L^{-1}2^{-(k+1)}},$$

where $B^R_r := B_R \setminus B_r$, we have

$$|u(y_k) - u(y_{k+1})| \le c\,b_k\,|y_k - y_{k+1}|^{1-2/p_k}, \tag{8.11}$$

where

$$p_k = 2 + c_\beta\left(\log(L2^{k+2})\right)^{-t}, \qquad c_\beta := \frac{\sin(2\beta)}{2} > 0, \qquad b_k := \frac{3}{p_k - 2}, \tag{8.12}$$

and c is an absolute constant depending on L and $\int_{S_k}(|Du|^{p(x)}+1)\,dx$, and hence on $\int_{B_1}|Du|^{p(x)}\,dx$. Note that we used the fact that $u \in W^{1,p_k}(S_k)$ since $p(x) \ge p_k$ whenever $x \in S_k$. For comments on the validity of the previous inequality, and, in

particular, the determination of the constant b_k, see Remark 8.5. Then for k large so that $(\pi+1)\, r_k < 1$, by (8.9) and (8.10) we have

$$\begin{aligned} |u(y_k) - u(y_{k+1})| &\leq c\, b_k\, |y_k - y_{k+1}|^{1-2/p_k} \\ &\leq c\, b_k\, (2^{-k})^{(c_\beta/3)\,(\log(L2^{k+2}))^{-t}} \\ &\leq c\, b_k\, e^{\widehat{c}\,(\log(2^{k+2}))^{-t}\,\log(2^{-k})} \\ &\leq c\,(\log(2^{k+2}))^t\, e^{-\widehat{c}\,(\log(2^{k+2}))^{1-t}}, \end{aligned}$$

where c and $\widehat{c}$ are positive constants depending on β, L, and $\int_{B_1}(|Du|^{p(x)}+1)\,dx$. Therefore,

$$\sum_{k=1}^{+\infty} |u(y_k) - u(y_{k+1})| \leq c \sum_{k=1}^{+\infty} (k+2)^t e^{-\widehat{c}\,(\log 2)^{1-t}\,(k+2)^{1-t}} < +\infty,$$

the last series being convergent by (8.4). Observe that the constant c depends on L and $\int_{B_1}(|Du|^{p(x)}+1)\,dx$ and $\widehat{c}$ depends on β, L; moreover $\widehat{c} \to 0$ as $\beta \to 0$ or when $L \to +\infty$ whereas, by the definition of b_k, it follows that $c \to +\infty$ as $\beta \to 0$. Therefore we have that the sequence $u(y_k)$ converges to a certain limit value $l < +\infty$.

Step 2*: Comparing dyadic type sequences.* Now take $\{y_k^1\}$ and $\{y_k^2\}$, two sequences as in (8.8) satisfying (8.9) with different constants L_1, L_2, and define $L = \max\{L_1, L_2\}$. Arguing as in the previous step, there exist l_1, l_2 such that $u(y_k^1) \to l_1$ and $u(y_k^2) \to l_2$. As in (8.10) we also deduce that

$$|y_k^1 - y_k^2| \leq c_3(L)/2^k.$$

With the same notation as in (8.12) (with everything adapted to the new value of the constant L), we find

$$\begin{aligned} |u(y_k^1) - u(y_k^2)| &\leq c\frac{|y_k^1 - y_k^2|^{1-2/p_k}}{p_k - 2} \\ &\leq c\,(k+2)^t e^{(-\widehat{c})\,(\log 2)^{1-t}\,(k+2)^{1-t}} \to 0, \end{aligned}$$

where $\widehat{c} \equiv \widehat{c}(\beta, L) > 0$ and c depends both on L and $\int_{B_1}(|Du|^{p(x)}+1)\,dx$, as in the previous step. Therefore we infer $l_1 = l_2$.

Step 3*: Conclusion.* It suffices to show that if $\{x_k\} \subset \overline{C}_r^\beta \setminus \{0_{\mathbb{R}^2}\}$ converges to $0_{\mathbb{R}^2}$, then $u(x_k) \to l$, where l is defined as the limit of any sequence of Step 1. Note that by Step 2 we have that l does not depend on such a choice. Therefore we pick $\lambda_1 := l$ in (8.6). In turn, it suffices to show that from $\{x_k\}$ it is possible to select a subsequence $\{z_k\}$ such that $u(z_k) \to l$. To this aim we let $x_k := \widetilde{\rho}_k(\cos\widetilde{\varphi}_k, \sin\widetilde{\varphi}_k)$ and we pass to a subsequence $z_k := \rho_k(\cos\varphi_k, \sin\varphi_k)$ such that $\rho_{k+1} \leq 4^{-1}\rho_k$. Next we consider the new sequence $\{y_k\}$, built as follows. First we define $\{\widetilde{r}_k\}$ as the decreasing rearrangement of the set $\{\rho_k \mid k \in \mathbb{N}\} \cup \{2^{-k} \mid k \in \mathbb{N}\} \equiv A \cup B$. Then, from this sequence we build yet another sequence by dropping certain terms: we delete $\widetilde{r}_h$ if and only if $\widetilde{r}_h \in B \setminus A$ and moreover $\widetilde{r}_{h+1} \in A$. (Roughly speaking, after rearranging the pieces of the original sequence A with those from the dyadic type sequence B, we delete all the terms from $B \setminus A$ which come immediately before a term of the sequence A. Observe that since we have chosen $\{\rho_k\}$ such that $\rho_{k+1} \leq \rho_k/4$, then between any two terms of the type 2^{-k} and 2^{-k-1} it falls at most one term of $A \setminus B$; moreover, in the new sequence all the terms of the original A do appear and, in the case in which they go to zero faster than 2^{-k}, they are interpolated by the dyadic numbers.)

Relabelling, we finally get $\{r_k\}$; then we set $y_k := r_k(\cos\theta_k, \sin\theta_k)$, where $\theta_k = \varphi_j$ if $r_k \in A$ and $r_k = \rho_j$ for some $j \in \mathbb{N}$, while $\theta_k = \pi/4$ otherwise. The sequence $\{y_k\}$ is now of the type in (8.8), for a suitable constant L. Therefore by Steps 1 and 2 it follows that $u(y_k) \to l$ and so does $\{u(z_k)\}$, being a subsequence of $\{u(y_k)\}$. The proof is now complete. □

Remark 8.5. Here we briefly justify the validity of the inequality (8.11). It is well known that if we let

$$R := (0,1) \times (\beta, \pi/2 - \beta),$$

then for any function $v \in W^{1,s}(R)$, with $s > 2$, Morrey's imbedding inequality takes the form

$$|v(x) - v(y)| \le \frac{c}{1 - 2/s} |x - y|^{1-2/s} \left(\int_R |Dv(z)|^s \, dz \right)^{1/s}, \tag{8.13}$$

where c is an absolute constant. This can be inferred from [5, p. 110]; similar inequalities are valid for general parallelepipeds in higher dimensions. Then we infer (8.11) from the previous inequality, letting of course $s := p_k$, via a simple change of variable argument and the use of polar coordinates. The details follow. Using the nonsingular map

$$\phi_k : (\rho, \theta) \in R \to g_k(\rho)(\cos\theta, \sin\theta) \in S_k,$$

$$g_k(\rho) := (L2^{-k} - L^{-1}2^{-(k+1)})\rho + L^{-1}2^{-(k+1)},$$

it turns out that

$$\Phi_k \equiv \phi_k^{-1} : (x_1, x_2) \in S_k \to (f_k(|x|), \arctan(x_2/x_1)) \in R,$$

where

$$f_k(\rho) := (L2^{-k} - L^{-1}2^{-(k+1)})^{-1}(t - L^{-1}2^{-(k+1)}),$$

and the following relations hold:

$$||D\phi_k||_\infty \le c\, 2^{-k}, \qquad ||D\Phi_k||_\infty \le c\, 2^k,$$

$$\det D\phi_k = g_k(\rho)(L2^{-k} - L^{-1}2^{-(k+1)}) > 0, \qquad |\det D\phi_k|^{-1} \le c4^k,$$

where $c \equiv c(L)$ denotes an absolute constant independent of $k \in \mathbb{N}$. Finally, if $u \in W^{1,s}(S_k)$, then $v := u \circ \phi_k \in W^{1,s}(R)$; therefore, using (8.13), the change of variable formula, and the previous relations, we get, with $x, y \in S_k$ and $\phi_k(\tilde{x}) = x$ and $\phi_k(\tilde{y}) = y$,

$$\begin{aligned}
|u(x) - u(y)| &= |v(\tilde{x}) - v(\tilde{y})| \\
&\le \frac{c\,s}{s-2} |\tilde{x} - \tilde{y}|^{1-2/s} ||Dv||_{L^s(R)} \\
&\le \frac{c\,s}{s-2} ||D\Phi_k||_\infty^{1-2/s} ||D\phi_k||_\infty |||\det D\phi_k|^{-1}||_\infty^{1/s} |x - y|^{1-2/s} \\
&\qquad \times \left(\int_R |Du(\phi_k(\rho,\theta))|^s \det D\phi_k \, d\rho \, d\theta \right)^{1/s} \\
&\le \frac{c\,s}{s-2} |x - y|^{1-2/s} \left(\int_{S_k} |Du(x)|^s \, dx \right)^{1/s},
\end{aligned}$$

and (8.11) follows by taking $s := p_k$, observing that $\max p(x) \le 3$; also, from the previous estimates, the constant c clearly depends on L and blows up when so L does. Regardless, c is independent of $k \in \mathbb{N}$.

Proof of Theorem 8.3. As for the proof of Theorem 7.6 we shall restrict ourselves, without loss of generality, to the case $f(x,z) := |z|^{p(x)}$. Arguing as for Theorem 7.4, it then suffices to prove the following proposition.

PROPOSITION 8.6. *Let $\{u_j\} \subset C^1(B_r)$ be such that $u_j \to u$ in $L^1(B_r)$ and a.e. in B_r and*

$$\sup_j \int_{C_r^\beta} |Du_j|^{p(x)}\,dx + \sup_j \int_{-C_r^\beta} |Du_j|^{p(x)}\,dx < +\infty.$$

Then, possibly passing to subsequences, $u_j \to u$ uniformly on the closure of $C_r^\beta \cup -C_r^\beta$. Therefore, $\lambda_1 = \lambda_2$ in (8.6).

Proof. Due to a.e. convergence $u_j \to u$, by the Ascoli–Arzelà theorem it suffices to show that $\{u_j\}$ is equi-uniformly continuous on the closure of C_r^β and of $-C_r^\beta$. We make use of the following lemma.

LEMMA 8.7. *Let $v \in C^1(B_r)$ be such that $\int_{C_r^\beta} |Dv|^{p(x)}\,dx < +\infty$; there exist a nondecreasing nonnegative function $g : \mathbb{R}^+ \to \mathbb{R}^+$, depending on $p(x)$ and β, with $g(R) \to 0^+$ as $R \to 0^+$ and constants $\widehat{c}_1$, depending only on $\int_{C_r^\beta} |Dv|^{p(x)}\,dx$, and β, $\widehat{c}_2$ depending only on β, both independent of the function v, such that*

$$|v(x) - v(y)| \le \widehat{c}_1 \min\left\{B(R)d^{\widehat{c_2}(\log(2/R))^{-t}},\, g(S)\right\} \tag{8.14}$$

for every $x, y \in \overline{C}_r^\beta \setminus \{0\}$. Here $d := |x-y| \le 1$, $S := \max\{|x|,|y|\}$, $R := \min\{|x|,|y|\}$, and $B(R)$ is a function depending only on R and such that it is bounded on every interval of the type $[R_0, 1]$, $R_0 > 0$. The same result holds replacing $\overline{C}_r^\beta$ by $-\overline{C}_r^\beta$.

Proof. We treat only the case of $\overline{C}_r^\beta$, the proof for $-\overline{C}_r^\beta$ being similar. By applying Morrey's theorem to the set $C_1^\beta \setminus C_{R/2}^\beta$ we infer

$$|v(x) - v(y)| \le c\,B(R)|x-y|^{c_\beta\,(\log(4/R))^{-t}},$$

where $c_\beta := \sin(2\beta)/2 > 0$, $B(R)$ depends only on R and the constant c depends on $\int_{C_r^\beta} |Du|^{p(x)}\,dx$; observe that $B(R)$ is given by Morrey's imbedding inequality, and it turns out that $B(R) \to +\infty$ when $R \to 0$; this gives the first estimate for (8.14). In order to get the second estimate we argue as follows: if $k \in \mathbb{N}$ is such that $2^{-k} \le |x| < 2^{-k+1}$, and $x_i := 2^{-i}(\cos(\pi/4), \sin(\pi/4))$, arguing as in Theorem 8.2 (compare Remark 8.5) and setting $p(\rho) := 2 + c_\beta\,(\log(2/\rho))^{-t}$, we also infer

$$\begin{aligned}
|v(x) - v(0_{\mathbb{R}^2})| &\le |v(x) - v(x_k)| + \sum_{i=k}^{+\infty} |v(x_i) - v(x_{i+1})| \\
&\le c\,\frac{|x - x_k|^{1-2/p(2^{-k})}}{p(2^{-k}) - 2} + c\sum_{i=k}^{+\infty} \frac{|x_i - x_{i+1}|^{1-2/p(2^{-(i+1)})}}{p(2^{-(i+1)}) - 2} \\
&\le c\sum_{i=k}^{+\infty} \log(2^{i+1})^t (2^{-(i+1)})^{(c_\beta/3)\log(2^{i+1})^{-t}} \\
&\le c\sum_{i=k}^{+\infty} (i+1)^t e^{-\widehat{c}\,(\log 2)^{1-t}\,(i+1)^{1-t}} \\
&=: cA_k.
\end{aligned}$$

Observe that, as for Theorem 8.2, the constant $c > 0$ depends only on $\int_{C_r^\beta} |Dv|^{p(x)}\,dx$ and β, while $\widehat{c} > 0$ depends only on β, and moreover $c \to +\infty$ when $\beta \to 0$ while $\widehat{c} \to 0$ when $\beta \to 0$; this follows by Remark 8.5 replacing S_k by $\widetilde{S}_k := C^\beta \cap B^{2^{-k+1}}_{2^{-k}}$ (thereby taking $L = 1$). As before, it has been used that $u \in W^{1,p(2^{-k})}(\widetilde{S}_k)$.

In the same way, if $2^{-h} \le |y| < 2^{-h+1}$ for some $h \in \mathbb{N}$, then

$$|v(y) - v(0_{\mathbb{R}^2})| \le A_h$$

and by the triangle inequality

$$|v(x) - v(y)| \le 2\max\{A_k, A_h\} = 2A_{\min\{k,h\}}.$$

Then, since $A_k \to 0$ as $k \to +\infty$ (see (8.4)), we conclude by setting

$$g(S) := \begin{cases} A_k & \text{if} \quad 2^{-k} \le S < 2^{-k+1}, \\ A_1 & \text{if} \quad S \ge 1/2; \end{cases}$$

clearly, $g(S) \to 0$ if and only if $S \to 0$. □

End of Proofs of Proposition 8.6 *and Theorem* 8.3. Let $\{u_j\} \subset C^1(B_r)$ be the sequence as in the statement; as explained at the beginning of the proof of Proposition 8.6, it suffices to prove that $\{u_j\}$ is equi-uniformly continuous on the closure of C_r^β, and in turn this is equivalent to proving the equi-uniform continuity on $\overline{C}_r^\beta \setminus \{0\}$. We have to prove that for any $\varepsilon > 0$ there exists $\delta \equiv \delta(\epsilon) > 0$ such that whenever $x, y \in \overline{C}_r^\beta \setminus \{0\}$ satisfy $|x - y| \le \delta$, then $|u_j(x) - u_j(y)| \le \epsilon$ for every $j \in \mathbb{N}$. We argue by contradiction; if it were not so, there would exist $\epsilon_0 > 0$ such that for any positive integer h there exist $j(h) \in \mathbb{N}$, $x_h, y_h \in \overline{C}_r^\beta \setminus \{0\}$ and a function $v_h := u_{j(h)}$ such that

$$d_h := |x_h - y_h| \le 1/h \tag{8.15}$$

but

$$|v_h(x_h) - v_h(y_h)| > \epsilon_0. \tag{8.16}$$

By the estimate (8.14), if we set $S_h := \max\{|x_h|, |y_h|\} > 0$, then (8.16) implies that $g(S_h) \ge \epsilon_0/c_1 > 0$ (where c_1 is independent of $h \in \mathbb{N}$, as observed in Lemma 8.7), and so there exists $S_0 > 0$ such that $S_h \ge S_0$ for every $h \in \mathbb{N}$. In turn, if we let $R_h := \min\{|x_h|, |y_h|\} > 0$, by (8.15) we get that $R_h \ge R_0 := S_0/2 > 0$ for every index $h > 2/S_0$. Hence, by (8.14) we get

$$B(R_h) d_h^{\widehat{c}_2 (\log(2/R_h))^{-t}} < \widetilde{B} d_h^{\widehat{c}_2 (\log(2/R_0))^{-t}} \to 0, \qquad \widetilde{B} = \max_{[R_0,1]} B(R),$$

and applying again (8.14) and (8.16) yields

$$0 < \epsilon_0 \le \limsup_{h \to +\infty} |v_h(x_h) - v_h(y_h)| \le \lim_{h \to +\infty} \widehat{c}_1 B(R_h) d_h^{\widehat{c}_2 (\log(2/R_h))^{-t}} = 0,$$

which is impossible. Therefore, $\{u_j\}$ is equi-uniformly continuous and the proofs are complete. □

Acknowledgments. G. M. acknowledges the hospitality of the Departments of Mathematics of Albert-Ludwig University (Freiburg) and Charles University (Prague), in May 2001 and May 2002, respectively. Last but not least, the authors acknowledge the invaluable work of the referees, which contributed substantially to improving the presentation of the paper.

REFERENCES

[1] E. Acerbi, G. Bouchitté, and I. Fonseca, *Relaxation of convex functionals: The gap problem,* Ann. Inst. H. Poincaré Anal. Non Linéaire, 20 (2003), pp. 359–390.

[2] E. Acerbi and N. Fusco, *Semicontinuity problems in the calculus of variations,* Arch. Ration. Mech. Anal., 86 (1984), pp. 125–145.

[3] E. Acerbi and G. Mingione, *Regularity results for a class of functionals with non-standard growth,* Arch. Ration. Mech. Anal., 156 (2001), pp. 121–140.

[4] E. Acerbi and G. Mingione, *Regularity results for stationary electro-rheological fluids,* Arch. Ration. Mech. Anal., 164 (2002), pp. 213–259.

[5] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[6] J. Ball, *Convexity conditions and existence theorems in nonlinear elasticity,* Arch. Rational Mech. Anal., 63 (1976/77), pp. 337–403.

[7] J. Ball and F. Murat, *$W^{1,p}$-quasiconvexity and variational problems for multiple integrals,* J. Funct. Anal., 58 (1984), pp. 225–253.

[8] G. Bouchitté, I. Fonseca, and J. Malý, *The effective bulk energy of the relaxed energy of multiple integrals below the growth exponent,* Proc. Roy. Soc. Edinburgh Sect. A, 128 (1998), pp. 463–479.

[9] A. Braides and A. Defranceschi, *Homogenization of multiple integrals*, Oxford Lecture Ser. Math. Appl. 12, Oxford University Press, Oxford, UK, 1998.

[10] G. Buttazzo, *Semicontinuity, Relaxation and Integral Representation in the Calculus of Variations,* Pitman Research Notes in Mathematical Series 207, Longman, Harlow, UK, 1989.

[11] G. Buttazzo and G. Dal Maso, *Integral representation and relaxation of local functionals,* Nonlinear Anal., 9 (1985), pp. 515–532.

[12] G. Buttazzo and V. Mizel, *Interpretation of the Lavrentiev phenomenon by relaxation,* J. Funct. Anal., 110 (1992), pp. 434–460.

[13] A. Coscia and D. Mucci, *Integral representation and Γ-convergence of variational integrals with $p(x)$-growth,* ESAIM: Control, Optim. Calc. Var., 7 (2002), pp. 495–519.

[14] B. Dacorogna, *Direct Methods in the Calculus of Variations*, Appl. Math. Sci. 78, Springer-Verlag, Berlin, 1989.

[15] B. Dacorogna, *Quasiconvexity and relaxation of nonconvex problems in the calculus of variations,* J. Funct. Anal., 46 (1982), pp. 102–118.

[16] G. Dal Maso, *An Introduction to Γ-Convergence*, Progr. Nonlinear Differential Equations Appl. 8, Birkäuser Boston, Boston, 1993.

[17] E. De Giorgi, *Teoremi di semicontinuitá nel calcolo delle variazioni*, INDAM, Rome, Italy, 1968-69.

[18] E. De Giorgi and G. Letta, *Une notion générale de convergence faible pour des fonctions croissantes d'ensemble,* Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 4 (1977), pp. 61–99.

[19] L. Diening, *Maximal function of generalized Lebesgue spaces $L^{p(x)}$*, Math. Inequal. Appl. 7 (2004).

[20] D. E. Edmunds and J. Rákosnik, *Sobolev embeddings with variable exponent,* Studia Math., 143 (2000), pp. 267–293.

[21] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North Holland, Amsterdam, 1976.

[22] L. Esposito, F. Leonetti, and G. Mingione, *Sharp regularity for functionals with (p,q)-growth,* J. Differential Equations, 204 (2004), pp. 5–55.

[23] H. Federer, *Geometric Measure Theory,* Grundleheren Math. Wiss. 153, Springer-Verlag, New York, 1969.

[24] I. Fonseca and J. Malý, *Relaxation of multiple integrals below the growth exponent,* Ann. Inst. H. Poincaré Anal. Non Linéare, 14 (1997), pp. 309–338.

[25] I. Fonseca and P. Marcellini, *Relaxation of multiple integrals in subcritical Sobolev spaces,* J. Geom. Anal., 7 (1997), pp. 57–81.

[26] M. Foss, *Examples of the Lavrentiev phenomenon with continuous Sobolev exponent dependence,* J. Convex Anal., 10 (2003), pp. 445–464.

[27] M. Fuchs and G. A. Seregin, *Variational Methods for Problems from Plasticity Theory and for Generalized Newtonian Fluids*, Lecture Notes in Mathematics 1749, Springer-Verlag, Berlin, 2000.
[28] N. Fusco, *On the convergence of integral functionals depending on vector-valued functions,* Ricerche Mat., 32 (1983), pp. 321–329.
[29] N. Fusco and C. Sbordone, *Higher integrability of the gradient of minimizers of functionals with nonstandard growth conditions,* Comm. Pure Appl. Math., 43 (1990), pp. 673–683.
[30] W. Gangbo, *On the weak lower semicontinuity of energies with polyconvex integrands,* J. Math. Pures Appl. (9), 73 (1994), pp. 455–469.
[31] P. Harjuletho and P. Hästo, *An overview of variable exponent Lebesgue and Sobolev spaces,* in Proceedings of the Workshop on Future Trends in Geometric Function Theory, Jyväskylä, Finland, 2003.
[32] A. D. Ioffe, *On lower semicontinuity of integral functionals* I, SIAM J. Control Optim., 15 (1977), pp. 521–538.
[33] J. Kristensen, *Lower semicontinuity in Sobolev spaces below the growth exponent of the integrand*, Proc. Roy. Soc. Edinburgh Sect. A, 127 (1997), pp. 797–817.
[34] J. Kristensen, *Lower semicontinuity of quasi-convex integrals in BV,* Calc. Var. Partial Differential Equations, 7 (1998), pp. 249–261.
[35] J. Kristensen, *Lower semicontinuity in spaces of weakly differentiable functions,* Math. Ann., 313 (1999), pp. 653–710.
[36] J. Kristensen, *A necessary and sufficient condition for lower semicontinuity*, Ann. Mat. Pura Appl, to appear.
[37] J. Malý, *Weak lower semicontinuity of polyconvex integrals,* Proc. Roy. Soc. Edinburgh Sect. A, 123 (1993), pp. 681–691.
[38] J. Malý, *Lower semicontinuity of quasiconvex integrals,* Manuscripta Math., 85 (1994), pp. 419–428.
[39] P. Marcellini, *Approximation of quasiconvex functions, and lower semicontinuity of multiple integrals,* Manuscripta Math., 51 (1985), pp. 1–28.
[40] P. Marcellini, *On the definition and the lower semicontinuity of certain quasiconvex integrals,* Ann. Inst. H. Poincaré Anal. Non Linéare, 3 (1986), pp. 391–409.
[41] C. B. Morrey, *Quasi-convexity and semicontinuity of multiple integrals,* Pacific J. Math., 2 (1952), pp. 25–53.
[42] D. Mucci, *Relaxation of variational functionals with piecewise constant growth conditions,* J. Convex Anal., 10 (2003), pp. 295–324.
[43] J. Musielak, *Orlicz Spaces and Modular Spaces*, Springer-Verlag, Berlin, 1983.
[44] M. Ružička, *Electrorheological Fluids: Modeling and Mathematical Theory,* Lecture Notes in Math. 1748, Springer-Verlag, Berlin, 2000.
[45] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.
[46] V. V. Zhikov, *On Lavrentiev's phenomenon,* Russian J. Math. Phys., 3 (1995), pp. 249–269.

SIAM J. MATH. ANAL.
Vol. 36, No. 5, pp. 1580–1628

NETWORK APPROXIMATION FOR EFFECTIVE VISCOSITY OF CONCENTRATED SUSPENSIONS WITH COMPLEX GEOMETRY*

LEONID BERLYAND†, LILIANA BORCEA‡, AND ALEXANDER PANCHENKO§

Abstract. We study suspensions of rigid particles (inclusions) in a viscous incompressible fluid. The particles are close to touching one another, so that the suspension is near the packing limit, and the flow at small Reynolds number is modeled by the Stokes equations. The objective is to determine the dependence of the effective viscosity $\langle\mu\rangle$ on the geometric properties of the particle array. We study spatially irregular arrays, for which the volume fraction alone is not sufficient to estimate the effective viscosity. We use the notion of the interparticle distance parameter δ, based on the Voronoi tessellation, and we obtain a discrete network approximation of $\langle\mu\rangle$, as $\delta \to 0$. The asymptotic formulas for $\langle\mu\rangle$, derived in dimensions two and three, take into account translational and rotational motions of the particles. The leading term in the asymptotics is rigorously justified in two dimensions by constructing matching upper and lower variational bounds on $\langle\mu\rangle$. While the upper bound is obtained by patching together local approximate solutions, the construction of the lower bound cannot be obtained by a similar local analysis because the boundary conditions at fluid-solid interfaces must be resolved for all particles simultaneously. We observe that satisfying these boundary conditions, as well as the incompressibility condition, amounts to solving a certain algebraic system. The matrix of this system is determined by the total number of particles and their coordination numbers (number of neighbors of each particle). We show that the solvability of this system is determined by the properties of the network graph (which is uniquely defined by the array of particles) as well as by the conditions imposed at the external boundary.

Key words. effective viscosity, discrete network, variational bounds, concentrated suspension

AMS subject classifications. 74Q, 35Q72, 74F10, 76T20

DOI. 10.1137/S0036141003424708

1. Introduction.

In this paper, we obtain and justify approximate formulas for the effective viscosity $\langle\mu\rangle$ of a highly concentrated suspension of solid particles in a viscous incompressible fluid. We study generic, nonperiodic spatial distributions of particles and focus on a particular type of highly concentrated suspension, which can be approximately modeled on the macroscale by a single phase fluid, called the effective fluid. The effective viscosity is determined from the equality of the viscous dissipation rates in the suspension and the effective fluid. This is a classical approach that goes back to Einstein [14], who approximated the effective viscosity in the limit of an infinitely small particle concentration (the so-called dilute limit). Further results for dilute suspensions can be found in [3] and the references therein.

While in the dilute limit the interactions between the particles are negligible, the case of finite (nonsmall) concentrations is much harder to analyze because these interactions must be taken into account. In [25], an asymptotic expansion of the effective viscosity was constructed assuming a periodic distribution of particles. In

*Received by the editors March 23, 2003; accepted for publication (in revised form) February 13, 2004; published electronically April 29, 2005.

http://www.siam.org/journals/sima/36-5/42470.html

†Department of Mathematics and Materials Research Institute, McAllister Building, Penn State University, University Park, PA 16802 (berlyand@math.psu.edu). Supported in part by NSF grant DMS-0204637.

‡Computational and Applied Mathematics, MS 134, Rice University, 6100 Main Street, Houston, TX 77005-1892 (borcea@caam.rice.edu). Supported in part by NSF grant DMS-9971209 and by ONR grant N00014-02-1-0088.

§Department of Mathematics, Penn State University. Current address: Department of Mathematics, Washington State University, Pullman, WA 99163 (panchenko@math.wsu.edu). Supported in part by ONR grant N00014-001-0853.

[25], the formal two-scale homogenization was carried out under the assumption that the number of particles tends to infinity while their total volume remains constant. In this case, the distances between the particles are of the order of their sizes, which is the key feature of the so-called finite (moderate) concentration regime.

By contrast, our interest lies in the high concentration regime, where the particles are close to touching one another, so the typical interparticle distances are much smaller than their sizes. In this case, the hydrodynamic interactions lead to the blow-up of the dissipation rate in the thin gaps between the closely spaced particles. Note that the effective fluid can be either Newtonian [15] or non-Newtonian [1, 2, 4, 24]. We consider only the former case, following [15, 16, 27] (see also [19] for a review of physical data). Also, we consider only noncolloidal suspensions, which means that hydrodynamic interactions are much stronger than Brownian interactions, so the latter can be neglected. For effective rheology of colloidal suspensions, one may consult [12].

For periodic arrays of particles [15, 16, 27], the estimation of the effective viscosity reduces to solving the flow problem locally, in a thin gap between two neighboring particles. In [15], this is done by a formal asymptotic method, similar to the well-known lubrication approximation, which takes into account only the translational motions of particles along the lines of their centers. The contributions of rotations and shear-type translations are neglected in [15]. In [16], a more general formula for the effective viscosity is obtained, which combines the results in [15] and [14] for dilute and high concentration regimes, respectively. In [27], the definition of the effective viscosity involves the traction exerted by the fluid on a single sphere. This traction satisfies an integral equation derived and solved (for a cubic periodic lattice) in [27]. Note in particular that the periodicity assumption in [27] reduces the boundary conditions on the surface of the particles to just a rigid body rotation (no translations).

In this paper, we consider generic, nonperiodic arrays, where different particles have different translational and rotational body motions. Since the rigid motions of the particles are not known a priori, the effective viscosity cannot be obtained simply by solving a local problem in the gap between adjacent particles. The motion of one particle influences the motion of all the particles in the array and, to find the effective viscosity, we must solve the global problem. A key ingredient in our method of solution is the so-called discrete network approximation.

Discrete network models have been used in the engineering and physics literature [21, 17, 28, 29], although the relation between the continuum problem and the discrete network has not been established. The first rigorous mathematical characterization of high-contrast media, in terms of discrete networks, was obtained for electromagnetic transport problems in [8, 9, 10, 11], where the electrical conductivity (and permittivity) were modeled as exponentials of the form $e^{S(x)/\epsilon}$. This continuum high-contrast model is due to Kozlov [22], where S is a smooth, Morse function and where $\epsilon \ll 1$, such that small variations of S are highly amplified by the exponential, thus giving the high contrast. Kozlov's model is especially useful in the context of imaging [10], where the medium is not known and it is approximated by a generic, high-contrast continuum. The high-contrast continuum model leads to an explicit characterization of two-dimensional flow of DC (AC) electric current in the material, in terms of a network of resistors (and capacitors), which is uniquely defined by the distribution of critical points of S. Explicitly, in the DC case, the nodes of the network are the local maxima of the electrical conductivity function (i.e., of S) and the branches of the network connect adjacent nodes through the saddle points of S. The resistor associated with each branch is determined by the conductivity and by the curvatures of S at

the saddle point, respectively. The boundary currents and voltages of the asymptotic network are also uniquely defined by S and by the boundary conditions specified for the continuum problem, so the asymptotic results in [9, 10, 11] give more than the homogenized electrical properties of the high contrast continuum. They give that the Neumann-to-Dirichlet map of the continuum problem is asymptotically equivalent to the discrete Neumann-to-Dirichlet map of the asymptotic network, in the limit $\epsilon \to 0$ [9, 11]. All the results in [8, 9, 10, 11] apply to the two-dimensional case for all smooth functions S with isolated, nondegenerate critical points. Extensions to three dimensions are straightforward for a special class of functions S, but for a general S, the network approximation may not apply.

In [5], another network approximation has been developed for a scalar, DC conductivity problem which models dispersive high contrast composites. In this case, $S(x)$ is the characteristic function of the particles, the high-contrast parameter is $\epsilon = 0$ (perfectly conducting particles), and the asymptotic analysis is carried out in the limit of the interparticle distance parameter δ tending to zero. The particle radii are not treated as small parameters, and the number of the particles is sufficiently large but bounded from above by $N_{\max}$, where $N_{\max}$ is the maximal close packing number. In [5], the connectivity patterns and the interparticle distance parameter for irregular spatial arrays of particles are rigorously defined using Voronoi tessellation. It is demonstrated that the network approximation is an efficient numerical tool, capable of capturing various percolation effects as well as effects due to the polydispersity of particles. This approach also allows for analytical error estimates, subsequently obtained in [6].

In [6, 8, 9, 10, 11, 5], the network approximation was rigorously justified by employing variational duality. The key point is the construction of trial functions, the electric potential and current density for the direct and dual variational problems, respectively. The choice of trial functions depends on both the mathematical and the physical features of the problem. For example, the construction of trial functions in [5] is essentially different from those in [8, 9, 10, 11], and it requires the development of new mathematical tools. While the upper bound can be obtained by patching together the appropriate test functions based on the local analysis of [20], such a straightforward approach does not work for the lower bound. The difficulty in obtaining the latter lies in the construction of trial functions for the dual problem, when the boundary conditions on the surfaces of the particles cannot be satisfied independently for each particle, and one must deal with all inclusions simultaneously. The dual (lower) bound was obtained in [5] by constructing an approximate, divergence-free trial electric current density in the gap between adjacent particles and extending it to zero elsewhere in the domain. Then, the network equations are used to choose the unknown parameters in the dual trial field, so that the boundary conditions on the surface of the particles are satisfied exactly. Note, however, that this construction is specialized to the scalar, electrical conductivity problem, and does not admit a generalization to vectorial problems.

In this work, we study the vectorial problem described by Stokes's flow in a closely packed suspension with rigid particles. Since the array of particles is irregular, our construction uses the interparticle distance parameter introduced in [5], based on the Voronoi tessellation. Due to the high concentration of particles of finite size, in a fixed volume, the particles are close to touching. Thus, we assume that distances δ^{ij} between adjacent particles $D^{(i)}$ and $D^{(j)}$ become infinitesimally small, but positive. More precisely, we say that $c\delta \leq \delta^{ij} \leq \delta$ for all pairs $D^{(i)}$, $D^{(j)}$ of neighboring particles, where $0 < c < 1$ is fixed and where δ is the small parameter of the problem.

We are interested in the asymptotics of the effective viscosity as $\delta \to 0$, while the particle radii a_i are kept fixed and the number of particles N approaches $N_{\max}$, from below.

The goal of this paper is twofold. The first objective is to obtain a method for estimating the effective viscosity which captures explicitly the effects of the complex geometry (the irregular distribution of the location and size of the particles). This is done in both two and three dimensions, and our derivation is based on the generalization of the lubrication approximation technique. We take into account all possible translations and rotations of the rigid particles in the suspension, which we quantify by constant vectors $\mathbf{T}^{(p)}$ and $\boldsymbol{\omega}^{(p)}$, respectively, for $1 \le p \le N$. Using the linearity of the problem, we approximate first the velocity, pressure, and stress in the gaps (necks) between the particles for translational and rotational motions, separately, and then we superpose the results. The lubrication analysis is local for each gap, and by summing the contribution of all the gaps, we obtain the discrete approximation of $\langle \mu \rangle$, parameterized in terms of the rigid body translational and rotational velocities $\mathbf{T}^{(p)}$ and $\boldsymbol{\omega}^{(p)}$, respectively, for $1 \le p \le N$. These rigid body motions are not arbitrary, but they are calculated by solving a system of linear equations, which corresponds to the conditions of mechanical equilibrium for all particles in the suspension.

For the reader interested mainly in numerical estimation of the effective viscosity, we describe our approach in Remark 3.1. (See also the forthcoming paper [7], where the effective viscosity was computed for several boundary conditions and various particle arrays by adapting the approach developed in this paper.)

The second objective of the paper is to provide a rigorous mathematical justification of the asymptotic approximation of the effective viscosity. The rigorous justification of the leading order term in the asymptotic approximation is done here in two dimensions. The most subtle part of this justification is the construction of the dual trial function for the lower bound on the effective viscosity. None of the techniques developed previously in [5, 6, 8, 9, 10, 11] for constructing trial functions for the dual problem work here. There are two main difficulties in the construction of the bounds on the effective viscosity. The first difficulty is that the trial functions must be divergence free in the fluid domain. The second difficulty is raised by the boundary conditions on fluid-solid interfaces. While these issues can be handled in the upper bound construction with an approach inspired by the work in [5, 6, 8, 9, 10, 11], the dual problem is significantly more challenging because the trial fields are tensors. In the dual problem, neither of the above two difficulties can be resolved by doing local analysis, that is, by choosing approximate solutions in each gap followed by patching these solution together. First, we must consider the global problem to ensure that the boundary conditions are satisfied for all inclusions at once. Second, we show that the divergence-free requirement on the stress trial fields is also global, analogous to the interface conditions. Then, we observe that the solvability of a certain algebraic system is sufficient to ensure that these two global requirements are satisfied. The size of the matrix of the linear system is determined by N, the total number of particles, and by their coordination number (number of neighbors). The solvability of the system, in turn, is determined by the connectivity and the coordination numbers of the network graph corresponding to the particle array, as well as by the conditions at the external boundary. We present geometric conditions for the network graph (topology) so that this linear system is solvable. In particular, we point out that these conditions are satisfied by network graphs which model typical close packing configurations.

The paper deals with irregular spatial arrays of particles. In this case, the total

volume fraction of particles (the only parameter in the formulas from [14, 15]) is not sufficient for estimating the effective viscosity. Instead of a formula, we give an algorithm, which essentially reduces computation of the effective viscosity to solving a linear algebraic system for translational and angular velocities of particles. The gain here is that we obtain an accurate yet computationally inexpensive approximation for the effective viscosity, which, unlike the above mentioned formulas, takes into account variable distances between neighboring particles. Note that variability in these distances for a fixed total volume fraction of particles may result in significant changes in the effective properties due to percolation effects (see [6]).

The focus of this paper is on derivation and, particularly, analytical justification of this algorithm, while its implementation will be investigated elsewhere. (See, for example, the forthcoming [7], where both shear and compression boundary conditions for various arrays of particles are investigated.) This paper, however, contains results of immediate practical interest, such as determination of the order of magnitude of the effective viscosity in the interparticle distance parameter δ. An interesting feature of the vectorial problem that distinguishes it from the scalar case considered in [5, 6, 8, 9, 10, 11] is that the order of magnitude of the effective viscosity depends crucially on the geometry of the particle array and on the boundary conditions. For the scalar problem, the order of magnitude is the same for all networks satisfying a natural connectedness assumption [5], which is not the case in our vectorial problem. In section 6.2.6 we give a sufficient condition on the particle array such that the effective viscosity blows up at the rate $\delta^{-3/2}$, in two dimensions, and the leading term in the asymptotics of $\langle\mu\rangle$ is given by the so-called spring network approximation, in which only the translational motions of adjacent particles, along the axis of their centers, are taken into account. In this case, the rotations of the particles do not contribute to the leading term of the asymptotics of $\langle\mu\rangle$. If an array does not satisfy this condition, the blow up rate may be a weaker ($\delta^{-1/2}$), in which case rotational contributions cannot be ignored. A detailed study of this phenomenon is presented in the forthcoming [7], where we also use network approximation to explain the discrepancy, observed in [31], between the effective shear viscosity formulas for periodic arrays and estimates obtained from experimental results and numerical simulations.

In this paper, we give the rigorous justification of the spring network approximation. An important physical problem is to calculate the second order term in $\langle\mu\rangle$, which depends on the rotational motions of particles. The two-term (formal) asymptotics obtained here provide physical insight and the quantitative estimate of the contributions of rotations, as well as the effects of variable size distribution. Rigorous justification of these formulas requires a more careful lower bound construction than we attempt here, and it remains an interesting and challenging open problem.

Our study is motivated by the problem of transport of highly concentrated slurries, which arises in numerous industrial applications ranging from construction engineering to combustion processes [30, 32]. It is often necessary to use slurries with high solid content (highly packed). The transport of such slurries is impeded by the fact that their effective viscosity is very high. Thus the goal is to find an optimal balance between the effective viscosity and the concentration of the solid phase. The first step in achieving this goal is to obtain relatively simple formulas which show how the effective viscosity depends on the control parameters (e.g., geometrical parameters, such as the particle size distributions, particle locations, and shapes). The network approximation we propose here can be used in the prediction of optimal properties of such slurries.

The paper is organized as follows. In section 2, we give the mathematical formu-

lation of the problem. Section 3 deals with the discrete network approximation of the effective viscosity. We also give here, and in section 4, the lubrication approximation of $\langle\mu\rangle$, in two and three dimensions. In section 5, we construct the upper bound on the effective viscosity, which accounts for both translational and rotational motions of the inclusions. In section 6, we give the rigorous justification of the spring network approximation of the effective viscosity. This accounts just for the leading order term in the asymptotics of the effective viscosity of the high contrast, closely packed suspension of particles. Finally, in section 7, we give a summary and conclusions.

2. Formulation of the problem.

2.1. The Stokes flow problem.

Consider a cube

$$\Omega = \left\{ \mathbf{x} = \sum_{j=1}^{n} x_j \, \mathbf{e}_j, \;\; -L \le x_j \le L, \; 1 \le j \le n \right\} \tag{2.1}$$

of volume $|\,\Omega\,| = (2L)^n$, where $n = 2$ or 3 and where $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ is an orthonormal basis. We suppose that Ω is filled with N, nonoverlapping, rigid balls (particles) $D^{(j)}$ of radius a_j, suspended in an incompressible fluid of viscosity μ. We study the Stokes flow of this suspension, where the fluid occupies the perforated, connected domain

$$\Omega_F = \Omega \setminus \bigcup_{j=1}^{N} D^{(j)}. \tag{2.2}$$

We are particularly interested in concentrated suspensions with volume fraction

$$\alpha = 1 - \frac{|\,\Omega_F\,|}{|\,\Omega\,|}, \tag{2.3}$$

close to maximal packing (neighboring particles are close to touching).

Let $\mathbf{u}(\mathbf{x})$ be the velocity field at point $\mathbf{x} \in \Omega_F$ and let $\mathcal{E}(\mathbf{x})$ be the rate of strain tensor

$$\mathcal{E}(\mathbf{x}) = \frac{1}{2}[\nabla\mathbf{u}(\mathbf{x}) + (\nabla\mathbf{u}(\mathbf{x}))^T], \tag{2.4}$$

which satisfies

$$\operatorname{trace}\mathcal{E}(\mathbf{x}) = \operatorname{div}\mathbf{u}(\mathbf{x}) = 0 \tag{2.5}$$

by incompressibility. The stress in the fluid is

$$\mathcal{S}(\mathbf{x}) = -P(\mathbf{x})I + 2\mu\mathcal{E}(\mathbf{x}), \tag{2.6}$$

where μ is the viscosity, P is the hydrostatic pressure, and I denotes the unit tensor. In the rigid balls, $\mathcal{E} = 0$. In the absence of external forces, the velocity field $\mathbf{u}(\mathbf{x})$ in the fluid satisfies Stokes's equation

$$\operatorname{div}\mathcal{S}(\mathbf{x}) = \mu\Delta\mathbf{u}(\mathbf{x}) - \nabla P(\mathbf{x}) = \mathbf{0} \tag{2.7}$$

and the incompressibility constraint (2.5).

Let us denote by $\partial\Omega^+$ and $\partial\Omega^-$ the top and bottom parts of the external boundary $\partial\Omega$, respectively,

$$\partial\Omega^+ = \{\mathbf{x} \in \partial\Omega : x_n = L\} \quad \text{and} \quad \partial\Omega^- = \{\mathbf{x} \in \partial\Omega : x_n = -L\}. \tag{2.8}$$

In this paper, we work with the model boundary conditions prescribed as follows. On $\partial\Omega^+ \cup \partial\Omega^-$, the velocity satisfies

$$\mathbf{u}(\mathbf{x}) = \mathbf{g}(x), \quad \text{where } \mathbf{g}(\mathbf{x}) = \begin{cases} -\dfrac{\mathbf{e_n}}{2L} & \text{on } \partial\Omega^-, \\ \dfrac{\mathbf{e_n}}{2L} & \text{on } \partial\Omega^+, \end{cases} \tag{2.9}$$

and the remaining part of $\partial\Omega$ is traction free,

$$\mathcal{S}(\mathbf{x})\ \mathbf{n}(\mathbf{x}) = \mathbf{0} \quad \text{for } \mathbf{x} \in \partial\Omega \setminus \left\{\partial\Omega^+ \cup \partial\Omega^-\right\}. \tag{2.10}$$

At the surface of each rigid ball $D^{(j)}$, the velocity satisfies

$$\mathbf{u}(\mathbf{x}) = \boldsymbol{\omega}^{(j)} \times (a_j \mathbf{n}^{(j)})(\mathbf{x}) + \mathbf{T}^{(j)} \quad \text{on } \partial D^{(j)}, \quad j = 1, 2, \ldots, N, \tag{2.11}$$

where $\boldsymbol{\omega}^{(j)}$, $\mathbf{T}^{(j)}$ are constant but unknown rotational and translational velocities of $D^{(j)}$ and where $\mathbf{n}^{(j)}(\mathbf{x})$ is the outer normal at $\partial D^{(j)}$. Finally, since each rigid ball is in equilibrium, the total force and torque exerted on $D^{(j)}$ by the fluid must be zero,

$$\int_{\partial D^{(j)}} \mathcal{S}\mathbf{n}^{(j)}\, ds = \mathbf{0} \ \text{ and } \ \int_{\partial D^{(j)}} \mathbf{n}^{(j)} \times \mathcal{S}\mathbf{n}^{(j)}\, ds = \mathbf{0} \text{ for } j = 1, 2, \ldots, N. \tag{2.12}$$

It is known that (2.7) and (2.5), with boundary conditions (2.9), (2.11), and (2.12), have a unique solution $\mathbf{u}(\mathbf{x})$, at least in the weak sense, with components in $H^1(\Omega_F)$.

2.2. The effective viscosity. The rate of viscous dissipation of the energy is given by [23]

$$E = \frac{1}{2}\int_{\Omega_F} (\mathcal{S}(\mathbf{x}), \mathcal{E}(\mathbf{x}))\ d\mathbf{x}, \tag{2.13}$$

where $(\cdot, \cdot)$ denotes the Frobenius tensor inner product

$$(\mathcal{S}(\mathbf{x}), \mathcal{E}(\mathbf{x})) = \sum_{i,j=1}^{n} \mathcal{S}_{ij}(\mathbf{x})\, \mathcal{E}_{ij}(\mathbf{x}). \tag{2.14}$$

Integrating by parts and using (2.5), (2.6), (2.9), (2.10), (2.11), and the identity

$$(\mathcal{S}, \mathcal{E}) = -P \operatorname{trace} \mathcal{E} + 2\mu(\mathcal{E}, \mathcal{E}) = \frac{\mu}{2}(\nabla\mathbf{u} + (\nabla\mathbf{u})^T, \nabla\mathbf{u} + (\nabla\mathbf{u})^T), \tag{2.15}$$

we obtain

$$E = \frac{1}{2}\int_{\partial\Omega^+\cup\partial\Omega^-} \frac{\mathbf{e}_n}{2L} \cdot \mathcal{S}(\mathbf{x})\mathbf{e}_n\, ds - \frac{1}{2}\sum_{j=1}^{N}\int_{\partial D^{(j)}} (\boldsymbol{\omega}^{(j)} \times \mathbf{n}^{(j)}(\mathbf{x}) + \mathbf{T}^{(j)}) \cdot \mathcal{S}(\mathbf{x})\mathbf{n}^{(j)}\, ds. \tag{2.16}$$

Furthermore, due to the balance equations (2.12), the integrals at the surface of the particles vanish and (2.13) can be rewritten as

$$E = \frac{1}{4L}\int_{\partial\Omega^+\cup\partial\Omega^-} \mathbf{e}_n \cdot \mathcal{S}(\mathbf{x})\, \mathbf{e}_n\, ds. \tag{2.17}$$

The effective viscosity $\langle\mu\rangle$ is defined by the equation

$$\frac{\langle\mu\rangle}{\mu} = \frac{E}{E^0} = \frac{\displaystyle\int_{\partial\Omega^+\cup\partial\Omega^-} \mathbf{e}_n \cdot \mathcal{S}(\mathbf{x})\,\mathbf{e}_n\,ds}{\displaystyle\int_{\partial\Omega^+\cup\partial\Omega^-} \mathbf{e}_n \cdot \mathcal{S}^0(\mathbf{x})\,\mathbf{e}_n\,ds}, \tag{2.18}$$

where $\mathcal{S}^0(\mathbf{x})$ is the stress tensor that would occur in Ω, in the absence of all the particles, under the same external boundary conditions (2.9), (2.10), and where E^0 is the corresponding rate of dissipation (see, for example, [15]). An equivalent definition of $\langle\mu\rangle$ can be obtained directly from (2.13) and (2.15) by equating the viscous dissipation rates

$$\langle\mu\rangle \int_\Omega \left(\mathcal{E}^0, \mathcal{E}^0\right) d\mathbf{x} = \mu \int_{\Omega_F} (\mathcal{E}, \mathcal{E})\, d\mathbf{x}. \tag{2.19}$$

We note that the definition of the effective viscosity, via the dissipation rate, is introduced in [3] for dilute suspensions, where the energy of the particulate phase is negligible. However, since the particles are rigid and condition (2.12) holds, the total mechanical energy of the particles is conserved. Thus, definitions (2.18) and (2.19) can be used as well for the suspensions considered in this paper.

2.3. The variational principles. The dissipation rate (2.13) or, equivalently, the effective viscosity (2.18), have a primal and dual variational formulation. The primal variational principle is widely known (see, for example, [13]),

$$E = \min_{\mathbf{u}\in\mathcal{U}} W_{\Omega_F}(\mathbf{u}), \text{ where } W_{\Omega_F}(\mathbf{u}) = \frac{\mu}{4}\sum_{i,j=1}^{n}\int_{\Omega_F}\left(\frac{\partial u_i(\mathbf{x})}{\partial x_j} + \frac{\partial u_j(\mathbf{x})}{\partial x_i}\right)^2 d\mathbf{x}, \tag{2.20}$$

and where the function space $\mathcal{U}$ of admissible velocity fields is

$$\mathcal{U} = \left\{\mathbf{u} = \sum_{j=1}^{n} u_j\,\mathbf{e}_j,\ u_j \in H^1(\Omega_F),\ j = 1\ldots n,\ \operatorname{div}\mathbf{u} = 0,\ (2.9) \text{ and } (2.11) \text{ hold}\right\}. \tag{2.21}$$

Note that the minimizer in (2.20) is the solution of the Stokes flow equation (2.7), where $P(\mathbf{x})$ is the Lagrange multiplier for the incompressibility constraint $\operatorname{div}\mathbf{u}(\mathbf{x})=0$.

The dual variational principle[1] is

$$E = \max_{\mathcal{S}\in\mathcal{F}}\left\{\frac{1}{2L}\int_{\partial\Omega^+\cup\partial\Omega^-}\mathbf{e}_n\cdot\mathcal{S}(\mathbf{x})\,\mathbf{e}_n\,ds - \frac{1}{4\mu}\int_{\Omega_F}\left[(\mathcal{S}(\mathbf{x}),\mathcal{S}(\mathbf{x})) - \frac{(\operatorname{trace}\mathcal{S}(\mathbf{x}))^2}{n}\right]d\mathbf{x}\right\}, \tag{2.22}$$

where we maximize over the space $\mathcal{F}$ of admissible stress fields

$$\begin{aligned}\mathcal{F} = \{&\mathcal{S}\in\mathbb{R}^{n\times n},\ \mathcal{S} = \mathcal{S}^T,\ \operatorname{div}\mathcal{S} = \mathbf{0},\ \mathcal{S}_{ij}\in L^2(\Omega_F),\\ &i,j = 1,\ldots,n,\ (2.10) \text{ and } (2.12) \text{ hold}\}.\end{aligned} \tag{2.23}$$

The maximizer in (2.22) is the stress field $\mathcal{S}(\mathbf{x})$, which determines the minimizing velocity field $\mathbf{u}(\mathbf{x})$ in (2.20) by Newton's law (2.6), where

$$P(\mathbf{x}) = -\frac{\operatorname{trace}\mathcal{S}(\mathbf{x})}{n}. \tag{2.24}$$

[1]For the derivation, see Appendix A in the preprint version of this article, available at http://www.math.wsu.edu/math/faculty/panchenko/welcome.html.

3. The discrete approximation of the effective viscosity. Intuitively, in highly packed suspensions, we expect that most energy is dissipated in the thin gaps between the rigid particles. Let us then define the local dissipation rate in a gap Π, between two adjacent particles in Ω_F, by

$$W_\Pi(\mathbf{u}) = \mu \int_\Pi (\mathcal{E}(\mathbf{u}), \mathcal{E}(\mathbf{u}))\, d\mathbf{x}. \tag{3.1}$$

In this paper, we show that, in the asymptotic limit of infinitesimally small gap thickness

$$\frac{\delta}{a} = \max_{j,k} \frac{\delta^{jk}}{a} \to 0, \tag{3.2}$$

the effective viscosity is determined by the sum of local dissipation rates (3.1) over all the gaps in Ω_F. In a highly packed suspension, the rate of dissipation of the energy can be written as an asymptotic series, in the limit (3.2), with the first and second terms blowing up at different rates (as powers or at least logarithmically in a/δ). The remainder of the series is $O(1)$.

3.1. Connectivity patterns for densely packed suspensions. In the case of regular (cubic, hexagonal, etc.) arrays of particles in Ω, the volume fraction is sufficient to describe the distance between the particles and therefore the effective behavior of the suspension. However, for general distributions of particles in highly packed suspensions, one has to consider irregular connectivity patterns.

Let us consider an arbitrary distribution of particles $D^{(i)}$, centered at $\mathbf{x}^{(i)} \in \Omega$, for $i = 1, 2, \ldots, N$. We suppose that N is close to $N_{\max}$ such that particles can get close to touching one another. The concept of adjacent particles is essential to the analysis, and, to make it rigorous, we use Voronoi tessellations.

DEFINITION 3.1. *The Voronoi cell V_i, corresponding to $\mathbf{x}^{(i)}$, is the polyhedron*

$$V_i = \{\mathbf{x} \in \overline{\Omega} \text{ such that } |\, \mathbf{x} - \mathbf{x}^{(i)} \,| \leq |\, \mathbf{x} - \mathbf{x}^{(j)} \,| \quad \text{for all } j = 1, 2, \ldots, N,\ j \neq i\}.$$

The plane faces of V_i can lie either on $\partial\Omega$ or in the interior of Ω. On each face of V_i that lies inside Ω,

$$|\, \mathbf{x} - \mathbf{x}^{(i)} \,| = |\, \mathbf{x} - \mathbf{x}^{(j)} \,| \quad \text{for some } i \neq j.$$

In Figure 1, we illustrate a Voronoi tessellation in two dimensions.

DEFINITION 3.2. *Given the Voronoi tessellation and an arbitrary $D^{(i)}$, for $i = 1, 2, \ldots, N$, we define the set of indices of its neighbors as $\mathcal{N}_i = \{j \in \mathbb{N}, j \neq i$, such that V_i and V_j have a common face$\}$. The coordination number of $D^{(i)}$ is equal to the cardinal number of $\mathcal{N}_i$.*

Neighboring particles $D^{(i)}$ and $D^{(j)}$ are separated by a gap (neck) Π^{ij} (see Figure 3) of minimum thickness,

$$\delta^{ij} = |\, \mathbf{x}^{(i)} - \mathbf{x}^{(j)} \,| - (a_i + a_j), \tag{3.3}$$

and width $R^{ij} = O(a^{ij})$, where

$$a^{ij} = \frac{2a_i a_j}{a_i + a_j}. \tag{3.4}$$

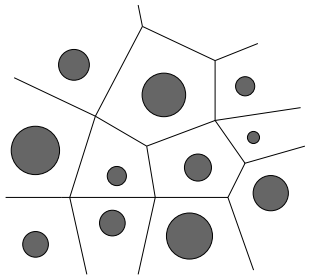

FIG. 1. *Two dimensional Voronoi tessellation.*

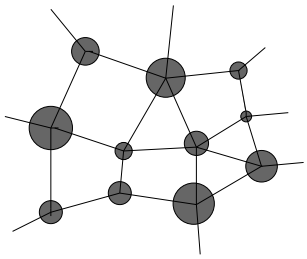

FIG. 2. *The Delaunay graph.*

Then the topology of network Γ, needed in the asymptotic approximation of $\langle\mu\rangle$, in the limit $\delta^{ij}/a^{ij} \to 0$, for $i = 1, \dots, N$ and $j \in \mathcal{N}_i$, is uniquely defined, as follows.

DEFINITION 3.3. *The interior vertices of the network (graph)* Γ *are given by* $\mathbf{x}^{(i)}$*, the locations of the centers of particles* $D^{(i)}$ *in* Ω *for* $i = 1, 2, \dots, N$*. The interior branches (edges)* b^{ij} *of the network connect vertices* $\mathbf{x}^{(i)}$ *and* $\mathbf{x}^{(j)}$ *(*$j \in \mathcal{N}_i$*) through the gaps (necks)* Π^{ij}*. For Voronoi cells* V_i *with faces belonging to* $\partial\Omega^+ \bigcup \partial\Omega^-$*, we join* $\mathbf{x}^{(i)}$ *with* $\partial\Omega^{\pm}$ *through a normal segment* $\tilde{b}^i$ *(exterior branch or edge) and we call the intersection* $\tilde{\mathbf{x}}^{(i)}$ *an exterior vertex. Finally, we let* $\mathcal{B}$ *be the set of indices* i *corresponding to the boundary Voronoi cells, that is, the cells at least one face of which belongs to* $\partial\Omega^+ \cup \partial\Omega^-$*.*

ASSUMPTION 3.1. *We assume that the distances between the neighboring balls are bounded below by* $c\delta$*, where* $c > 0$ *is fixed and* δ *is the small parameter of the problem. Thus the length of each edge in the graph is larger than* $2A + c\delta$*, where* A *is the smallest ball radius.*

Note that Γ is the Delaunay graph, which is dual to the Voronoi tessellation. The Delaunay graph for the two-dimensional tessellation of Figure 1 is shown in Figure 2. Note also the following properties of Γ,[2] which we use in the analysis.

Property 3.1. Γ is connected in the sense that each pair of interior vertices can be connected by a path consisting entirely of interior edges.

Property 3.2. Suppose there exists a Voronoi cell contained strictly inside Ω. Then there exists a closed path consisting entirely of interior edges.

Property 3.3. At least two edges originate at each interior vertex of Γ.

3.2. The two-term discrete asymptotic approximation. The asymptotic approximation of the viscous dissipation rate in the high-contrast suspension is obtained by summing the local dissipation rates $W_{\Pi^{ij}}$ in the gaps Π^{ij} between $D^{(i)}$ and $D^{(j)}$ for $i = 1, \dots, N$ and $j \in \mathcal{N}_i$. Then, focusing attention on one such gap (see Figure 3), we introduce a local system of coordinates $(x_1, \dots, x_n)$ in Π^{ij}, with the origin at $(\mathbf{x}^{(i)} + \mathbf{x}^{(j)})/2$ and coordinate x_n measured along the axis of the centers, pointing from $\mathbf{x}^{(j)}$ toward $\mathbf{x}^{(i)}$. The width of the gap is $R^{ij} = O(a^{ij})$ and the height (thickness) is

$$h(r) = \delta^{ij} + a_i\left(1 - \sqrt{1 - \frac{r^2}{a_i^2}}\right) + a_j\left(1 - \sqrt{1 - \frac{r^2}{a_j^2}}\right), \quad r = \sqrt{\sum_{k=1}^{n-1} x_k^2} \le R^{ij}. \tag{3.5}$$

[2]For the proof, see Appendix B in the preprint version of this article, available at http://www.math.wsu.edu/math/faculty/panchenko/welcome.html.

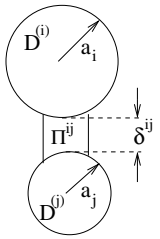

FIG. 3. *Two nearby particles $D^{(i)}$ and $D^{(j)}$ of radii a_i and a_j, respectively, separated by a gap δ^{ij}.*

The dissipation rate density $\mu(\mathcal{E},\mathcal{E})$ is expected to be highest at radial distances $r \ll \min(a_i, a_j)$, so, in the calculation of $W_{\Pi^{ij}} = \mu \int_{\Pi^{ij}} (\mathcal{E},\mathcal{E}) d\mathbf{x}$, we can approximate the spherical surfaces by paraboloids and the thickness of the gap by

$$h(r) \approx \delta^{ij} + \frac{r^2}{a^{ij}}. \tag{3.6}$$

3.2.1. The two-dimensional discrete approximation of $\langle\mu\rangle$. Clearly, $W_{\Pi^{ij}}$ depends on the velocity at the top and bottom surfaces of the gap, where $x_2 = \pm h(x_1)/2$, in two dimensions. Using boundary conditions (2.11) at $\partial D^{(i)}$ and $\partial D^{(j)}$, and approximating the outer normals as $\mathbf{n}^{(i)} \approx \frac{x_1}{a_i}\mathbf{e}_1 - \mathbf{e}_2$ and $\mathbf{n}^{(j)} \approx \frac{x_1}{a_j}\mathbf{e}_1 + \mathbf{e}_2$, by the normal vectors to the parabolas touching the disks (see (3.6)), we have

$$\begin{aligned}\mathbf{u}\left(x_1, \pm\frac{h(x_1)}{2}\right) \approx{}& \pm (T_2^{(i)} - T_2^{(j)})\frac{\mathbf{e}_2}{2} \pm (T_1^{(i)} - T_1^{(j)} + a_i\omega^{(i)} + a_j\omega^{(j)})\frac{\mathbf{e}_1}{2} \\ & \pm (\omega^{(i)} - \omega^{(j)}) x_1 \mathbf{e}_2 + \boldsymbol{\mathcal{R}},\end{aligned} \tag{3.7}$$

where

$$\boldsymbol{\mathcal{R}} = [T_2^{(i)} + T_2^{(j)} + (\omega^{(i)} + \omega^{(j)})x_1]\frac{\mathbf{e}_2}{2} + (T_1^{(i)} + T_1^{(j)} + a_i\omega^{(i)} - a_j\omega^{(j)})\frac{\mathbf{e}_1}{2}. \tag{3.8}$$

Equation (3.7) can be viewed as a decomposition of $\mathbf{u}$ in the following elementary velocity fields:

1. The first elementary velocity field in (3.7) is $\mathbf{u}_{\rm sp}$, and it solves the Stokes equations in Π^{ij} with boundary conditions

$$\mathbf{u}_{\rm sp}\left(x_1, \pm\frac{h(x_1)}{2}\right) = \pm(T_2^{(i)} - T_2^{(j)})\frac{\mathbf{e}_2}{2}. \tag{3.9}$$

We can associate $\mathbf{u}_{\rm sp}$ with the oscillatory motion, along $\mathbf{e}_2$, of two particles joined by a spring, with elastic constant $C_{\rm sp}^{ij} = O((a^{ij}/\delta^{ij})^{3/2})$ (see sections 4, 5, and 6). The velocity $(T_2^{(i)} - T_2^{(j)})/2$ of the particles is constant and unknown, so far. It is to be determined later from the global conditions of mechanical equilibrium of all inclusions in the suspension.

2. The second term in (3.7), denoted by $\mathbf{u}_{\rm sh}$, satisfies the Stokes equations in Π^{ij} with boundary conditions

$$\mathbf{u}_{\rm sh}\left(x_1, \pm\frac{h(x_1)}{2}\right) = \pm(T_1^{(i)} - T_1^{(j)} + a_i\omega^{(i)} + a_j\omega^{(j)})\frac{\mathbf{e}_1}{2}. \tag{3.10}$$

This accounts for a shear strain in the gap, where the fluid moves to the right and left, at the top and bottom surfaces of Π^{ij}, respectively, at the constant, unknown velocity $(T_1^{(i)} - T_1^{(j)} + a_i\omega^{(i)} + a_j\omega^{(j)})/2$. The contribution of this term to the dissipation rate is $C_{\rm sh}^{ij} = O(\sqrt{a^{ij}/\delta^{ij}})$ (see sections 4, 5).

3. The third term in (3.7) corresponds to a shear strain in the gap due to rotations. The boundary conditions are given by

$$\mathbf{u}_{\rm rot}\left(x_1, \pm\frac{h(x_1)}{2}\right) = \pm(\omega^{(i)} - \omega^{(j)})x_1\mathbf{e}_2, \tag{3.11}$$

as if the fluid were pushed and pulled, in direction $\mathbf{e}_2$, on the left and right sides of Π^{ij}, respectively (see Figure 6). The contribution of this term to the dissipation rate is $C_{\rm rot}^{ij} = O(\sqrt{a^{ij}/\delta^{ij}})$ (see sections 4, 5).

4. Finally, the remainder $\boldsymbol{\mathcal{R}}$ corresponds to a constant, $O(1)$ shear strain in the gap and, as such, it gives an $O(1)$ contribution to $W_{\Pi^{ij}}$ (see sections 4, 5).

In section 4, we obtain the formal asymptotic approximation

$$W_{\Pi^{ij}} \approx C_{\rm sp}^{ij}(T_2^{(i)} - T_2^{(j)})^2 + C_{\rm sh}^{ij}(T_1^{(i)} - T_1^{(j)} + a_i\omega^{(i)} + a_j\omega^{(j)})^2 + C_{\rm rot}^{ij}(\omega^{(i)} - \omega^{(j)})^2 + O(1), \tag{3.12}$$

where

$$C_{\rm sp}^{ij} = \frac{3\pi\mu}{4}\left(\frac{a^{ij}}{\delta^{ij}}\right)^{\frac{3}{2}} + \frac{12\pi\mu}{5}\sqrt{\frac{a^{ij}}{\delta^{ij}}},\quad C_{\rm sh}^{ij} = \frac{\pi\mu}{2}\sqrt{\frac{a^{ij}}{\delta^{ij}}},\ \text{and}\ C_{\rm rot}^{ij} = \frac{9\pi\mu}{16}\sqrt{\frac{a^{ij}}{\delta^{ij}}}. \tag{3.13}$$

The approximation (3.12) applies to interior inclusions $D^{(i)}$. For $i \in \mathcal{B}$, we have $D^{(i)}$ joined to a fictitious disk of infinite radius (i.e., $\partial\Omega^+$ or $\partial\Omega^-$) and the harmonic average of the radii is $a^i = 2a_i$. Given boundary conditions (2.9) at $\partial\Omega^{\pm}$, we have, similar to (3.12),

$$W_{\Pi^i} \approx C_{\rm sp}^{i}(T_2^{(i)} - \mathbf{g}\cdot\mathbf{e}_2)^2 + C_{\rm sh}^{i}(T_1^{(i)} - \mathbf{g}\cdot\mathbf{e}_1 + a_i\omega^{(i)})^2 + C_{\rm rot}^{i}(2\omega^{(i)})^2 + O(1), \tag{3.14}$$

where $C_{\rm sp}^i$, $C_{\rm sh}^i$, and $C_{\rm rot}^i$ are given by (3.13), with a^{ij} replaced by $a^i = 2a_i$ and δ^{ij} replaced by δ^i, the distance between $\partial D^{(i)}$ and the upper or lower boundary $\partial\Omega^{\pm}$.

Next, we approximate E by summing the local dissipation rates in all gaps Π^{ij} for $i = 1, \ldots, N$, $i \notin \mathcal{B}$, $j \in \mathcal{N}_i$, and Π^i for $i \in \mathcal{B}$. For this purpose, let us rename the orthonormal basis vectors in each gap Π^{ij} as

$$\mathbf{q}^{ij} = \frac{\mathbf{x}^{(i)} - \mathbf{x}^{(j)}}{|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}|}$$

and $\mathbf{p}^{ij}$ = the rotated $\mathbf{q}^{ij}$, clockwise, by $\pi/2$, in the two-dimensional plane. In the boundary gaps Π^i, joining a particle $D^{(i)}$ with $\partial\Omega^{\pm}$, these vectors are called $\mathbf{q}^i$ and $\mathbf{p}^i$, respectively. The discrete approximation of $\langle\mu\rangle$ is given by (2.19), with the right-hand side

$$\begin{aligned} E \approx \min_{\mathbf{T},\omega} \sum_{i=1}^{N} \sum_{\substack{j\in\mathcal{N}_i \\ j<i}} &\{C_{\rm sp}^{ij}[(\mathbf{T}^{(i)} - \mathbf{T}^{(j)})\cdot\mathbf{q}^{ij}]^2 + C_{\rm sh}^{ij}[(\mathbf{T}^{(i)} - \mathbf{T}^{(j)})\cdot\mathbf{p}^{ij} + a_i\omega^{(i)} + a_j\omega^{j}]^2 \\ &+ C_{\rm rot}^{ij}(\omega^{(i)} - \omega^{(j)})^2\} + \sum_{i\in\mathcal{B}}\{C_{\rm sp}^{i}[(\mathbf{T}^{(i)} - \mathbf{g})\cdot\mathbf{q}^{i}]^2 + C_{\rm rot}^{i}(2\omega^{(i)})^2 \\ &+ C_{\rm sh}^{i}[(\mathbf{T}^{(i)} - \mathbf{g})\cdot\mathbf{p}^{i} + a_i\omega^{(i)}]^2\}. \end{aligned} \tag{3.15}$$

Note that in (3.15) we minimize a quadratic functional, over translational and rotational velocities $\mathbf{T}^{(i)}$ and $\omega^{(i)}$, for $i = 1, \ldots, N$, respectively. This is equivalent to solving the Euler–Lagrange equations

$$\sum_{j\in\mathcal{N}_i} \{C^{ij}_{\rm sp}[(\mathbf{T}^{(i)} - \mathbf{T}^{(j)})\cdot\mathbf{q}^{ij}]\mathbf{q}^{ij} + C^{ij}_{\rm sh}[(\mathbf{T}^{(i)} - \mathbf{T}^{(j)})\cdot\mathbf{p}^{ij} + a_i\omega^{(i)} + a_j\omega^{(j)}]\mathbf{p}^{ij}\}$$
$$+\,\mathbf{F}_{\mathcal{B}}(\mathbf{T}^{(i)},\omega^{(i)}) = \mathbf{0}, \tag{3.16}$$

$$\sum_{j\in\mathcal{N}_i} \{C^{ij}_{\rm sh}[(\mathbf{T}^{(i)} - \mathbf{T}^{(j)})\cdot\mathbf{p}^{ij} + \omega^{(i)} + \omega^{(j)}] + C^{ij}_{\rm rot}(\omega^{(i)} - \omega^{(j)})\}$$
$$+\,\mathcal{M}_{\mathcal{B}}(\mathbf{T}^{(i)},\omega^{(i)}) = 0 \tag{3.17}$$

for all $i = 1, \ldots, N$, where

$$\mathbf{F}_{\mathcal{B}}(\mathbf{T}^{(i)},\omega^{(i)}) = \begin{cases} C^{i}_{\rm sp}[(\mathbf{T}^{(i)} - \mathbf{g})\cdot\mathbf{q}^{i}]\mathbf{q}^{i} + C^{i}_{\rm sh}[(\mathbf{T}^{(i)} - \mathbf{g})\cdot\mathbf{p}^{i} + a_i\omega^{(i)}]\mathbf{p}^{i} & \text{if } i\in\mathcal{B}, \\ \mathbf{0} \ \text{ otherwise,} \end{cases} \tag{3.18}$$

$$\mathcal{M}_{\mathcal{B}}(\mathbf{T}^{(i)},\omega^{(i)}) = \begin{cases} C^{i}_{\rm sh}[(\mathbf{T}^{(i)} - \mathbf{g})\cdot\mathbf{p}^{i} + a_i\omega^{(i)}] + 4C^{i}_{\rm rot}\omega^{(i)} & \text{if } i\in\mathcal{B}, \\ 0 \ \text{ otherwise.} \end{cases} \tag{3.19}$$

These are the equations of force and torque balance of the inclusions, and the minimization in (3.15) ensures that the rigid body translational and rotational velocities are chosen in such a way that the suspension is in mechanical equilibrium.

3.2.2. The three-dimensional discrete approximation of $\langle\boldsymbol{\mu}\rangle$. Consider the local system of coordinates described at the beginning of section 3.2 in three dimensions. Similar to our two-dimensional calculation, we write

$$\begin{aligned} \mathbf{u}\left(x_1, x_2, \pm\frac{h(x_1,x_2)}{2}\right) \approx{}& \pm(T_3^{(i)} - T_3^{(j)})\frac{\mathbf{e}_3}{2} \pm (T_1^{(i)} - T_1^{(j)} - a_i\omega_2^{(i)} - a_j\omega_2^{(j)})\frac{\mathbf{e}_1}{2} \\ & \pm (T_2^{(i)} - T_2^{(j)} + a_i\omega_1^{(i)} + a_j\omega_1^{(j)})\frac{\mathbf{e}_2}{2} \pm (\omega_1^{(i)} - \omega_1^{(j)})\frac{x_2\mathbf{e}_3}{2} \\ & \mp (\omega_2^{(i)} - \omega_2^{(j)})\frac{x_1\mathbf{e}_3}{2} + \boldsymbol{\mathcal{R}}, \end{aligned} \tag{3.20}$$

where the remainder is

$$\begin{aligned} \boldsymbol{\mathcal{R}} = {}& (T_3^{(i)} + T_3^{(j)})\frac{\mathbf{e}_3}{2} + (T_1^{(i)} + T_1^{(j)} - a_i\omega_2^{(i)} + a_j\omega_2^{(j)})\frac{\mathbf{e}_1}{2} \\ & + (T_2^{(i)} + T_2^{(j)} + a_i\omega_1^{(i)} - a_j\omega_1^{(j)})\frac{\mathbf{e}_2}{2} + (\omega_1^{(i)} + \omega_1^{(j)})\frac{x_2\mathbf{e}_3}{2} \\ & - (\omega_2^{(i)} - \omega_2^{(j)})\frac{x_1\mathbf{e}_3}{2}. \end{aligned} \tag{3.21}$$

As in two dimensions, we associate the first term in (3.20), due to the motion of the inclusions along the axis of their centers, with the oscillatory motion of two particles

joined by a spring of elastic constant $C^{ij}_{\rm sp} = O(a^{ij}/\delta^{ij})$. The next two terms in (3.20) correspond to shear strains in the gap, where the fluid is pulled in the positive and negative directions of $\mathbf{e_1}$ and $\mathbf{e}_2$ at the top and bottom surfaces of Π^{ij}, respectively. The contribution of these terms to the dissipation rate is $C^{ij}_{\rm sh} = O(\ln a^{ij}/\delta^{ij})$. The fourth and fifth terms in (3.20) correspond to a shear strain in the gap as well, but now the fluid is pushed and pulled, in direction $\mathbf{e}_3$, on opposite sides of the axis of Π^{ij}, respectively. The contribution of these terms to the dissipation rate is $C^{ij}_{\rm rot} = O(\ln a^{ij}/\delta^{ij})$. Finally, the remainder $\boldsymbol{\mathcal{R}}$ gives an $O(1)$ contribution to $W_{\Pi^{ij}}$.

A formal asymptotic analysis, which is very similar to the two-dimensional one in section 4 and, as such, is not detailed here, gives

$$\begin{aligned} W_{\Pi^{ij}} \approx\ & C^{ij}_{\rm sp}(T_3^{(i)} - T_3^{(j)})^2 + C^{ij}_{\rm rot}[(\omega_1^{(i)} - \omega_1^{(j)})^2 + (\omega_2^{(i)} - \omega_2^{(j)})^2] \\ & + C^{ij}_{\rm sh}[(T_1^{(i)} - T_1^{(j)} - a_i\omega_2^{(i)} - a_j\omega_2^{(j)})^2 \\ & \qquad + (T_2^{(i)} - T_2^{(j)} + a_i\omega_1^{(i)} + a_j\omega_1^{(j)})^2] + O(1), \end{aligned} \tag{3.22}$$

where

$$C^{ij}_{\rm sp} = \frac{3\pi\mu a^{ij}}{4}\left(\frac{a^{ij}}{\delta^{ij}}\right) + \frac{9\pi\mu a^{ij}}{5}\ln\frac{a^{ij}}{\delta^{ij}},\quad C^{ij}_{\rm sh} = \frac{\pi\mu a^{ij}}{2}\ln\frac{a^{ij}}{\delta^{ij}},\ \text{and}\ C^{ij}_{\rm rot} = \frac{9\pi\mu a^{ij}}{16}\ln\frac{a^{ij}}{\delta^{ij}}(a^{ij})^2. \tag{3.23}$$

If $i \in \mathcal{B}$, $D^{(i)}$ is joined to $\partial\Omega^{\pm}$ through gap Π^i and we obtain (see section 3.2.1)

$$\begin{aligned} W_{\Pi^i} \approx\ & C^i_{\rm sp}(T_3^{(i)} - \mathbf{g}\cdot\mathbf{e}_3)^2 + C^i_{\rm rot}[(2\omega_1^{(i)})^2 + (2\omega_2^{(i)})^2] \\ & + C^i_{\rm sh}[(T_1^{(i)} - \mathbf{g}\cdot\mathbf{e}_1 - a_i\omega_2^{(i)})^2 + (T_2^{(i)} - \mathbf{g}\cdot\mathbf{e}_2 + a_i\omega_1^{(i)})^2] + O(1) \end{aligned} \tag{3.24}$$

with constants $C^i_{\rm sp}$, $C^i_{\rm sh}$, and $C^i_{\rm rot}$ given by (3.23), where a^{ij} is replaced by $a^i = 2a_i$ and δ^{ij} is replaced by δ^i, the distance between $\mathbf{x}^{(i)}$ and the upper or lower boundary.

The discrete approximation of the viscous dissipation rate E in the suspension is given by the sum of the local dissipation rates in all gaps Π^{ij} for $i = 1, \ldots, N,\ i \notin \mathcal{B}$, $j \in \mathcal{N}_i$, and Π^i, for $i \in \mathcal{B}$. Let us then introduce, in gap Π^{ij}, the orthonormal vectors

$$\mathbf{q}^{ij} = \frac{\mathbf{x}^{(i)} - \mathbf{x}^{(j)}}{|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}|},\ \mathbf{r}^{ij} = \text{ the rotated } \mathbf{q}^{ij},$$
$$\text{in the plane } \mathcal{P}^{ij},\ \text{clockwise, by } \frac{\pi}{2},\ \text{and } \mathbf{p}^{ij} = \mathbf{r}^{ij} \times \mathbf{q}^{ij},$$

where $\mathbf{p}^{ij}$, $\mathbf{r}^{ij}$, and $\mathbf{q}^{ij}$ play the role of $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$, respectively, in the above calculation of $W_{\Pi^{ij}}$. Since plane $\mathcal{P}^{ij}$ is not uniquely defined by $\mathbf{q}^{ij}$, there are infinitely many choices of $\mathbf{r}^{ij}$, although they give the same dissipation rate $W_{\Pi^{ij}}$. Let us then pick $\mathcal{P}^{ij}$ arbitrarily but ensure, at the same time, that when indices i and j are interchanged, we have

$$\mathcal{P}^{ij} = \text{span}\{\mathbf{q}^{ij}, \mathbf{r}^{ij}\} = \text{span}\{\mathbf{q}^{ji}, \mathbf{r}^{ji}\} = \mathcal{P}^{ji}$$

or, equivalently,

$$\mathbf{q}^{ij} = -\mathbf{q}^{ji},\ \mathbf{r}^{ij} = -\mathbf{r}^{ji},\ \text{and}\ \mathbf{p}^{ij} = \mathbf{p}^{ji} \quad \text{for all } i = 1, \ldots, N,\ i \neq \mathcal{B},\ j \in \mathcal{N}_i.$$

For $i \in \mathcal{B}$, $D^{(i)}$ is joined with $\partial\Omega^{\pm}$ by gap Π^i, and the unit vectors are denoted by $\mathbf{p}^i$, $\mathbf{r}^i$, and $\mathbf{q}^i$, respectively. Then, the discrete approximation of the effective viscosity is given by (2.19), with the right-hand side

$$
\begin{aligned}
E \approx \min_{\mathbf{T},\omega} \sum_{i=1}^{N} \sum_{\substack{j \in \mathcal{N}_i \\ j<i}} & \{C_{\rm sp}^{ij}[(\mathbf{T}^{(i)} - \mathbf{T}^{(j)}) \cdot \mathbf{q}^{ij}]^2 + C_{\rm rot}^{ij}[(\boldsymbol{\omega}^{(i)} - \boldsymbol{\omega}^{(j)}) \cdot \mathbf{p}^{ij}]^2 \\
& + C_{\rm rot}^{ij}[(\boldsymbol{\omega}^{(i)} - \boldsymbol{\omega}^{(j)}) \cdot \mathbf{r}^{ij}]^2 + C_{\rm sh}^{ij}[(\mathbf{T}^{(i)} - \mathbf{T}^{(j)}) \cdot \mathbf{p}^{ij} \\
& - (a_i \boldsymbol{\omega}^{(i)} + a_j \boldsymbol{\omega}^{(j)}) \cdot \mathbf{r}^{ij}]^2 \\
& + C_{\rm sh}^{ij}[(\mathbf{T}^{(i)} - \mathbf{T}^{(j)}) \cdot \mathbf{r}^{ij} + (a_i \boldsymbol{\omega}^{(i)} + a_j \boldsymbol{\omega}^{(j)}) \cdot \mathbf{p}^{ij}]^2\} \\
& + \sum_{i \in \mathcal{B}} \{C_{\rm sp}^{i}[(\mathbf{T}^{(i)} - \mathbf{g}) \cdot \mathbf{q}^{i}]^2 \\
& \qquad + C_{\rm rot}^{i}[(2\boldsymbol{\omega}^{(i)} \cdot \mathbf{p}^{i})^2 + (2\boldsymbol{\omega}^{(i)} \cdot \mathbf{r}^{i})^2] \\
& \qquad + C_{\rm sh}^{i}[(\mathbf{T}^{(i)} - \mathbf{g}) \cdot \mathbf{p}^{i} - a_i \boldsymbol{\omega}^{(i)} \cdot \mathbf{r}^{i}]^2 \\
& \qquad + C_{\rm sh}^{i}[(\mathbf{T}^{(i)} - \mathbf{g}) \cdot \mathbf{r}^{i} + a_i \boldsymbol{\omega}^{(i)} \cdot \mathbf{p}^{i}]^2\} + O(1).
\end{aligned} \tag{3.25}
$$

Finally, as in section 3.2.1, the minimization in (3.25), over translational and rotational velocities $\mathbf{T}^{(i)}$ and $\boldsymbol{\omega}^{(i)}$ for $i = 1, \ldots, N$, ensures that all the inclusions in the suspension are in mechanical equilibrium.

Remark 3.1 (computation of the effective viscosity). We now summarize the steps necessary to compute the effective viscosity in problem (2.7)–(2.12). First, compute the approximate dissipation rate E by minimizing the quadratic functional (3.15) (in two dimensions) or (3.25) (in three dimensions). Next, solve the Stokes equations in the domain Ω (see (2.1)) with viscosity equal to one and boundary conditions given by (2.9), (2.10). Then compute the corresponding strain rate $\mathcal{E}^0 = 1/2(\nabla \mathbf{u}^0 + \nabla^T \mathbf{u}^0)$ and the normalized dissipation rate $\int_\Omega \mathcal{E}_{ij}^0 \mathcal{E}_{ij}^0 d\mathbf{x}$. Finally, compute the approximate value of the effective viscosity by the formula

$$
\langle \mu \rangle = \frac{E}{\int_\Omega \mathcal{E}_{ij}^0 \mathcal{E}_{ij}^0 d\mathbf{x}}. \tag{3.26}
$$

When the contributions of rotations can be neglected, the leading term in (3.26) is given by the leading term in the formula (6.85). Note that this term corresponds to the spring network approximation, which takes into account only motions of particles along the line of their centers. Detailed analysis of computational formulas for $\langle \mu \rangle$, based on the approach developed in this paper, for various boundary conditions and different arrays of particles, is presented in [7].

4. The local dissipation rate in a gap between two adjacent particles. Formal asymptotics in two dimensions. We begin our estimation of E with a formal asymptotic analysis which extends the lubrication approximations in [15, 16, 27] beyond the leading term by accounting for all possible rigid body motions of the inclusions in the suspension. To find E, we construct a velocity field in Ω_F which satisfies boundary conditions (2.11) but solves the Stokes equations approximately in the following sense: Since the density $\mu(\mathcal{E}, \mathcal{E})$ of the viscous dissipation rate is very high near the axis of the centers of adjacent inclusions $D^{(i)}$ and $D^{(j)}$, we approximate $\mathbf{u}$ in each gap Π^{ij} by the solution of the Stokes problem between two parallel plates, at distance h (which we pretend is a constant) apart, and we calculate the corresponding

rate of strain $\mathcal{E}$. Then we integrate over the gap to obtain the local dissipation rate

$$W_{\Pi^{ij}} \approx \int_{-a^{ij}}^{a^{ij}} dx_1 \int_{-\frac{h(x_1)}{2}}^{\frac{h(x_1)}{2}} dx_2 \, \mu\left(\mathcal{E},\mathcal{E}\right).$$

Since most energy is dissipated in the gaps, we expect that the contribution to E from the region outside the gaps remains uniformly bounded in the limit $\delta \to 0$.

Let us denote by E^{a} the approximation of the dissipation rate, obtained with the formal asymptotic, lubrication type, approach. Since E^{a} is a heuristic estimate, it requires rigorous justification, which we give in sections 5 and 6, where we calculate upper and lower variational bounds on E that match E^{a} to leading order. Nevertheless, both bounds are inspired to some extent by the calculation of E^{a}, so we describe next, in detail, our formal asymptotic analysis.

We begin by recalling the local system of coordinates (x_1, x_2) in gap Π^{ij}, as defined in section 3.2. At the surface of $D^{(i)}$, the velocity is given by

$$\mathbf{u}\mid_{\partial D^{(i)}} = (T_1^{(i)} + a_i\omega^{(i)})\mathbf{e}_1 + (T_2^{(i)} + a_i\omega^{(i)}n_1^{(i)})\mathbf{e}_2 - a_i\omega^{(i)}(n_2^{(i)}+1)\mathbf{e}_1,$$

and the two components of the outer normal at $\partial D^{(i)}$ are $n_1^{(i)} = \frac{x_1}{a_i}$ and $n_2^{(i)} = -\sqrt{1-\frac{x_1^2}{a_i^2}}$. Similarly,

$$\mathbf{u}\mid_{\partial D^{(j)}} = (T_1^{(j)} - a_j\omega^{(j)})\mathbf{e}_1 + (T_2^{(j)} + a_j\omega^{(j)}n_1^{(j)})\mathbf{e}_2 - a_j\omega^{(j)}(n_2^{(j)}-1)\mathbf{e}_1,$$

where $n_1^{(j)} = \frac{x_1}{a_j}$ and $n_2^{(j)} = \sqrt{1-\frac{x_1^2}{a_j^2}}$. Equivalently, we rewrite the boundary conditions on $\mathbf{u}$ as

(4.1)

$$\begin{aligned}\mathbf{u}\left(x_1, \pm\frac{h}{2}\right) = &\pm\left(T_1^{(i)} - T_1^{(j)} + \omega^{(i)}\sqrt{a_i^2-x_1^2} + \omega^{(j)}\sqrt{a_j^2-x_1^2}\right)\frac{\mathbf{e}_1}{2} \pm (T_2^{(i)} - T_2^{(j)})\frac{\mathbf{e}_2}{2}\\ &\pm(\omega^{(i)}-\omega^{(j)})\frac{x_1\mathbf{e}_2}{2} + \left(T_1^{(i)} + T_1^{(j)} + \omega^{(i)}\sqrt{a_i^2-x_1^2} - \omega^{(j)}\sqrt{a_j^2-x_1^2}\right)\frac{\mathbf{e}_1}{2}\\ &+(T_2^{(i)} + T_2^{(j)})\frac{\mathbf{e}_2}{2} + (\omega^{(i)}+\omega^{(j)})\frac{x_1\mathbf{e}_2}{2} \approx \mathbf{u}^{\mathrm{a}}\left(x_1, \pm\frac{h}{2}\right),\end{aligned}$$

where $\mathbf{u}^{\mathrm{a}}$ is an approximation of the velocity field near the axis of the gap (i.e., for $x_1/a^{ij} \ll 1$, where the density of the dissipation rate is highest). The boundary conditions on $\mathbf{u}^{\mathrm{a}}$ are

$$\begin{aligned}\mathbf{u}^{\mathrm{a}}\left(x_1, \pm\frac{h}{2}\right) = &\pm(T_1^{(i)} - T_1^{(j)} + a_i\omega^{(i)} + a_j\omega^{(j)})\frac{\mathbf{e}_1}{2}\\ &\pm(T_2^{(i)} - T_2^{(j)})\frac{\mathbf{e}_2}{2} \pm (\omega^{(i)}-\omega^{(j)})\frac{x_1\mathbf{e}_2}{2}\\ &+(T_1^{(i)} + T_1^{(j)} + a_i\omega^{(i)} - a_j\omega^{(j)})\frac{\mathbf{e}_1}{2}\\ &+(T_2^{(i)} + T_2^{(j)})\frac{\mathbf{e}_2}{2} + (\omega^{(i)}+\omega^{(j)})\frac{x_1\mathbf{e}_2}{2},\end{aligned}$$

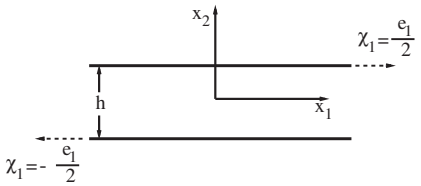

FIG. 4. *The setup for the calculation of* $\boldsymbol{\chi}_1$.

and, using the linearity of the problem, we decompose $\mathbf{u}^{\mathrm{a}}$ as

$$\mathbf{u}^{\mathrm{a}} = (T_1^{(i)} - T_1^{(j)} + a_i\omega^{(i)} + a_j\omega^{(j)})\boldsymbol{\chi}_1 + (T_2^{(i)} - T_2^{(j)})\boldsymbol{\chi}_2 + (\omega^{(i)} - \omega^{(j)})\boldsymbol{\lambda} + \boldsymbol{\mathcal{R}}, \tag{4.2}$$

where $\boldsymbol{\chi}_1, \boldsymbol{\chi}_2, \boldsymbol{\lambda}$, and $\boldsymbol{\mathcal{R}}$ are elementary velocity fields satisfying

$$\boldsymbol{\chi}_k\left(x_1, \frac{h}{2}\right) = -\boldsymbol{\chi}_k\left(x_1, -\frac{h}{2}\right) = \frac{1}{2}\mathbf{e}_k, \quad k = 1, 2, \tag{4.3}$$

$$\boldsymbol{\lambda}\left(x_1, \frac{h}{2}\right) = -\boldsymbol{\lambda}\left(x_1, -\frac{h}{2}\right) = \frac{x_1}{2}\mathbf{e}_2, \tag{4.4}$$

$$\begin{aligned}\boldsymbol{\mathcal{R}}\left(x_1, \pm\frac{h}{2}\right) &= [T_2^{(i)} + T_2^{(j)} + (\omega^{(i)} + \omega^{(j)})x_1]\frac{\mathbf{e}_2}{2} \\ &\quad + (T_1^{(i)} + T_1^{(j)} + a_i\omega^{(i)} - a_j\omega^{(j)})\frac{\mathbf{e}_1}{2}.\end{aligned} \tag{4.5}$$

We approximate all elementary velocity fields in the decomposition (4.2) by solving the simplified Stokes flow problem between two parallel plates at distance h (treated as constant) apart.

Clearly, velocity field

$$\boldsymbol{\mathcal{R}}(x_1, x_2) = [T_2^{(i)} + T_2^{(j)} + (\omega^{(i)} + \omega^{(j)})x_1]\frac{\mathbf{e}_2}{2} + (T_1^{(i)} + T_1^{(j)} + a_i\omega^{(i)} - a_j\omega^{(j)})\frac{\mathbf{e}_1}{2} \tag{4.6}$$

is divergence free and satisfies (2.7) for a constant pressure field. Moreover, its rate of strain is uniformly bounded in the asymptotic limit $\delta^{ij}/a^{ij} \to 0$ and the contribution of $\boldsymbol{\mathcal{R}}$ to the local dissipation rate in Π^{ij} is negligible.

Velocity field $\boldsymbol{\chi}_1$. As we zoom in near the axis of the centers $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$, the top and bottom boundaries of Π^{ij} which belong to ∂D_i and $\partial D^{(j)}$, respectively, are approximated by parallel planes which move in opposite directions, as shown in Figure 4. Using separation of variables, we find

$$\boldsymbol{\chi}_1(x_1, x_2) = \left[\frac{x_2}{h} + \frac{C}{2\mu}\left(x_2^2 - \frac{h^2}{4}\right)\right]\mathbf{e}_1. \tag{4.7}$$

Integrating,[3] we obtain

$$W^{\chi_1}_{\Pi^{ij}} = \frac{\mu}{4}\int_{-a^{ij}}^{a^{ij}} dx_1 \int_{-\frac{h(x_1)}{2}}^{\frac{h(x_1)}{2}} dx_2 \left(\nabla\boldsymbol{\chi}_1 + \nabla\boldsymbol{\chi}_1^T, \nabla\boldsymbol{\chi}_1 + \nabla\boldsymbol{\chi}_1^T\right) \approx \frac{\pi\mu}{2}\sqrt{\frac{a^{ij}}{\delta^{ij}}} + O(1). \tag{4.8}$$

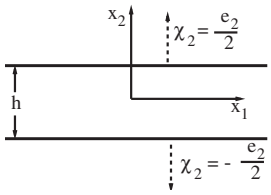

FIG. 5. *The setup for the calculation of* $\boldsymbol{\chi}_2$.

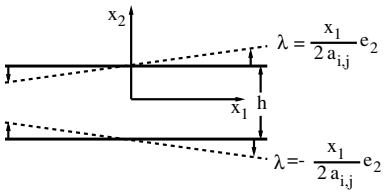

FIG. 6. *The setup for the calculation of* $\boldsymbol{\lambda}$.

Velocity field $\boldsymbol{\chi}_2$. We approximate $\boldsymbol{\chi}_2$ by the velocity of an incompressible fluid between two parallel plates which move at constant speed away from each other along the axis $\mathbf{e}_2$ (see Figure 5). Separating variables, we obtain

$$\boldsymbol{\chi}_2(x_1,x_2) \approx \frac{6x_1}{h}\left(\frac{x_2^2}{h^2}-\frac{1}{4}\right)\mathbf{e}_1 + \left[\frac{3x_2}{2h} - 2\left(\frac{x_2}{h}\right)^3\right]\mathbf{e}_2, \tag{4.9}$$

$$p(x_1,x_2) \approx \frac{6\mu x_1^2}{h^3} - \frac{6\mu}{h}\left(\frac{x_2^2}{h^2}-\frac{1}{4}\right) + C. \tag{4.10}$$

Then,

$$\begin{aligned} W_{\Pi^{ij}}^{\chi_2} &= \frac{\mu}{4}\int_{-a^{ij}}^{a^{ij}} dx_1 \int_{-\frac{h(x_1)}{2}}^{\frac{h(x_1)}{2}} dx_2 \left(\nabla\boldsymbol{\chi}_2 + \nabla\boldsymbol{\chi}_2^T, \nabla\boldsymbol{\chi}_2 + \nabla\boldsymbol{\chi}_2^T\right) \\ &\approx \frac{3\pi\mu}{4}\left(\frac{a^{ij}}{\delta^{ij}}\right)^{\frac{3}{2}} + \frac{12\pi\mu}{5}\sqrt{\frac{a^{ij}}{\delta^{ij}}} + O(1). \end{aligned} \tag{4.11}$$

Velocity field $\boldsymbol{\lambda}$. The setup for the calculation of $\boldsymbol{\lambda}$ is shown in Figure 6. In this case, the approximate solution is

$$\boldsymbol{\lambda}(x_1,x_2) \approx \frac{3x_1^2}{h}\left(\frac{x_2^2}{h^2}-\frac{1}{4}\right)\mathbf{e}_1 + \left(\frac{3x_1}{2}\frac{x_2}{h} - 2\frac{x_1x_2^3}{h^3}\right)\mathbf{e}_2, \tag{4.12}$$

$$p(x_1,x_2) \approx \frac{2\mu x_1^3}{h^3} - \frac{6\mu x_1}{h}\left(\frac{x_2^2}{h^2}-\frac{1}{4}\right), \tag{4.13}$$

and

$$W_{\Pi^{ij}}^{\lambda} = \frac{\mu}{4}\int_{x_1=-a^{ij}}^{a^{ij}} dx_1 \int_{x_2=-\frac{h(x_1)}{2}}^{\frac{h(x_1)}{2}} dx_2 (\nabla\boldsymbol{\lambda} + \nabla\boldsymbol{\lambda}^T, \nabla\boldsymbol{\lambda} + \nabla\boldsymbol{\lambda}^T) \approx \frac{9\pi\mu}{16}\sqrt{\frac{a^{ij}}{\delta}}(a^{ij})^2 + O(1). \tag{4.14}$$

[3]The integration is done in MAPLE.

Finally, using straightforward MAPLE calculations, we find that the contributions to $W_{\Pi^{ij}}$ of all the cross terms such as $(\nabla\boldsymbol{\chi}_1 + \nabla\boldsymbol{\chi}_1^T, \nabla\boldsymbol{\chi}_2 + \nabla\boldsymbol{\chi}_2^T)$ is $O(1)$. Gathering all the results, we have

$$
\begin{aligned}
W_{\Pi^{ij}} \approx & \left[\frac{3\pi\mu}{4}\left(\frac{a^{ij}}{\delta^{ij}}\right)^{\frac{3}{2}} + \frac{12\pi\mu}{5}\sqrt{\frac{a^{ij}}{\delta^{ij}}}\right](T_2^{(i)} - T_2^{(j)})^2 \\
& + \frac{\pi\mu}{2}\sqrt{\frac{aij}{\delta^{ij}}}(T_1^{(i)} - T_1^{(j)} + a_i\omega^{(i)} + a_j\omega^{(j)})^2 \\
& + \frac{9\pi\mu}{16}\sqrt{\frac{a^{ij}}{\delta^{ij}}}(a^{ij})^2(\omega^{(i)} - \omega^{(j)})^2 + O(1).
\end{aligned} \tag{4.15}
$$

This is precisely the result (3.12), and the approximation E^{a} of the viscous dissipation rate in the suspension is obtained by summing contributions (4.15) of all the gaps, as explained in section 3.2.

5. The upper bound. Any test velocity field $\mathbf{u} \in \mathcal{U}$ gives an upper bound on the viscous dissipation rate E when used in variational principle (2.20). However, of all choices of $\mathbf{u}$, we are interested in those that give tight, correct-to-leading-orders bounds on E. In this section, we give the construction of such a velocity field in two dimensions. We begin with the construction of $\mathbf{u}$ in the gap Π^{ij} between two adjacent particles $D^{(i)}$ and $D^{(j)}$ (see section 5.1), and, to capture the important features of the flow, we use the formal asymptotic analysis of section 4 as a guide. Then in section 5.2, we extend $\mathbf{u}$ to the remainder of the domain, where the flow is diffuse and, as such, contributes to $O(1)$ terms in E.

5.1. Definition of the trial velocity field u in a gap $\boldsymbol{\Pi^{ij}}$. The local construction of section 4 captures the important features of the flow in the gap Π^{ij} between adjacent particles $D^{(i)}$ and $D^{(j)}$. However, since the gap thickness $h(x_1)$ is not a constant (as it is treated in section 4), $\mathbf{u}^{\mathrm{a}}$ derived in section 4 is not divergence free and, therefore, it is not an admissible trial field in variational principle (2.20). In this section, we modify the velocity field calculated in section 4 in such a way that the incompressibility condition is satisfied and yet the effect of the corrections on E is minimal.

Using (4.1), (4.2), and the linearity of the problem, we decompose trial velocity field $\mathbf{u}$ as

$$
\begin{aligned}
\mathbf{u}(\mathbf{x}) = & (T_1^{(i)} - T_1^{(j)} + a_i\omega^{(i)} + a_j\omega^{(j)})\boldsymbol{\chi}_1(\mathbf{x}) + (T_2^{(i)} - T_2^{(j)})\boldsymbol{\chi}_2(\mathbf{x}) \\
& + (\omega^{(i)} - \omega^{(j)})\boldsymbol{\lambda}(2) + \boldsymbol{\mathcal{R}}(\mathbf{x}) + \boldsymbol{\mathcal{C}}(\mathbf{x}),
\end{aligned} \tag{5.1}
$$

where $\boldsymbol{\mathcal{R}}$ is given by (4.6), $\boldsymbol{\chi}_1, \boldsymbol{\chi}_2$, and $\boldsymbol{\lambda}$ satisfy boundary conditions (4.3) and (4.4), and

$$
\begin{aligned}
\boldsymbol{\mathcal{C}}\left(x_1, \pm\frac{h(x_1)}{2}\right) = & \left(\sqrt{1 - \frac{x_1^2}{a_i^2}} + \sqrt{1 - \frac{x_1^2}{a_j^2}} - 2\right)\left(\pm\frac{\omega^{(i)} + \omega^{(j)}}{2} + \frac{\omega^{(i)} - \omega^{(j)}}{2}\right)\frac{\mathbf{e}_1}{2} \\
& + \left(\sqrt{1 - \frac{x_1^2}{a_i^2}} - \sqrt{1 - \frac{x_1^2}{a_j^2}}\right)\left(\pm\frac{\omega^{(i)} - \omega^{(j)}}{2} + \frac{\omega^{(i)} + \omega^{(j)}}{2}\right)\frac{\mathbf{e}_1}{2}.
\end{aligned} \tag{5.2}
$$

Elementary velocity fields $\boldsymbol{\chi}_1, \boldsymbol{\chi}_2$, and $\boldsymbol{\lambda}$ have been approximated in section 4. Here, we modify their expression to ensure that they are divergence free in the gap of variable thickness $h(x_1)$. Velocity field $\boldsymbol{\mathcal{C}}$ accounts for the curvature of the gap and it has been omitted in section 4. In this section, we calculate an admissible field $\boldsymbol{\mathcal{C}}$ and we show that its influence on E is $O(1)$.

Velocity field $\boldsymbol{\chi}_1$. Using the formal asymptotic analysis of section 4, we have that

$$\boldsymbol{\chi}_1(x_1, x_2) \sim \frac{x_2}{h(x_1)}\mathbf{e}_1. \tag{5.3}$$

However, the right-hand side in (5.3) is not divergence free, so we correct (5.3) as

$$\boldsymbol{\chi}_1(x_1, x_2) = \nabla^\perp F(x_1, x_2), \quad \text{where } F(x_1, x_2) = -\frac{x_2^2}{2h(x_1)} - \frac{h(x_1)}{8} \tag{5.4}$$

and $\nabla^\perp = (-\partial/\partial x_2, \partial/\partial x_1)$. Then,

$$\boldsymbol{\chi}_1(x_1, x_2) = \frac{x_2}{h(x_1)}\mathbf{e}_1 + \frac{1}{2}\frac{dh(x_1)}{dx_1}\left(\frac{x_2^2}{h^2(x_1)} - \frac{1}{4}\right)\mathbf{e}_2, \quad \text{div}\boldsymbol{\chi}_1(x_1, x_2) = 0 \tag{5.5}$$

and, on the top/bottom parts of boundary $\partial\Pi^{ij}$, $\boldsymbol{\chi}_1(x_1, x_2 = \pm h(x_1)/2) = \pm\mathbf{e}_1/2$. The calculation of local rate of dissipation $W^{\chi_1}_{\Pi^{ij}}$ is now straightforward and the result coincides with (4.8).

Velocity field $\boldsymbol{\chi}_2$. The formal asymptotic analysis of section 4 gives

$$\boldsymbol{\chi}_2(x_1, x_2) \sim \frac{6x_1}{h(x_1)}\left(\frac{x_2^2}{h^2(x_1)} - \frac{1}{4}\right)\mathbf{e}_1 + \left[\frac{3x_2}{2h(x_1)} - 2\left(\frac{x_2}{h(x_1)}\right)^3\right]\mathbf{e}_2, \tag{5.6}$$

but, since h is, in truth, a function of x_1, (5.6) is not divergence free and cannot be used as such in the upper bound calculation. Instead, we define the trial field

$$\boldsymbol{\chi}_2(x_1, x_2) = \nabla^\perp F(x_1, x_2), \quad \text{where } F(x_1, x_2) = -\frac{2x_1x_2^3}{h^3(x_1)} + \frac{3x_1x_2}{2h(x_1)}. \tag{5.7}$$

The corresponding local dissipation rate $W^{\chi_2}_{\Pi^{ij}}$ is

$$W^{\chi_2}_{\Pi^{ij}} = \frac{3\pi\mu}{4}\left(\frac{a^{ij}}{\delta^{ij}}\right)^{\frac{3}{2}} + \frac{27\pi\mu}{10}\sqrt{\frac{a^{ij}}{\delta^{ij}}} + O(1) \tag{5.8}$$

and we note that it coincides, with leading order, with (4.11).

Velocity field $\boldsymbol{\lambda}$. We define a divergence free trial field $\boldsymbol{\lambda}$, which is approximately equal to (4.12), as

$$\boldsymbol{\lambda}(x_1, x_2) = \nabla^\perp F(x_1, x_2), \quad \text{where } F(x_1, x_2) = \left(\frac{3x_1^2x_2}{4h(x_1)} - \frac{x_1^2x_2^3}{h^3(x_1)}\right), \tag{5.9}$$

Then $\boldsymbol{\lambda}(x_1, x_2 = \pm h(x_1)/2) = \pm\frac{x_1}{2}\mathbf{e}_2$, and the corresponding local rate of dissipation $W^{\lambda}_{\Pi^{ij}}$ is given by (4.14).

Velocity field $\boldsymbol{\mathcal{C}}(\mathbf{x})$. We define trial field $\boldsymbol{\mathcal{C}}(\mathbf{x})$ as

$$\boldsymbol{\mathcal{C}}(x_1, x_2) = (\omega^{(i)} + \omega^{(j)})\nabla^\perp F(x_1, x_2) + (\omega^{(i)} - \omega^{(j)})\nabla^\perp G(x_1, x_2), \tag{5.10}$$

where

$$F(x_1,x_2) = \left(1-\frac{1}{2}\sqrt{1-\frac{x^2}{a_i^2}}-\frac{1}{2}\sqrt{1-\frac{x^2}{a_j^2}}\right)\left(\frac{x_2^2}{2h(x_1)}+\frac{h(x_1)}{8}\right)$$

$$+\int_0^{x_1}\frac{h(s)}{8}\frac{d}{ds}\left(\sqrt{1-\frac{s^2}{a_i^2}}+\sqrt{1-\frac{s^2}{a_i^2}}\right)ds$$

$$+\left(\frac{1}{2}\sqrt{1-\frac{x_1^2}{a_i^2}}-\frac{1}{2}\sqrt{1-\frac{x_1^2}{a_j^2}}\right)\left(-\frac{x_2}{2}+\frac{3h(x_1)x_2}{2}-\frac{2x_2^3}{h(x_1)}\right)$$

$$-\int_0^{x_1}\left(\frac{3h(s)x_2}{2h(x_1)}-\frac{2h(s)x_2^3}{h^3(x_1)}\right)\times\left[\left(\frac{1}{4}-\frac{h(s)}{2}\right)\left(\frac{s/a_i^2}{\sqrt{1-s^2/a_i^2}}-\frac{s/a_j^2}{\sqrt{1-s^2/a_j^2}}\right)\right.$$

$$\left.+\frac{dh(s)}{ds}\left(\sqrt{1-s^2/a_i^2}-\sqrt{1-s^2/a_j^2}\right)\right]ds$$

and

$$G(x_1,x_2) = -\left(\frac{1}{2}\sqrt{1-\frac{x_1^2}{a_i^2}}-\frac{1}{2}\sqrt{1-\frac{x_1^2}{a_j^2}}\right)\left(\frac{x_2^2}{2h(x_1)}-\frac{h(x_1)}{4}\right)$$

$$-\frac{1}{4}\int_0^{x_1}\frac{dh(s)}{ds}\frac{\sqrt{1-s^2/a_i^2}-\sqrt{1-s^2/a_j^2}}{2}ds$$

$$+\frac{x_2}{2}+\frac{\sqrt{1-x_1^2/a_i^2}+\sqrt{1-x_1^2/a_j^2}}{2}\left(-\frac{x_2}{2}+\frac{3h(x_1)x_2}{2}-\frac{2x_2^3}{h(x_1)}\right)$$

$$-\int_0^{x_1}\left(\frac{3h(s)x_2}{2h(x_1)}-\frac{2h(s)x_2^3}{h^3(x_1)}\right)\times\left[\left(\frac{1}{4}-\frac{h(s)}{2}\right)\left(\frac{s/a_i^2}{\sqrt{1-s^2/a_i^2}}+\frac{s/a_j^2}{\sqrt{1-s^2/a_j^2}}\right)\right.$$

$$\left.+\frac{dh(s)}{ds}\left(\sqrt{1-s^2/a_i^2}+\sqrt{1-s^2/a_j^2}\right)\right]ds.$$

Although the expression (5.10) is rather complicated, it can be checked with straightforward calculations (which we have done in MAPLE) that it satisfies boundary conditions (5.2) and, as such, it is an admissible trial field, which gives a local rate of dissipation $W^{\mathcal{C}}_{\Pi^{ij}} = O(1)$.

Finally, we find through MAPLE calculations that the contribution of the cross terms to $W_{\Pi^{ij}}$ is $O(1)$.

We have now defined a trial velocity field that satisfies the exact boundary conditions on the top and bottom boundaries of the gap Π^{ij}, is divergence free, and, most important, gives an upper bound on the gap dissipation rate which agrees, with leading order, with the formal asymptotic result of section 4.

5.2. Extension of the trial velocity field u outside the gaps between the particles in suspension. Let us denote the union of all gaps by U_Π and define the complement in Ω_F of the union of all gaps

$$U_E = \Omega_F \setminus U_\Pi. \tag{5.11}$$

We wish to extend the trial velocity field $\mathbf{u}$ from the gaps Π^{ij} to U_E, so that the leading order terms of the dissipation rate are not affected. Clearly, when there are many particles in the suspension, the set U_E is the union of many disjoint, connected components, which we denote by C_j. Let us then focus attention on one such component and drop the subscript j. To avoid boundary corners in the connected component C, we take a slightly larger domain $\tilde{C} \subset \Omega_F$ such that $C \subset \tilde{C}$ and $\partial\tilde{C}$ is smooth.[4] Note that the construction of section 5.1 gives a trial velocity of the form $\mathbf{u} = \nabla^\perp F$ and, since the gap thickness is $h = O(a) \gg \delta^{ij}$ at $\partial\Pi^{ij} \cap \partial\tilde{\mathcal{C}}$, the first and second derivatives of F are uniformly bounded on $\partial\tilde{\mathcal{C}}$, as $\delta \to 0$. We now wish to extend $\mathbf{u}$ to the interior of $\tilde{C}$.

Let us take a $\gamma > 0$, independent of δ, and define the boundary layer

$$C_\gamma = \{\mathbf{x} \in \tilde{C} \text{ such that } \operatorname{dist}(\mathbf{x}, \partial\tilde{C}) < \gamma\}. \tag{5.12}$$

Since the arcs in $\partial\tilde{C}$ are independent of δ, we can choose a cover $\mathcal{P}_j$, $j = 1, 2, \ldots, J$ independent of δ, and a subordinate partition of unity ϕ_j with support $\phi_j = \tilde{\mathcal{P}}_j \subset \mathcal{P}_j$ such that $\tilde{\mathcal{P}}_j \cap \tilde{C} \subset C_\gamma$. Then let us extend $\mathbf{u}$ in $\tilde{\mathcal{P}}_j$ and, for simplicity of notation, drop the index j.

In $\tilde{\mathcal{P}}$, define local coordinates $\mathbf{y} = (y_1, y_2)$, such that $y_2 = 0$ at $\partial\tilde{\mathcal{P}} \cap \partial\tilde{\mathcal{C}}$, and $\tilde{\mathcal{P}} \cap \tilde{\mathcal{C}}$ is mapped into a tensor product of intervals $I_1(y_1) \times I_2(y_2)$, for $y_2 > 0$. Take then a smooth function $g(y_2)$, which vanishes outside interval $I_2(y_2)$ and, at $y_2 = 0$, $g(0) = 1$, and define the extension of F, from $\tilde{\mathcal{P}} \cap \partial\tilde{\mathcal{C}}$ to $\tilde{\mathcal{P}} \cap \tilde{\mathcal{C}}$, as[5]

$$F(y_1, y_2) = g(y_2)\left[F(y_1, 0) + y_2 \frac{\partial F(y_1, 0)}{\partial y_2}\right]. \tag{5.13}$$

Clearly, the extended F is smooth and its first derivatives are equal to the previously specified values on $\tilde{\mathcal{P}} \cap \partial\tilde{\mathcal{C}}$. We also have

$$\|F(y_1, y_2)\|_{H^2(I_1 \times I_2)} \le A \tag{5.14}$$

with A, independent of δ. Repeating the same procedure, we extend F to all $\tilde{\mathcal{P}}_j \cap \tilde{\mathcal{C}}$, $j = 1, \ldots, J$, or, equivalently, to $\tilde{\mathcal{C}}$. Then, taking $\mathbf{u} = \nabla^\perp F$, we have div $\mathbf{u} = 0$, and the strain tensor $\mathcal{E}(\mathbf{u})$ with components in $L^2(\tilde{C})$ and a corresponding viscous dissipation rate

$$\int_{\tilde{C}} \mu\left(\mathcal{E}(\mathbf{u}), \mathcal{E}(\mathbf{u})\right) d\mathbf{x} \le A|\tilde{C}|. \tag{5.15}$$

We end this section with the remark that it is not necessary that $\tilde{C}$ lie inside Ω_F for estimate (5.15) to hold (see Figure 7). Indeed, even if the connected component C

[4] $\partial\tilde{C}$ is the union of arcs which lie either inside a gap Π^{ij} or on the boundary of a surrounding particle.

[5] Note that (5.13) is a simplified version of the classic Borel construction in [18, Theorem 1.2.6].

intersects the exterior boundary $\partial\Omega$, we can always extend F to a smooth H^2 function supported away from the corners of $\partial\Omega$, and (5.15) follows.

Gathering all the results in this section, we have, in the notation of section 3.2.1, the upper bound

(5.16)
$$
\begin{aligned}
E \le \min_{\mathbf{T},\omega} \sum_{i=1}^{N} \sum_{\substack{j\,\in\,\mathcal{N}_i \\ j\,<\,i}} & \left\{ \left[\frac{3\pi\mu}{4} \left(\frac{a^{ij}}{\delta^{ij}} \right)^{\frac{3}{2}} + \frac{27\pi\mu}{10} \sqrt{\frac{a^{ij}}{\delta^{ij}}} \right] [(\mathbf{T}^{(i)} - \mathbf{T}^{(j)}) \cdot \mathbf{q}^{ij}]^2 \right. \\
& + \frac{\pi\mu}{2} \sqrt{\frac{a^{ij}}{\delta^{ij}}} [(\mathbf{T}^{(i)} - \mathbf{T}^{(j)}) \cdot \mathbf{p}^{ij} + a_i \omega^{(i)} + a_j \omega^{j}]^2 \\
& \left. + \frac{9\pi\mu}{16} \sqrt{\frac{a^{ij}}{\delta^{ij}}} \left(\omega^{(i)} a^{ij} - \omega^{(j)} a^{ij} \right)^2 \right\} \\
& + \sum_{i \in \mathcal{B}} \left\{ \left[\frac{3\pi\mu}{4} \left(\frac{2a_i}{\delta^i} \right)^{\frac{3}{2}} + \frac{27\pi\mu}{10} \sqrt{\frac{2a_i}{\delta^i}} \right] [(\mathbf{T}^{(i)} - \mathbf{g}) \cdot \mathbf{q}^{i}]^2 \right. \\
& \qquad + \frac{9\pi\mu}{16} \sqrt{\frac{2a_i}{\delta^i}} (2\omega^{(i)})^2 (a^{ij})^2 \\
& \qquad \left. + \frac{\pi\mu}{2} \sqrt{\frac{2a_i}{\delta^i}} [(\mathbf{T}^{(i)} - \mathbf{g}) \cdot \mathbf{p}^{i} + a_i \omega^{(i)}]^2 \right\} + O(1),
\end{aligned}
$$

where, for the boundary nodes $i \in \mathcal{B}$, δ^i is the distance between $\partial D^{(i)}$ and the upper or lower boundary $\partial\Omega^{\pm}$.

6. Rigorous justification of the leading-order spring network approximation. In this section, we derive and justify rigorously the spring network approximation in two dimensions (recall section 3.2.1) by constructing a lower bound on E, which agrees with (5.16), to $O\big(\big(\frac{a}{\delta}\big)^{\frac{3}{2}}\big)$.

6.1. A simplified upper bound. Since the leading order term is not affected by the rotations of the particles, we set in (5.16) $\omega^{(i)} = 0$ for all $i = 1, \dots, N$, and we obtain a less precise but simplified upper bound

(6.1)
$$
\begin{aligned}
E \le W_{\Omega_F}(\mathbf{u}) = \min_{\mathbf{T}} \sum_{i=1}^{N} \sum_{\substack{j\,\in\,\mathcal{N}_i \\ j\,<\,i}} & \left\{ \left[\frac{3\pi\mu}{4} \left(\frac{a^{ij}}{\delta^{ij}} \right)^{\frac{3}{2}} + \frac{27\pi\mu}{10} \sqrt{\frac{a^{ij}}{\delta^{ij}}} \right] [(\mathbf{T}^{(i)} - \mathbf{T}^{(j)}) \cdot \mathbf{q}^{ij}]^2 \right. \\
& \left. + \frac{\pi\mu}{2} \sqrt{\frac{a^{ij}}{\delta^{ij}}} [(\mathbf{T}^{(i)} - \mathbf{T}^{(j)}) \cdot \mathbf{p}^{ij}]^2 \right\} + \sum_{i \in \mathcal{B}} \left\{ \left[\frac{3\pi\mu}{4} \left(\frac{2a_i}{\delta^i} \right)^{\frac{3}{2}} + \frac{27\pi\mu}{10} \sqrt{\frac{2a_i}{\delta^i}} \right] [(\mathbf{T}^{(i)} - \mathbf{g}) \cdot \mathbf{q}^{i}]^2 \right. \\
& \qquad \left. + \frac{\pi\mu}{2} \sqrt{\frac{2a_i}{\delta^i}} [(\mathbf{T}^{(i)} - \mathbf{g}) \cdot \mathbf{p}^{i}]^2 \right\} + O(1).
\end{aligned}
$$

Except for the $O(1)$ term, the right-hand side of (6.1) involves the minimization of a quadratic form in the translation velocities $\mathbf{T}^{(i)}$ for $i = 1, \dots, N$, and the minimum

is achieved by the solution of the linear system of equations,

$$\sum_{j\in\mathcal{N}_i}\left\{\left[\frac{3\pi\mu}{4}\left(\frac{a^{ij}}{\delta^{ij}}\right)^{\frac{3}{2}}+\frac{27\pi\mu}{10}\sqrt{\frac{a^{ij}}{\delta^{ij}}}\right][(\mathbf{T}^{(i)}-\mathbf{T}^{(j)})\cdot\mathbf{q}^{ij}]\mathbf{q}^{ij}\right.$$
$$\left.+\frac{\pi\mu}{2}\sqrt{\frac{a^{ij}}{\delta^{ij}}}[(\mathbf{T}^{(i)}-\mathbf{T}^{(j)})\cdot\mathbf{p}^{ij}]\mathbf{p}^{ij}\right\}+\mathbf{F}_{\mathcal{B}}(\mathbf{T}^{(i)})=\mathbf{0}\quad\text{for } 1\le i\le N, \tag{6.2}$$

where

$$\mathbf{F}_{\mathcal{B}}(\mathbf{T}^{(i)})$$
$$=\begin{cases}\left[\dfrac{3\pi\mu}{4}\left(\dfrac{2a_i}{\delta^i}\right)^{\frac{3}{2}}+\dfrac{27\pi\mu}{10}\sqrt{\dfrac{2a_i}{\delta^i}}\right][(\mathbf{T}^{(i)}-\mathbf{g})\cdot\mathbf{q}^i]\mathbf{q}^i+\dfrac{\pi\mu}{2}\sqrt{\dfrac{2a_i}{\delta^i}}[(\mathbf{T}^{(i)}-\mathbf{g})\cdot\mathbf{p}^i]\mathbf{p}^i & \text{if } i\in\mathcal{B},\\ \mathbf{0} \text{ otherwise.}\end{cases} \tag{6.3}$$

Next, we prove the unique solvability of these equations.

PROPOSITION 6.1. *The linear system of* (6.2) *has a unique solution,*

$$\boldsymbol{\tau}=(T_1^{(1)},T_2^{(1)},\dots,T_1^{(N)},\dots,T_2^{(N)})^T\in\mathbb{R}^{2N},$$

where superscript T stands for transpose.

Proof. Let us write the upper bound (6.1) in compact form as

$$E\le\min_{\boldsymbol{\tau}}\,(\boldsymbol{\tau}\cdot A\boldsymbol{\tau}-2\boldsymbol{\tau}\cdot\mathbf{f})+r+O(1), \tag{6.4}$$

where matrix $A\in\mathbb{R}^{2N\times 2N}$ is symmetric, $\mathbf{f}\in\mathbb{R}^{2N}$, and $r\in\mathbb{R}$. We prove the unique solvability of (6.2) (i.e., of $A\boldsymbol{\tau}=\mathbf{f}$) by showing that A is positive definite. Since we take the limit $\delta\to 0$, we have from (6.1) and (6.4) that

$$\begin{aligned}\boldsymbol{\tau}\cdot A\boldsymbol{\tau}&=\sum_{i=1}^{N}\sum_{\substack{j\in\mathcal{N}_i\\ j<i}}\left\{\left[\frac{3\pi\mu}{4}\left(\frac{a^{ij}}{\delta^{ij}}\right)^{\frac{3}{2}}+\frac{27\pi\mu}{10}\sqrt{\frac{a^{ij}}{\delta^{ij}}}\right][(\mathbf{T}^{(i)}-\mathbf{T}^{(j)})\cdot\mathbf{q}^{ij}]^2\right.\\ &\quad\left.+\frac{\pi\mu}{2}\sqrt{\frac{a^{ij}}{\delta^{ij}}}[(\mathbf{T}^{(i)}-\mathbf{T}^{(j)})\cdot\mathbf{p}^{ij}]^2\right\}\\ &\quad+\sum_{i\in\mathcal{B}}\left\{\left[\frac{3\pi\mu}{4}\left(\frac{2a_i}{\delta^i}\right)^{\frac{3}{2}}+\frac{27\pi\mu}{10}\sqrt{\frac{2a_i}{\delta^i}}\right](\mathbf{T}^{(i)}\cdot\mathbf{q}^i)^2+\frac{\pi\mu}{2}\sqrt{\frac{2a_i}{\delta^i}}(\mathbf{T}^{(i)}\cdot\mathbf{p}^i)^2\right\}\\ &\ge C\delta^{-\frac{1}{2}}\sum_{i=1}^{N}\sum_{\substack{j\in\mathcal{N}_i\\ j<i}}[(\mathbf{T}^{(i)}-\mathbf{T}^{(j)})\cdot\mathbf{q}^{ij}]^2+[(\mathbf{T}^{(i)}-\mathbf{T}^{(j)})\cdot\mathbf{p}^{ij}]^2\\ &\quad+C\delta^{-\frac{1}{2}}\sum_{i\in\mathcal{B}}[(\mathbf{T}^{(i)}\cdot\mathbf{q}^i)^2+(\mathbf{T}^{(i)}\cdot\mathbf{p}^i)^2]\\ &=C\delta^{-\frac{1}{2}}\left(\sum_{i=1}^{N}\sum_{\substack{j\in\mathcal{N}_i\\ j<i}}|\,\mathbf{T}^{(i)}-\mathbf{T}^{(j)}\,|^2+\sum_{i\in\mathcal{B}}|\,\mathbf{T}^{(i)}\,|^2\right)=\boldsymbol{\tau}\cdot\tilde{A}\boldsymbol{\tau},\end{aligned} \tag{6.5}$$

where C is independent of δ and matrix $\tilde{A}$ is clearly symmetric, nonnegative definite. To show that $\tilde{A}$ is, in fact, positive definite, let us suppose that there exists a nontrivial $\boldsymbol{\tau}$ in the null space of $\tilde{A}$. Then, by (6.5), we have $\mathbf{T}^{(i)} - \mathbf{T}^{(j)} = 0$ for $i = 1, \dots, N$, $j \in \mathcal{N}_i$, and $\mathbf{T}^{(i)} = 0$ for $i \in \mathcal{B}$. Since the graph Γ is connected (Property 3.1), this implies $\mathbf{T}^{(i)} = 0$ for all $i = 1, \dots, N$ or, equivalently, $\boldsymbol{\tau} = \mathbf{0}$. However, this contradicts our initial assumption on $\boldsymbol{\tau}$, so the null space of $\tilde{A}$ must be trivial. This implies, in turn, that A is positive definite and that the linear system of (6.2) is uniquely solvable. ∎

Remark 6.1. In the remainder of this paper, we denote by $\mathbf{u}$ the trial velocity field constructed in section 5, where all rotational velocities $\omega^{(i)}$ are set to zero and where translational velocities $\mathbf{T}^{(i)}$ solve linear system of equations (6.2) for $1 \le i \le N$. In particular, in gap Π^{ij}, connecting adjacent disks $D^{(i)}$ and $D^{(j)}$, the trial field is

$$\mathbf{u}(\mathbf{x}) = [(\mathbf{T}^{(i)} - \mathbf{T}^{(j)}) \cdot \mathbf{e}_1]\boldsymbol{\chi}_1(\mathbf{x}) + [(\mathbf{T}^{(i)} - \mathbf{T}^{(j)}) \cdot \mathbf{e}_2]\boldsymbol{\chi}_2(\mathbf{x}) + \frac{1}{2}(\mathbf{T}^{(i)} + \mathbf{T}^{(j)}), \tag{6.6}$$

where $\boldsymbol{\chi}_1$ and $\boldsymbol{\chi}_2$ are given by (5.5) and (5.7), respectively. Note that this trial field yields upper bound (see (6.1))

$$E \le W_{\Omega_F}(\mathbf{u}) \tag{6.7}$$

when used in variational principle (2.20).

6.2. Lower Bound.

6.2.1. Outline of the construction. Given a subdomain M of Ω_F, define a functional

$$W_M^*(\mathcal{S}) = \int_{\partial\Omega \cap \bar{M}} \mathbf{g} \cdot \mathcal{S}\mathbf{n} ds - \int_M F(\mathcal{S}) d\mathbf{x}, \tag{6.8}$$

where $\bar{M}$ is the closure of M, $\mathbf{g}$ is defined by (2.9),

$$F(\mathcal{S}) = \frac{1}{4\mu}\left[(\mathcal{S}, \mathcal{S}) - \frac{1}{2}(\text{trace } \mathcal{S})^2\right], \tag{6.9}$$

and $\mathcal{S}$ is a symmetric (stress) tensor in $\mathcal{F}$. In the context of this paper, subdomain M stands for either a gap Π^{ij} between adjacent particles or a connected component $\mathcal{C}$ in U_E, where the flow is diffuse (see section 5.2). Then $W_{\Omega_F}^*$ is given by the sum of $W_M^*(\mathcal{S})$ for all such disjoint subdomains in Ω_F. For any $\mathcal{S} \in \mathcal{F}$, we have by dual variational principle (2.22)

$$W_{\Omega_F}^\star(\mathcal{S}) \le E \le W_{\Omega_F}(\mathbf{u}). \tag{6.10}$$

Our goal in this section is to construct a trial tensor $\mathcal{S} \in \mathcal{F}$ such that $W_{\Omega_F}^\star(\mathcal{S})$ matches leading order upper bound $W_{\Omega_F}(\mathbf{u})$.

The construction of the trial tensor $\mathcal{S}$ proceeds as follows.

Step 1. In an ideal case, where $\hat{\mathbf{u}}$ and $\hat{\mathcal{S}}$, the minimizer and maximizer of the direct and dual problems (2.20) and (2.22), respectively, would be known, the constitutive equations for the incompressible fluid would give

$$\int_M F(\hat{\mathcal{S}}) d\mathbf{x} = W_M(\hat{\mathbf{u}}) = \mu \int_M (\mathcal{E}(\mathbf{u}), \mathcal{E}(\mathbf{u}))\, d\mathbf{x}, \tag{6.11}$$

and, by integration by parts,

$$\int_{\partial\Omega\cap\bar{M}} \mathbf{g}\cdot\hat{\mathcal{S}}\mathbf{n}\,ds = 2W_M(\hat{\mathbf{u}}). \tag{6.12}$$

However, we don't know $\hat{\mathbf{u}}$, so we use instead trial velocity field $\mathbf{u}$ described in Remark 6.1. With this $\mathbf{u}$, we find, as a first step in our construction, an approximate pressure P and the corresponding approximate stress tensor $\mathcal{S}_0 = 2\mu\mathcal{E}(\mathbf{u}) - P\mathcal{I}$.

For this purpose, let us focus on a gap Π^{ij}, where $\mathcal{S}_0$ satisfies

$$\int_{\Pi^{ij}} F(\mathcal{S}_0)d\mathbf{x} = W_{\Pi^{ij}}(\mathbf{u}) = O\left(\left(\frac{a^{ij}}{\delta^{ij}}\right)^{\frac{3}{2}}\right). \tag{6.13}$$

Note, however, that $\mathcal{S}_0 \notin \mathcal{F}$ because $\mathrm{div}\mathcal{S}_0 \neq \mathbf{0}$, so we define the trial tensor in Π^{ij} as

$$\mathcal{S} = \mathcal{S}_0 - \mathcal{K},$$

where $\mathcal{K}$ is a compensating tensor chosen such that $\mathrm{div}\,(\mathcal{S}_0 - \mathcal{K}) = 0$ in Π, and

$$\int_{\Pi^{ij}} F(\mathcal{S})d\mathbf{x} = W_{\Pi^{ij}}(\mathbf{u}) + O\left(\sqrt{\frac{a^{ij}}{\delta^{ij}}}\right). \tag{6.14}$$

Step 2. This is the crucial step in the construction of the lower bound. In Step 1, we obtained tensor $\mathcal{S}(\mathbf{x})$ in Π^{ij} and, in particular, on the portion of $\partial D^{(j)}$ which belongs to the neck Π^{ij}. In the second step, we extend $\mathcal{S}$ to the remaining parts of $\partial D^{(j)}$, so that the net force and torque conditions (2.12) hold. Such an extension cannot be constructed for each $\partial D^{(j)}$ individually. Recall that U_Π is the union of all gaps. For each connected component $\mathcal{C}$ of the set $U_E = \Omega_F\backslash U_\Pi$, where the flow is diffuse, we must have

$$\int_{\partial\mathcal{C}} \mathcal{S}\mathbf{n}\,ds = \mathbf{0} \tag{6.15}$$

for any divergence free extension of $\mathcal{S}$, from Π to $\mathcal{C}$. But, since each $\partial\mathcal{C}$ contains parts of the boundaries of several neighboring disks, the extensions of $\mathcal{S}$ to the boundaries of these disks must be coupled. An attempt to satisfy the balance of forces and torques (2.12) for an individual disk $D^{(j)}$ influences the balance on all neighboring disks. Since these disks have other neighbors as well (recall that the graph Γ is connected), the extension of $\mathcal{S}$ from the necks Π^{ij} to the remaining parts of $\partial D^{(j)}$, for $1 \le j \le N$, is a global problem.

Note that a similar difficulty arises in the scalar problem of electrical conduction [5], where a simple construction of the dual trial field (which is a vector flux) is given as follows. In a gap Π^{ij}, the dual trial field is taken as the vector $\mathbf{j} = (0, \zeta(x_1))$, where ζ is a smooth function of x_1, the local coordinate in the direction orthogonal to the axis of symmetry of the gap. Outside the union of all gaps, the dual trial field is extended to $\mathbf{0}$. While this choice satisfies the divergence-free condition locally in each subdomain of Ω_F, one must ensure that the total flux through $\partial D^{(j)}$ intersected with the union of all the gaps connected with $D^{(j)}$ is zero for all $1 \le j \le N$ as well. The latter condition is satisfied in [5] by setting

$$\int_{\partial D^{(i)}\cap\Pi^{ij}} \mathbf{j}\cdot\mathbf{n}^{(i)}\,ds = J_{\Pi^{ij}},$$

where J_Π is the net current flowing through the corresponding branch of the asymptotic network (graph Γ). More explicitly, the condition of flux balance at $\partial D^{(j)}$ is formulated as Kirchhoff's current law at the node $\mathbf{x}^{(j)}$ in the asymptotic network.

While in the scalar problem, the two conditions (divergence free and flux balance) on the dual trial field can be dealt with separately in the vectorial problem that we consider here, they appear to be coupled, and one cannot simply generalize the construction in [5] to find an admissible $\mathcal{S} \in \mathcal{F}$. We introduce in section 6.2.3 our novel construction of the extension of $\mathcal{S}$, which is divergence free and satisfies the momentum balance equations for all disks.

Step 3. Extend the tensor $\mathcal{S}$, defined so far in the gaps and at $\partial D^{(j)}$ for $1 \leq j \leq N$, to the whole Ω_F. The main point of this step is to control the energy of the extension in such a way that

$$W^*_{\Omega_F \setminus U_\Pi}(\mathcal{S}) \ll O(\delta^{-\frac{3}{2}}). \tag{6.16}$$

Step 4. In this step we gather all the results of the previous steps and show that $W_{\Omega_F}(\mathbf{u})$ and $W^\star_{\Omega_F}(\mathcal{S})$ are the same leading order.

6.2.2. The trial field $\mathcal{S}$ in a gap. We begin our construction of $\mathcal{S}$ in a gap Π^{ij}, with the help of velocity field (6.6). Recall from sections 4 and 5 that (6.6) is divergence free and, furthermore, it is an approximate solution of Stokes's equations in the following sense: if the gap thickness h were a constant, we would have curl $\triangle \mathbf{u} = 0$, the pressure would be well defined by $\mu \triangle u = \nabla P$, and the stress

$$\mathcal{S}_0 = 2\mu \mathcal{E}(\mathbf{u}) - P\mathcal{I}$$

would be divergence free. However, in truth, gap Π^{ij} is not flat and the condition $\mathrm{div}\mathcal{S} = 0$ that any dual trial field $\mathcal{S}$ must satisfy needs to be ensured for the variable thickness $h(x_1)$. In that case, Δu is not a gradient of a scalar function, so we introduce an approximate pressure P and a compensating symmetric tensor $\mathcal{K}$ such that

$$\mathcal{S} = \mathcal{S}_0 - \mathcal{K} \tag{6.17}$$

is divergence free. Because $\mathrm{div}\,\mathbf{u} = 0$, we have

$$\begin{aligned} F(\mathcal{S}_0) &= \frac{1}{4\mu}\left[(\mathcal{S}_0, \mathcal{S}_0) - \frac{1}{2}(\mathrm{trace}\ \mathcal{S}_0)^2\right] \\ &= \frac{1}{4\mu}\left[(2\mu\mathcal{E}(\mathbf{u}) - P\mathcal{I}, 2\mu\mathcal{E}(\mathbf{u}) - P\mathcal{I}) - 2P^2\right] = \mu\left(\mathcal{E}(\mathbf{u}), \mathcal{E}(\mathbf{u})\right) \end{aligned}$$

and

$$\int_{\Pi^{ij}} F(\mathcal{S}_0) d\mathbf{x} = W_{\Pi^{ij}}(\mathbf{u}) = O\left(\frac{a^{ij}}{\delta^{ij}}^{\frac{3}{2}}\right), \tag{6.18}$$

so to get a lower bound that matches the upper one to leading order, we wish that

$$\int_{\Pi^{ij}} \left[F(\mathcal{S}) - F(\mathcal{S}_0)\right] d\mathbf{x} = O\left(\sqrt{\frac{a^{ij}}{\delta^{ij}}}\right). \tag{6.19}$$

This can be accomplished, for example, by choosing P and $\mathcal{K}$ to satisfy

$$\int_{\Pi^{ij}} (\mathcal{S}_0, \mathcal{K})\, d\mathbf{x} = O\left(\sqrt{\frac{a^{ij}}{\delta^{ij}}}\right) \text{ and } \int_{\Pi^{ij}} (\mathcal{K}, \mathcal{K})\, d\mathbf{x} = O\left(\sqrt{\frac{a^{ij}}{\delta^{ij}}}\right) \tag{6.20}$$

since

$$F(\mathcal{S}) - F(\mathcal{S}_0) = -2\,(\mathcal{S}_0, \mathcal{K}) + (\mathcal{K}, \mathcal{K}) + \text{trace } \mathcal{S}_0 \text{ trace } \mathcal{K} - \frac{1}{2}(\text{trace } \mathcal{K})^2.$$

Let us then begin our search for $\mathcal{K}$ by rewriting equation $\text{div}\mathcal{S} = \mathbf{0}$, in terms of the components of $\mathcal{K}$, as

$$\begin{aligned} \partial_{x_1}\mathcal{K}_{11} + \partial_{x_2}\mathcal{K}_{12} &= R_1, \\ \partial_{x_1}\mathcal{K}_{12} + \partial_{x_2}\mathcal{K}_{22} &= R_2, \end{aligned} \tag{6.21}$$

where the discrepancy vector

$$\mathbf{R} = \text{div } \mathcal{S}_0 = \mu\triangle\mathbf{u} - \nabla P \tag{6.22}$$

depends on the choice of P. We define the approximate pressure by

$$P(\mathbf{x}) = \mu\int_{-h/2}^{x_2} \triangle u_2(s_1, s_2)ds_2 + \mu\int_{-R^{ij}}^{x_1} r_1(s_1)ds_1, \tag{6.23}$$

where $r_1(x_1)$ is given in terms of the Laplacian of the first component of $\mathbf{u}$ as

$$\Delta u_1(x_1, x_2) = r_1(x_1) + r_2(x_1, x_2). \tag{6.24}$$

Then we set the first entry $\mathcal{K}_{11}$ in the compensating tensor to zero, and we find from (6.21) that

$$\mathcal{K}_{12}(\mathbf{x}) = \int_{-h/2}^{x_2} R_1(s_1, s_2)ds_2, \qquad \mathcal{K}_{22}(\mathbf{x}) = -\int_{-R^{ij}}^{x_1} R_1(s_1, s_2)ds_2 \tag{6.25}$$

for discrepancy vector

$$\mathbf{R}(\mathbf{x}) = \mu\triangle\mathbf{u}(\mathbf{x}) - \nabla P(\mathbf{x}) = \begin{pmatrix} \mu\triangle u_1(\mathbf{x}) - \partial_{x_1}P(\mathbf{x}) \\ 0 \end{pmatrix}. \tag{6.26}$$

Now, to verify that estimates (6.20) hold, we note that the components of $\Delta\mathbf{u}$ are sums of terms of the form

$$\text{const } \frac{x_1^k x_2^l}{h(x_1)^m} \tag{6.27}$$

for some nonnegative integers k, l, m, and that we have the following estimate.

LEMMA 6.1. *For k even, there exists a positive constant c such that*

$$\int_{\Pi^{ij}} \frac{x_1^k x_2^l}{h^m} d\mathbf{x} \le c\int_{-R^{ij}}^{R^{ij}} h^{\frac{k}{2}+l+1-m} dx_1. \tag{6.28}$$

If k is odd, then

$$\int_{\Pi^{ij}} \frac{x_1^k x_2^l}{h^m} d\mathbf{x} = 0. \tag{6.29}$$

Moreover, for any positive integer p, we have

$$\int_{-R^{ij}}^{R^{ij}} \frac{dx_1}{(\delta^{ij} + x_1^2/a^{ij})^p} = O\left(\left(\frac{a^{ij}}{\delta^{ij}}\right)^{p-\frac{1}{2}}\right). \tag{6.30}$$

Proof. To prove (6.30), we write $\int_{-R^{ij}}^{R^{ij}}(\delta^{ij}+x_1^2/a^{ij})^{-p}dx_1 = \mathcal{I}_1+\mathcal{I}_2$, where

$$\mathcal{I}_1 = \int_{-\sqrt{\delta^{ij}}}^{\sqrt{\delta^{ij}}} \frac{dx_1}{(\delta^{ij}+x_1^2/a^{ij})^p}.$$

Scaling x_1 by $\sqrt{\delta^{ij}}$, we get

$$I_1 = (\delta^{ij})^{1/2-p}\int_{-1}^{1}\frac{dt}{(1+t^2/a^{ij})^p} = c_1(p,a^{ij})(\delta^{ij})^{1/2-p},$$

where c_1 is independent of δ^{ij}. For $\mathcal{I}_2$, we have

$$\begin{aligned}\mathcal{I}_2 &= 2\int_{\sqrt{\delta^{ij}}}^{R^{ij}}\frac{dx_1}{(\delta^{ij}+x_1^2/a^{ij})^p} \le 2\int_{\sqrt{\delta^{ij}}}^{R^{ij}}\left(\frac{a^{ij}}{x_1^2}\right)^p dx_1\\ &= \frac{2}{2p-1}[(\delta ij)^{\frac{1}{2}-p}-(R^{ij})^{1-2p}] \le c_2(p)(\delta^{ij})^{1/2-p}\end{aligned}$$

and the proof of (6.30) is complete. Identity (6.29) follows immediately because the integrand is an odd function of x_1. Finally, (6.29) and $x_1^2/a^{ij} < h(x_1)$ imply (6.28). ∎

In light of Lemma 6.1, we obtain with explicit calculations that (6.20) holds and, therefore,

$$\int_{\Pi^{ij}} F(\mathcal{S})d\mathbf{x} = W_{\Pi^{ij}}(\mathbf{u}) + O\left(\sqrt{\frac{a^{ij}}{\delta^{ij}}}\right). \tag{6.31}$$

In the next section, we extend $\mathcal{S}$ from Π^{ij} to $\partial D^{(i)}$ and $\partial D^{(j)}$ in such a way that the net force and torque on $D^{(i)}$ and $D^{(j)}$ vanish. For that purpose, we need to examine the integrals of $\mathcal{S}\mathbf{n}$ over various parts of $\partial\Pi^{ij}$. We show that, roughly speaking, the integrals of $\mathcal{S}\mathbf{n}$ over opposite sides of $\partial\Pi^{ij}$ cancel each other. To make this precise, let us denote the lateral parts of $\partial\Pi^{ij}$ by

$$L_\pm = \left\{(x_1,x_2) : x_1 = \pm R^{ij}, -\frac{1}{2}h(R^{ij})) < x_2 < \frac{1}{2}h(R^{ij})\right\}.$$

Proposition 6.2.

$$\int_{L_+}\mathcal{S}\mathbf{n}\,ds + \int_{L_-}\mathcal{S}\mathbf{n}\,ds = 0.$$

Proof. Since $\mathcal{S}\mathbf{n} = (\mathcal{S}_{11}(\pm R^{ij},x_2),\mathcal{S}_{12}(\pm R^{ij},x_2))^T$ on $L_\pm$, we must show that

$$\int_{-\frac{h(R^{ij})}{2}}^{\frac{h(R^{ij})}{2}}\mathcal{S}_{1k}(-R^{ij},x_2)dx_2 = \int_{-\frac{h(R^{ij})}{2}}^{\frac{h(R^{ij})}{2}}\mathcal{S}_{1k}(R^{ij},x_2)dx_2 \quad \text{for } k=1,2. \tag{6.32}$$

This, in turn, follows by direct calculation from the expression of trial stress field $\mathcal{S}$ constructed above. ∎

Remark 6.2. Since $\operatorname{div}\mathcal{S}=0$ in Π^{ij}, $\int_{\partial\Pi^{ij}}\mathcal{S}\mathbf{n}\,ds = 0$ and, by Proposition 6.2, we have for the top and bottom parts of $\partial\Pi^{ij}$,

$$\int_{\partial D^{(i)}\cap\overline{\Pi^{ij}}}\mathcal{S}\mathbf{n}^{(i)}\,ds = -\int_{\partial D^{(j)}\cap\overline{\Pi^{ij}}}\mathcal{S}\mathbf{n}^{(j)}\,ds. \tag{6.33}$$

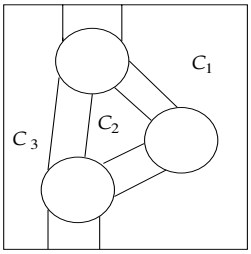

FIG. 7. *Three-disk network. Connected components.*

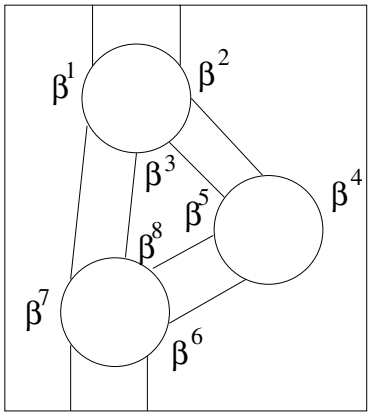

FIG. 8. *Three-disk network. Designation of the vector integrals* $\boldsymbol{\beta}^j$.

6.2.3. Extension of $\mathcal{S}$ to the boundaries of the disks. In section 6.2.2, we defined $\mathcal{S}$ in Π^{ij} and, in particular, on $\partial D^{(j)} \cap \overline{\Pi^{ij}}$. Here, we wish to extend $\mathcal{S}$ to the whole boundary $\partial D^{(j)}$ in such a way that

$$\int_{\partial D^{(j)}} \mathcal{S}\mathbf{n}^{(j)}\, ds = \mathbf{0} \quad \text{and} \quad \int_{\partial \mathcal{C}} \mathcal{S}\mathbf{n}\, ds = \mathbf{0} \tag{6.34}$$

for any connected component $\mathcal{C}$ of diffuse flow in $U_E = \Omega_F \setminus U_\Pi$ and for all $j = 1, \ldots, N$. We note that $\partial D^{(j)} \cap \overline{U}_E$ is a union of circular, complementary arcs, and we let vectors $\boldsymbol{\beta}^k$ denote the unknown integrals of $\mathcal{S}\mathbf{n}^{(j)}$ over various parts of $\partial D^{(j)} \cap \overline{U}_E$ for $1 \le j \le N$. We begin by showing that there exist vectors $\boldsymbol{\beta}^k$ consistent with (6.34). This is done first for a simple, three-disk network and is generalized later to N disks. Then we construct $\mathcal{S}$ on $\partial D^{(j)} \cap \overline{U}_E$ for $1 \le j \le N$ so the integral of $\mathcal{S}\mathbf{n}^{(j)}$ over the kth complementary arc is equal to $\boldsymbol{\beta}^k$ for all k.

Part I: A simple, three-disk network. To simplify the presentation, let us begin by considering a simple three-disk network, as shown in Figure 7, where there are three connected regions $\mathcal{C}_1$, $\mathcal{C}_2$, and $\mathcal{C}_3$ of diffuse flow.

The unknown integrals of $\mathcal{S}\mathbf{n}$ over the complementary arcs in $\partial D^{(j)} \cap \overline{U}_E$ for $j = 1, 2, 3$ are denoted by $\boldsymbol{\beta}^k$, $1 \le k \le 8$ (see Figure 8). We also let $\mathbf{F}^k$, and $\mathbf{B}^k$ for $1 \le k \le 5$ be the known integrals of $\mathcal{S}\mathbf{n}$ over the parts of $\partial D^{(j)} \cap \bar{U}_\Pi$ and the lateral segments of the gaps, respectively (see Figure 9 and recall Proposition 6.2). Finally, for connected components $\mathcal{C}_1$ and $\mathcal{C}_3$, we need $\mathcal{S}\mathbf{n}$ on the exterior boundaries $\partial\Omega \setminus \overline{U_\Pi}$ of the domain. On the vertical segments of the external boundary, we set $\mathcal{S} = 0$,

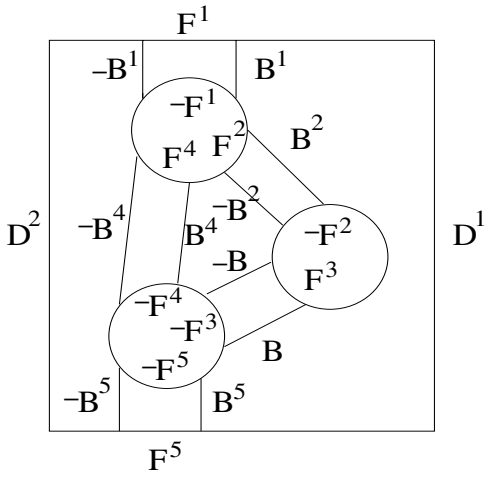

FIG. 9. *Three-disk network. Right-hand sides.*

and on the horizontal segments, we let $\mathcal{S}$ be constant. Letting $\mathbf{D}^1$ and $\mathbf{D}^2$ be the net traction over $\partial\mathcal{C}_3 \cap \partial\Omega$ and $\partial\mathcal{C}_1 \cap \partial\Omega$, respectively, we can now write (6.34) as

$$\mathbf{F}^1 + \mathbf{D}^1 + \mathbf{F}^5 + \mathbf{D}^2 = 0, \tag{6.35}$$

$$\begin{array}{llll} \boldsymbol{\beta}^1+ \ \boldsymbol{\beta}^2+ \ \boldsymbol{\beta}^3 = \ \mathbf{Q}^1, & & \mathbf{Q}^1 = -\mathbf{F}^2 \ -\mathbf{F}^4 \ +\mathbf{F}^1, \\ \boldsymbol{\beta}^4+ \ \boldsymbol{\beta}^5 = \ \mathbf{Q}^2, & \text{where} & \mathbf{Q}^2 = -\mathbf{F}^3 \ +\mathbf{F}^2, \\ \boldsymbol{\beta}^6+ \ \boldsymbol{\beta}^7+ \ \boldsymbol{\beta}^8 = \ \mathbf{Q}^3, & & \mathbf{Q}^3 = \ \mathbf{F}^3 \ +\mathbf{F}^4 \ +\mathbf{F}^5, \end{array} \tag{6.36}$$

and

$$\begin{array}{llll} \boldsymbol{\beta}^2+ \ \boldsymbol{\beta}^4+ \ \boldsymbol{\beta}^6 = \ \mathbf{Q}^4, & & \mathbf{Q}^4 = -\mathbf{B}^1 \ -\mathbf{B}^2 \ -\mathbf{B}^3 \ -\mathbf{B}^5 \ -\mathbf{D}^1, \\ \boldsymbol{\beta}^3+ \ \boldsymbol{\beta}^5+ \ \boldsymbol{\beta}^8 = \ \mathbf{Q}^5, & \text{where} & \mathbf{Q}^5 = -\mathbf{B}^4 \ +\mathbf{B}^2 \ +\mathbf{B}^3, \\ \boldsymbol{\beta}^1+ \ \boldsymbol{\beta}^7 = \ \mathbf{Q}^6, & & \mathbf{Q}^6 = \ \mathbf{B}^1 \ +\mathbf{B}^4 \ +\mathbf{B}^5 \ -\mathbf{D}^2. \end{array} \tag{6.37}$$

We now have an undetermined system of six vectorial equations (6.36), (6.37), with eight unknown vectors $\boldsymbol{\beta}^k$ for $1 \le k \le 8$ with right-hand sides satisfying constraint (6.35).

PROPOSITION 6.3. *There exist solutions of the linear system of* (6.36), (6.37).

Proof. Note that the vector system (6.36), (6.37) is equivalent to two scalar systems with the same matrix for the components of $\boldsymbol{\beta}^j$. It is therefore sufficient to prove the proposition for any one of the two scalar systems. The matrix A is

$$A = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}, \tag{6.38}$$

and we denote its rows by $\mathbf{a}_j$ for $1 \le j \le 6$. We call the first three rows in A disk-rows or simply d-rows, and the last three rows c-rows (in reference to the connected

components $\mathcal{C}_j$). We observe that each arc in $\partial D^{(j)} \cap \overline{U_E}$ belongs to exactly one disk and one connected component and thus exactly two entries in each column are equal to 1. Moreover, matrix A possesses the following two-ones property: one of these unit entries appears in a d-row and another appears in a c-row.

Next, we show that each equation in (6.36), (6.37) is a linear combination of the other five. Indeed, summing up first the equations in (6.36) and then those in (6.37), we obtain

$$\sum_{j=1}^{8} \boldsymbol{\beta}^j = \mathbf{F}_1 + \mathbf{F}_5 \quad \text{and} \quad \sum_{j=1}^{8} \boldsymbol{\beta}^j = -(\mathbf{D}_1 + \mathbf{D}_2),$$

respectively. These equations are consistent by (6.35), which gives $\mathbf{F}_1 + \mathbf{F}_5 = -(\mathbf{D}_1 + \mathbf{D}_2)$, and the rows of A are clearly linearly dependent.

Let us then eliminate from the original system of equations one equation, say, the first one, and show that the reduced system is solvable. Let A_R be the matrix of the reduced system. The rows of A_R are $\mathbf{a}_j$ with $2 \le j \le 6$. We show next that the rows of A_R are linearly independent (i.e., $\text{rank} A_R = 5$), and then that the existence of solutions follows from standard linear algebra.

Arguing by contradiction, suppose that there exists a k, between 2 and 6, such that $\mathbf{a}_k$ is a linear combination of $\mathbf{a}_p$ for $2 \le p \le 6$, $p \ne k$. Explicitly, we have

$$\mathbf{a}_k = \sum_{m=2, m\ne k}^{6} \lambda_m \mathbf{a}_m, \tag{6.39}$$

where not all λ_m are zero. By the two-ones property, three rows $\mathbf{a}_4, \mathbf{a}_5, \mathbf{a}_6$ have a unit entry at a column where all other rows of A_R have zeros. Take, for example, row $\mathbf{a}_4$. It has a unit entry in column 2, whereas the other four remaining rows $\mathbf{a}_j, 2 \le j \le 6, j \ne 4$, have zeros in this column. Hence, (6.39) implies that $k \ne 4$. The same argument shows that $k \ne 5$, $k \ne 6$. When $k = 2$ or 3, direct inspection of the first column of A_R shows that λ_6 from (6.39) is zero. Similarly we obtain $\lambda_4 = \lambda_5 = 0$. Then (6.39) reduces to $\mathbf{a}_2 = \lambda \mathbf{a}_3$, which is impossible since $\mathbf{a}_2$ and $\mathbf{a}_3$ are linearly independent. ∎

Part II: A general, N disk network with M connected components. Analogous to (6.36), we write the momentum balance equations on the boundary of each disk $D^{(j)}$ for $1 \le j \le N$. We call these equations d-equations. Furthermore, analogous to (6.37), we write the equations for each connected component $\mathcal{C}_p$, $1 \le p \le M$, where $M \ge N$. These are referred to as c-equations. As above, we consider the scalar system of $N + M$ equations for the components of unknown vectors $\boldsymbol{\beta}^p$, $1 \le p \le P$ ($P = 8$ in the example with three disks). This linear system is referred to as the d-c-system. Similar to the case of three disks, the right-hand side of the system involves integrals of $\mathcal{S}\mathbf{n}$ over parts of $\partial\Omega$ which do not belong to gaps. We assume that $\mathcal{S}$ is extended to the external boundary $\partial\Omega$ so that condition $\int_{\partial\Omega} \mathcal{S}\mathbf{n}\, ds = \mathbf{0}$ holds.

The solvability of the d-c-system is determined by matrix A, which has $M + N$ rows $\mathbf{a}_i$, $i = 1, \dots, M + N$, and P columns. The rows of A that correspond to d-equations are called d-rows, and those remaining are called c-rows. The entries of A are again either 0 or 1. Since each complementary arc in $\partial D^{(j)} \cap \overline{U_E}$ belongs to exactly one disk and one connected component, we observe that in each column of A, exactly two entries are equal to 1. One of these entries appears in a d-row and the other in a c-row. (A has the two-ones property.)

In what follows, we recall from section 3.1 that network Γ is a Delaunay graph corresponding to a Voronoi tessellation of Ω. We restrict our attention to the case of large N (for technical reasons it is sufficient to have $N \geq 3$) and consider only Voronoi tessellations with at least one Voronoi cell being strictly inside Ω. We also make use of Properties 3.1 to 3.3 of Γ.

THEOREM 6.1. *The d-c-system has a solution.*

Proof. First, we show that the d-c-system is underdetermined (i.e., $P > M + N$). Indeed, by Property 3.2, at least two edges of Γ originate from each interior vertex (which is the center of some disk $D^{(i)}$). Then $P \geq 2N$. Next, by Property 3.3, there exists a closed path which consists of interior edges. Therefore, there exists a connected component $\mathcal{C}_j$ with its closure disjoint from $\partial\Omega$ and, as such, there are at least three edges and three arcs in $\partial\mathcal{C}_j$. If the connected component would contain parts of $\partial\Omega$, there would be at least two arcs in its boundary. Thus[6] $2M < P$ and, since $P \geq 2N$, $P > M + N$. ∎

Next, we show that matrix A of the d-c-system has linearly dependent rows. Indeed, similar to the case of three disks, we have that the sum of the d-equations is equal to the sum of k-equations. Then we eliminate the first equation in the d-c-system and we denote by A_R the reduced $(M + N - 1) \times P$ matrix. To finish the proof of the theorem, we now show that the reduced system is full rank.

LEMMA 6.2. *The rank of A_R is $M + N - 1$.*

Proof. We argue by contradiction. Assume that the rows of A_R are linearly dependent, that is, for some $k > 1$,

$$\mathbf{a}_k = \sum_{m \neq k, m=2}^{M+N} \lambda_m \mathbf{a}_m, \tag{6.40}$$

where at least one λ_m is nonzero. The strategy of the proof is as follows. We introduce a multistep procedure where on each consecutive step l we have a set X_l of d-rows and a set Y_l of c-rows. We show that the rows from $X_l \cup Y_l$ cannot appear on the left-hand side of (6.40). Furthermore, we show that if either of these rows are present in the right-hand side of (6.40), then the coefficients λ_m in front of these rows in (6.40) must be zero. The process is stopped after L steps, when either all d-rows are included in $\cup_{l=1}^{L} X_l$ or all c-rows belong to $\cup_{l=1}^{L} Y_l$. At that point, (6.40) contains only d-rows (or only c-rows). Then the lemma follows from the linear independence of the d-rows and (c-rows), respectively.

Before giving the multistep procedure, let us introduce some notation. Given a collection of disks $S = \{D^{(i_1)}, D^{(i_2)}, \ldots, D^{(i_k)}\}$, denote by $\mathcal{C}(S)$ the set of all connected components of $\Omega_F \setminus U_\Pi$ adjacent to a disk in S. Also, given a collection Q of connected components $\mathcal{C}_j$, denote by $D(Q)$ the set of all disks having an arc in common with the boundary of an element of Q. Moreover, since there is a one-to-one correspondence between a disk and a d-row, use X_l to denote both the sets of disks and the corresponding sets of d-rows. Similarly, use the same notation for the set Y_l of connected components and the corresponding set of c-rows.

The multistep procedure is as follows.

Step 1. Set $X_1 = D^{(1)}$ and $Y_1 = \mathcal{C}(X_1)$. The set Y_1 consists of all connected components adjacent to $D^{(1)}$. We also identify X_1 with the d-row $\mathbf{a}_1$. Recall that X_1 is eliminated in the above reduction. The two-ones property implies that for each $\mathbf{a}_j \in Y_1$, there is a column of A_R with the only nonzero entry belonging to row $\mathbf{a}_j$.

[6] Note that, in fact, for large N and M, we have $P > 3M - O(1)$ as $M \to \infty$.

This is the single-one property and it follows from the two-ones property after the elimination of X_1. This shows that if $\mathbf{a}_k \in Y_1$, it cannot appear in the left-hand side of (6.40) and so it appears in the right-hand side of (6.40) with coefficient $\lambda^k = 0$.

Step 2. Let $X_2 = D(Y_1) \setminus X_1$ and observe that X_2 consists of all disks except $D^{(1)}$, which have a part of the boundary in common with one of the connected components in Y_1. Then define $Y_2 = \mathcal{C}(X_2) \setminus Y_1$. The elements of Y_2 are connected components which do not belong to Y_1 and whose boundary intersects the boundary of some disk from X_2. Again, none of the vectors in $X_2 \cup Y_2$ can be in the left-hand side of (6.40) and so they must be in the right-hand side of (6.40), with corresponding coefficients λ_m equal to zero.

Step 3. Define Y_l, X_l recursively by

$$X_l = D(Y_{l-1}) \setminus X_{l-1}, \quad Y_l = \mathcal{C}(X_l) \setminus Y_{l-1}.$$

The elements of X_l are disks that do not belong to X_{l-1} and whose boundary intersects the boundary of some connected component in Y_{l-1}. The set Y_l consists of the connected components which do not belong to Y_{l-1} and whose boundary intersects the boundary of some disk in X_l. Repeating the argument used in the previous step, we show that all corresponding λ_m must be zero.

By Property 3.2 of graph Γ, sets Y_l and X_l are nonempty, unless for some L, $Y_{L-1} = Y_L = \Omega_F \setminus U_\Pi$, or $X_L = X_{L-1} = \{D^{(1)}, \dots, D^{(N)}\}$. Then we stop the process and note that the rows remaining in (6.40) are either all d-rows or all c-rows. By the two-ones property, we obtain that if a d-row has a unit entry in some column, the other d-rows have zeros in the same column. Hence all d-rows are linearly independent. The same reasoning yields linear independence of all c-rows. This means that by the time we stop the process, all vectors possibly remaining in (6.40) are linearly independent and that the coefficients λ_m in front of these rows must be zero. This finishes the proof of Lemma 6.2 and of Theorem 6.1. □

Remark 6.3. The above iterative procedure can be illustrated as follows. Remove a disk $D^{(1)}$ from Ω. This disk has adjacent connected components, say, three of them, if there are three edges originating from $\mathbf{x}^{(1)}$. Remove these connected components. Now, the just-removed connected components were adjacent to three disks (second generation of disks) which are neighbors of $D^{(1)}$. Remove the second generation of disks and consider the remaining connected components (second generation of connected components) adjacent to them. Remove the second generation of connected components. The remaining neighbors of second generation disks are called third generation disks. Remove them and consider the remaining connected components adjacent to third-generation disks. Continue removing objects from Ω until there is nothing left. Due to the connectedness of the graph, the process does not stop until all the disks and all the connected components are removed.

Part III. Extending $\boldsymbol{\mathcal{S}}$ to $\boldsymbol{\partial D^{(j)} \cap \overline{U_E}}$. We wish to define a trial tensor $\mathcal{S}$ along the pth complementary arc in $\partial D^{(j)} \cap \overline{U_E}$ such that its integral is equal to $\boldsymbol{\beta}^p$ for some $1 \le j \le N$ and for $1 \le p \le P$. This ensures that conditions (6.34) hold and the existence of vectors $\boldsymbol{\beta}^p$ has been proved in Parts I and II. However, the trial stress tensor must also satisfy the balance of angular momentum

$$\int_{\partial D^{(j)}} \mathbf{n}^{(j)} \times \mathcal{S}\mathbf{n}^{(j)}\, ds = \mathbf{0} \tag{6.41}$$

for all $1 \le j \le N$. Let us then focus attention on one disk, say, $D^{(j)}$, of radius a_j centered at $\mathbf{x}^{(j)}$. At $\partial D^{(j)}$, $\mathbf{x} = \mathbf{x}^{(j)} + a_j \mathbf{n}^{(j)}$, so we rewrite (6.41) as

$$\mathbf{x}_0 \times \int_{\partial D^{(j)}} \mathcal{S}\, \mathbf{n}^{(j)}\, ds + a_j \int_{\partial D^{(j)}} \mathbf{n}^{(j)} \times \mathcal{S}\mathbf{n}^{(j)}\, ds = \mathbf{0}.$$

Due to (6.34), the first integral is zero, and (6.41) reduces to

$$\int_{\partial D} \mathbf{n}^{(j)} \times \mathcal{S}\mathbf{n}^{(j)}\, ds = \mathbf{0}.$$

Let $\boldsymbol{\tau}$ be the tangent unit vector at $\partial D^{(j)}$, pointing in the clockwise direction. Since $\mathcal{S}\mathbf{n}^{(j)} = (\mathcal{S}\mathbf{n}^{(j)} \cdot \boldsymbol{\tau})\boldsymbol{\tau} + (\mathcal{S}\mathbf{n}^{(j)} \cdot \mathbf{n}^{(j)})\mathbf{n}^{(j)}$ and $\boldsymbol{\tau} \cdot \mathbf{n}^{(j)} = 0$, we have

$$0 = \int_{\partial D^{(j)}} \mathbf{n}^{(j)} \times \mathcal{S}\mathbf{n}^{(j)}\, ds = \mathbf{k} \int_{\partial D^{(j)}} \mathcal{S}\mathbf{n}^{(j)} \cdot \boldsymbol{\tau}\, ds,$$

where $\mathbf{k} = \mathbf{n}^{(j)}(\mathbf{x}) \times \boldsymbol{\tau}(\mathbf{x})$ is a constant (independent of $\mathbf{x}$) unit vector, orthogonal to the two-dimensional plane and pointing into it. Therefore, any tensor $\mathcal{S}$ obeying the balance of angular momentum (6.41) satisfies

$$\int_{\partial D^{(j)} \cap \overline{U_E}} \mathcal{S}\mathbf{n}^{(j)} \cdot \boldsymbol{\tau}\, ds = - \int_{\partial D^{(j)} \cap \overline{U}_\Pi} \mathcal{S}\mathbf{n}^{(j)} \cdot \boldsymbol{\tau}\, ds. \tag{6.42}$$

Now, since $\mathcal{S}$ is already defined in U_Π, we estimate the integral in the right-hand side of (6.42). In the local coordinates of gap Π^{ij}, a complementary arc in $\partial D^{(j)} \cap \overline{\Pi^{ij}}$ is given by equation $f(x_1, x_2) = x_2 - \delta^{ij}/2 - a_j + \sqrt{a_j^2 - x_1^2} = 0$. Then

$$\mathbf{n}^{(j)} = \frac{1}{a_j} \begin{pmatrix} x_1 \\ -(a_j^2 - x_1^2)^{1/2} \end{pmatrix}, \qquad \boldsymbol{\tau} = \frac{1}{a_j} \begin{pmatrix} (a_j^2 - x_1^2)^{1/2} \\ x_1 \end{pmatrix}. \tag{6.43}$$

Using the explicit expression of $\mathcal{S}$ from section 6.2.2 and Lemma 6.1, we obtain

$$\int_{\partial D^{(j)} \cap \overline{\Pi^{ij}}} \mu \mathcal{E}(\mathbf{u})\mathbf{n}^{(j)} \cdot \boldsymbol{\tau}\, ds = O\left(\sqrt{\frac{a^{ij}}{\delta^{ij}}}\right) \tag{6.44}$$

and

$$\int_{\partial D^{(j)} \cap \overline{\Pi^{ij}}} \mathcal{K}\mathbf{n}^{(j)} \cdot \boldsymbol{\tau}\, ds = O(1). \tag{6.45}$$

Let $\mathcal{A}_{\Pi^{ij}}$ be a complementary arc from $\partial D^{(j)} \cap \overline{U_E}$ adjacent to gap Π^{ij} and oriented in the clockwise direction. We wish to construct tensor $\mathcal{S}$ on $\mathcal{A}_{\Pi^{ij}}$ so that

$$\int_{\mathcal{A}_{\Pi^{ij}}} \mathcal{S}\mathbf{n}^{(j)}\, ds = \boldsymbol{\beta} \tag{6.46}$$

and

$$\int_{\mathcal{A}_{\Pi^{ij}}} \mathcal{S}\mathbf{n}^{(j)} \cdot \boldsymbol{\tau} ds = -\rho. \tag{6.47}$$

Here, $\boldsymbol{\beta}$ is found by solving the d-c-system, and ρ stands for the sum of the integrals in (6.44), (6.45). Parameterize $\mathcal{A}_{\Pi^{ij}}$ as follows:

$$\mathcal{A}_{\Pi^{ij}} = \{(x_1, x_2) \in \partial D^{(j)} : x_1 = a_j \cos t, x_2 = a_j \sin t,\ t \in [0, \alpha]\}. \tag{6.48}$$

Then we rewrite (6.46), (6.47) as

$$a\int_0^\alpha (\mathcal{S}_{11}(t)\cos t - \mathcal{S}_{12}\sin t)dt = \beta_1, \tag{6.49}$$

$$a_j\int_0^\alpha (\mathcal{S}_{12}(t)\cos t - \mathcal{S}_{22}\sin t)dt = \beta_2,$$

$$a_j\int_0^\alpha (\mathcal{S}_{11}(t)\cos t\sin t - \mathcal{S}_{22}\sin t\cos t)dt = -\rho.$$

To accomplish the task of this section, we now set the components $\mathcal{S}_{kl}$ of our trial tensor on $\mathcal{A}_{\Pi^{ij}}$ to constant values satisfying (6.49). This implies that

$$M\begin{pmatrix}\mathcal{S}_{11}\\ \mathcal{S}_{12}\\ \mathcal{S}_{13}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\beta}_1/a_j\\ \boldsymbol{\beta}_2/a_j\\ -\rho/a_j\end{pmatrix}, \tag{6.50}$$

where

$$M = \begin{pmatrix}\sin\alpha & \cos\alpha - 1 & 0\\ 0 & \sin\alpha & \cos\alpha - 1\\ 1/2\sin^2\alpha & 0 & -1/2\sin^2\alpha\end{pmatrix}$$

and $\det(M) = \sin^2\alpha\cos\alpha\,(1-\cos\alpha)$. Thus, unless $\alpha = 0, \pi/2$ or π, (6.50) is uniquely solvable, and our extension of $\mathcal{S}$ to the complementary arc $\mathcal{A}_{\Pi^{ij}}$ is complete. Finally, all cases of α that make M singular can be eliminated. Indeed, $\alpha = 0$ is discarded by the observation that it implies an empty $\mathcal{A}_{\Pi^{ij}}$. The other cases, $\alpha = \pi/2$ or π, can also be avoided by modifying the length of $\mathcal{A}_{\Pi^{ij}}$, i.e., by changing slightly the widths of gaps Π^{ij} adjacent to $\mathcal{A}_{\Pi^{ij}}$.

6.2.4. Extension of $\mathcal{S}$ in the set U_E of connected components. The goal of this section is to extend $\mathcal{S}$ outside the gaps in such a way that the dual dissipation rate in U_E is much smaller than $O(\delta^{-3/2})$. First, we show that the components of the extended $\mathcal{S}$ at complementary arcs $\mathcal{A}_{\Pi^{ij}} \in \partial D^{(j)}$ (see section 6.2.3) for $1 \le j \le N$ and $i \in \mathcal{N}_j$ are bounded, pointwise by $C_{kl}\delta^{-1/2}$. Then we consider the extension of $\mathcal{S}$ from $\mathcal{A}^+ = \overline{U_\Pi} \cap \partial\Omega^+$ and $\mathcal{A}^- = \overline{U_\Pi} \cap \partial\Omega^-$ (the parts of $\partial\Omega^\pm$ included in gaps) to the whole $\partial\Omega$ and we prove the pointwise estimate $|\mathcal{S}_{kl}| \le C_{kl}\delta^{-1/2}$ for $k, l = 1, 2$. Finally, we extend $\mathcal{S}$ in the interior of U_E and show that the dissipation rate there is at most $O(\delta^{-1})$. Once this is done, the first three steps in the outline of section 6.2.1 would be completed. The fourth step in section 6.2.1 is accomplished in section 6.2.6, where we give the main theorem of the paper.

Part I: Estimates on the boundaries of the disks. To prove the desired pointwise estimates of the components of $\mathcal{S}$ at $\partial D^{(i)}$ for $1 \le i \le N$, we need the following proposition.

PROPOSITION 6.4. *For each disk $D^{(i)}$, we have*

$$\sum_{j\in\mathcal{N}_i}\int_{\partial D^{(i)}\cap\overline{\Pi^{ij}}} \mathcal{S}\mathbf{n}^{(i)}\,ds = O(\delta^{-1/2}) \quad \textit{for } i = 1, \dots, N.$$

This proposition states that if we fix a disk $D^{(i)}$ and consider the forces that act on each arc in $\partial D^{(i)} \cap \overline{U_\Pi}$, then the sum of these forces over all the arcs is $O(\delta^{-1/2})$,

whereas the force on each disk may be of order $\delta^{-3/2}$. Thus, we have a cancellation of terms due to the fact that the forces depend on translation velocities $\mathbf{T}^{(i)}$, the solutions of network equations (6.2). This is yet another manifestation of the global nature of the lower bound construction, which cannot be obtained by simply patching together trial functions obtained in each gaps Π^{ij}.

Proof of Proposition 6.4. Fix a gap Π^{ij} which joins disks $D^{(i)}$ and $D^{(j)}$ and use (6.22) to calculate

$$\int_{\Pi^{ij}} (\mu\triangle\mathbf{u} - \nabla P)\cdot\mathbf{u}d\mathbf{x} = \int_{\Pi^{ij}} \text{div}\,(2\mu\mathcal{E}(\mathbf{u}) - P\mathcal{I})\cdot\mathbf{u}d\mathbf{x},$$

where $\mathbf{u}$ is defined by (6.6). The approximate pressure P is defined by (6.23). Integrating by parts, using the symmetry of $\mathcal{E}$ and the incompressibility of $\mathbf{u}$, we have

$$\int_{\Pi^{ij}} (\mu\triangle\mathbf{u} - \nabla P)\cdot\mathbf{u}d\mathbf{x} = -2\mu\int_{\Pi^{ij}} \mathcal{E}(\mathbf{u})\cdot\mathcal{E}(\mathbf{u})d\mathbf{x} + \int_{\partial\Pi^{ij}} \mathcal{S}_0\mathbf{n}\cdot\mathbf{u}\,ds. \tag{6.51}$$

Recall that $\mu\triangle\mathbf{u} - \nabla P = \text{div}\,\mathcal{S}_0 = \text{div}\,\mathcal{K}$, where $\mathcal{K}$ is the compensating tensor defined in (6.25). Write $\mathcal{S}_0 = \mathcal{S} + \mathcal{K}$ and integrate by parts to obtain

$$2\mu\int_{\Pi^{ij}} \mathcal{E}(\mathbf{u})\cdot\mathcal{E}(\mathbf{u})d\mathbf{x} = \int_{\partial\Pi^{ij}} \mathcal{S}\mathbf{n}\cdot\mathbf{u}d\Gamma + \int_{\Pi^{ij}} \mathcal{K}\cdot\mathcal{E}(\mathbf{u})d\mathbf{x}. \tag{6.52}$$

In section 6.2.2 we showed that $\int_{\Pi} \mathcal{K}\cdot\mathcal{E}(\mathbf{u})d\mathbf{x} = O(\delta^{-1/2})$. Then rewrite the first integral in the right-hand side of (6.52) as

$$\int_{\partial\Pi^{ij}} \mathcal{S}\mathbf{n}\cdot\mathbf{u}ds = \int_{\partial D^{(i)}\cap\partial\Pi^{ij}} \mathcal{S}\mathbf{n}^{(i)}\cdot\mathbf{u}\,ds + \int_{\partial D^{(j)}\cap\partial\Pi^{ij}} \mathcal{S}\mathbf{n}^{(j)}\cdot\mathbf{u}\,ds + \int_{\partial U_E\cap\partial\Pi^{ij}} \mathcal{S}\mathbf{n}\cdot\mathbf{u}\,ds$$

$$= \int_{\partial D^{(i)}\cap\partial\Pi^{ij}} \mathcal{S}\mathbf{n}^{(i)}\cdot\mathbf{T}^{(i)}\,ds + \int_{\partial D^{(j)}\cap\partial\Pi^{(ij)}} \mathcal{S}\mathbf{n}^{(j)}\cdot\mathbf{T}^{(j)}\,ds + \int_{\partial U_E\cap\partial\Pi^{ij}} \mathcal{S}\mathbf{n}\cdot\mathbf{u}\,ds$$

$$+ \int_{\partial D^{(i)}\cap\partial\Pi^{ij}} \mathcal{S}\mathbf{n}^{(i)}\cdot(\mathbf{u} - \mathbf{T}^{(i)})ds + \int_{\partial D^{(j)}\cap\partial\Pi^{ij}} \mathcal{S}\mathbf{n}^{(i)}\cdot(\mathbf{u} - \mathbf{T}^{(j)})ds.$$

However, by Proposition 6.2, we have

$$\int_{\partial D^{(i)}\cap\partial\Pi^{ij}} \mathcal{S}\mathbf{n}^{(i)}\cdot\mathbf{u}ds = -\int_{\partial D^{(j)}\cap\partial\Pi^{ij}} \mathcal{S}\mathbf{n}^{(j)}\cdot\mathbf{u}ds,$$

and using the constructed $\mathcal{S}$ and $\mathbf{u}$ (see sections 6.1 and 6.2.2) and Lemma 6.1 gives

$$\int_{\partial\Pi^{ij}} \mathcal{S}\mathbf{n}\cdot\mathbf{u}ds = (\mathbf{T}^{(i)} - \mathbf{T}^{(j)})\cdot\int_{\partial D^{(i)}\cap\partial\Pi^{ij}} \mathcal{S}\mathbf{n}^{(i)}ds + O(\delta^{-1/2}).$$

Finally, combining this with (6.52) yields

$$W_{\Pi^{ij}}(\mathbf{u}) = \frac{1}{2}(\mathbf{T}^{(i)} - \mathbf{T}^{(j)})\cdot\int_{\partial D^{(i)}\cap\partial\Pi^{ij}} \mathcal{S}\mathbf{n}^{(i)}ds + O(\delta^{-1/2}). \tag{6.53}$$

Next, recall that $W_{\Pi^{ij}}(\mathbf{u})$ in the left-hand side of (6.53) is a quadratic form in $\mathbf{T}^{(i)} - \mathbf{T}^{(j)}$ for $i = 1,\dots,N$ and $j \in \mathcal{N}_i$ (see section 6.1). Denote the matrix of this

quadratic form by $A(\delta)$. From the definition (6.6) of $\mathbf{u}$, it follows that $\mathcal{S}$ is a linear function of $\mathbf{T}^i - \mathbf{T}^j$, so we write

$$\frac{1}{2}\int_{\partial D^{(i)}\cap\partial\Pi^{ij}} \mathcal{S}\mathbf{n}^{(i)}ds = B(\delta)(\mathbf{T}^{(i)} - \mathbf{T}^{(j)}), \tag{6.54}$$

where the matrix $B(\delta)$ is independent of $\mathbf{T}^i, \mathbf{T}^j$. Then the first term in the right-hand side of (6.53) is a quadratic form in $\mathbf{T}^{(i)} - \mathbf{T}^{(j)}$. Replacing the terms in (6.53) with the corresponding quadratic forms, we obtain

$$(\mathbf{T}^{(i)} - \mathbf{T}^{(j)}) \cdot A(\delta)(\mathbf{T}^{(i)} - \mathbf{T}^{(j)}) = (\mathbf{T}^{(i)} - \mathbf{T}^{(j)}) \cdot B(\delta)(\mathbf{T}^{(i)} - \mathbf{T}^{(j)}) + O(\delta^{-1/2}). \tag{6.55}$$

Summing up over all disks $D^{(i)}, i = 1, \dots, N$, and then differentiating with respect to the components of a fixed vector $\mathbf{T}^{(i)}$, we have

$$\sum_{j\in\mathcal{N}_i} A(\delta)(\mathbf{T}^{(i)} - \mathbf{T}^{(j)}) = \sum_{j\in\mathcal{N}_i} B(\delta)(\mathbf{T}^{(i)} - \mathbf{T}^{(j)}) + O(\delta^{-1/2}) \tag{6.56}$$

for each disk $D^{(i)}, i = 1, \dots, N$. However, by the network equations (6.2), the left-hand side in (6.56) is zero and so

$$0 = \sum_{j=1}^{J_i} B(\delta)(\mathbf{T}^i - \mathbf{T}^j) + O(\delta^{-1/2}).$$

This completes the proof of Proposition 6.4. ∎

Part II: Controlled extension of $\mathcal{S}$ to $\partial\Omega$. In this step, we deal with the extension of $\mathcal{S}$ from $\mathcal{A}^{\pm}$ to $\partial\Omega$.

PROPOSITION 6.5.

$$\int_{\mathcal{A}^+} \mathcal{S}\mathbf{n}\, ds + \int_{\mathcal{A}^-} \mathcal{S}\mathbf{n}\, ds = O(\delta^{-1/2}).$$

Proof. Since $\mathcal{S}$ is divergence free in each gap Π^{ij}, we have

$$\sum_{ij}\int_{\partial\Pi^{ij}} \mathcal{S}\mathbf{n}\, ds = \mathbf{0}, \tag{6.57}$$

where the sum is taken over all gaps. If a gap is connected to $\partial\Omega$, its boundary consists of a segment from $\mathcal{A}^+$ or $\mathcal{A}^-$, an arc which belongs to one of the disks, and two lateral segments. The boundary of an interior gap contains two disk arcs and two lateral segments. By Proposition 6.2, the sum of the integrals over the lateral parts of $\partial\Pi^{ij}$ is zero. Thus (6.57) reduces to

$$\int_{\mathcal{A}^+\cup\mathcal{A}^-} \mathcal{S}\mathbf{n}\, ds + \sum_{j=1}^{N}\int_{\partial D^{(j)}\cap\overline{U}_\Pi} \mathcal{S}\mathbf{n}\, ds = 0,$$

where the sum is taken over all disks. By Proposition 6.4,

$$\sum_{j=1}^{N}\int_{\partial D^{(j)}\cap\overline{U}_\Pi} \mathcal{S}\mathbf{n}\, ds = O(\delta^{-1/2}),$$

and Proposition 6.5 follows. ∎

To obtain pointwise estimates on the complementary arcs, we restrict our attention to the example of a three-disk network from Figure 7 and to the corresponding algebraic system (6.36), (6.37). Generalizing our arguments to the general case of N disks is straightforward.

Let us denote the vector $\mathbf{F}^1 + \mathbf{F}^5$ by $-\mathbf{P}$. By Proposition 6.5, $\mathbf{P} = O(\delta^{-1/2})$. Define $\mathcal{S} = 0$ on the lateral part of $\partial\Omega$. On $\partial\Omega^+ \setminus \mathcal{A}^+$ (or $\partial\Omega^- \setminus \mathcal{A}^-$), we choose the constant components $\mathcal{S}_{11} = 0$,

$$\mathcal{S}_{12} = \pm\frac{1}{2}\frac{\mathbf{P}_1}{|\partial\Omega^+ \setminus \mathcal{A}^+|}, \quad \mathcal{S}_{22} = \pm\frac{1}{2}\frac{\mathbf{P}_2}{|\partial\Omega^+ \setminus \mathcal{A}^+|},$$

where $|\cdot|$ denotes the length of a curve. Then

$$\int_{\partial\Omega^+\setminus\mathcal{A}^+} \mathcal{S}\mathbf{n}\, ds = \int_{\partial\Omega^-\setminus\mathcal{A}^-} \mathcal{S}\mathbf{n}\, ds = \frac{1}{2}\mathbf{P}, \quad \int_{\partial\Omega} \mathcal{S}\mathbf{n}\, ds = 0,$$

and

$$\sup_{\partial\Omega\setminus(\mathcal{A}^+\cup\mathcal{A}^-)} |\mathcal{S}_{kl}| \le c_{kl}\delta^{-1/2}, \quad k,l = 1,2, \tag{6.58}$$

with c_{kl} independent of δ. Thus, we can define $\mathcal{S}$ on $\partial\Omega \setminus (\mathcal{A}^+ \cup \mathcal{A}^-)$ so that (6.35) is satisfied and

$$\mathbf{D}^1 = O(\delta^{-1/2}), \quad \mathbf{D}^1 = O(\delta^{-1/2}). \tag{6.59}$$

From the definitions of $\mathbf{u}, P$, and $\mathcal{S}$, it follows that $|\mathcal{S}_{kl}|$ are pointwise bounded independent of δ on the lateral parts of the gap boundaries. Hence,

$$\mathbf{B}^j = O(1) \quad \text{for } 1 \le j \le 4. \tag{6.60}$$

Moreover, by Proposition 6.4,

$$-\mathbf{F}^1 + \mathbf{F}^3 + \mathbf{F}^4 = O(\delta^{-1/2}), \quad -\mathbf{F}^2 + \mathbf{F}^3 = O(\delta^{-1/2}), \quad -\mathbf{F}^4 - \mathbf{F}^3 - \mathbf{F}^5 = O(\delta^{-1/2}), \tag{6.61}$$

and, combining (6.59)–(6.61), we see that the components of the right-hand side of the algebraic system (6.36), (6.37) are bounded by $c\delta^{-1/2}$ with c independent of δ. Since the matrix of this algebraic system is independent of δ as well, we can choose a solution of (6.36), (6.37) so that all its components are bounded by $c\delta^{-1/2}$ with c independent of δ. This means that for each complementary arc in $\partial D^{(j)} \cap \partial U_E$, $1 \le j \le 8$, $\int_{\partial D^{(j)}\cap\partial U_E} \mathcal{S}\mathbf{n} ds = O(\delta^{-1/2})$. The latter implies that the right-hand side of algebraic system (6.50) is bounded by $c\delta^{-1/2}$ and, since matrix M is independent of δ and invertible, we can find a stress field $\mathcal{S}$ at the boundaries of the disks which is bounded by $c\delta^{-1/2}$ with c independent of δ. Then for each complementary arc, we have

$$\sup_{\partial D^{(j)}\cap\partial U_E} |\mathcal{S}_{kl}| \le c_{ij}\delta^{-1/2} \quad \text{for } 1 \le j \le N \tag{6.62}$$

with c_{ij} independent of δ.

The boundary of each connected component $\mathcal{C}_j$ of U_E consists of complementary arcs, pieces of the external boundary $\partial\Omega$, and lateral parts of gap boundaries. Thus

the estimates (6.58), (6.62) and the uniform estimates on the lateral segments yield the following.

PROPOSITION 6.6. *For each connected component $\mathcal{C}_j$ of U_E, we have*

$$\sup_{\partial\mathcal{C}_j} |\mathcal{S}_{kl}| \le c_j \delta^{-1/2}, \quad k,l = 1,2,$$

with c_j independent of δ.

Part III: Extension from the boundary, to the connected components. Now, $\mathcal{S}$ is defined on the boundary of each connected component $\mathcal{C}_j$ of $U_E = \Omega_F \setminus U_\Pi$. Moreover, we have

$$\int_{\partial\mathcal{C}_j} \mathcal{S}\mathbf{n}\, ds = 0 \tag{6.63}$$

and

$$\sup_{\partial\mathcal{C}_j} |\mathcal{S}(\mathbf{x})| \le c\delta^{-1/2} \tag{6.64}$$

with c independent of δ. Next, we construct a divergence-free extension of $\mathcal{S}$ from ∂U_E to U_E.

PROPOSITION 6.7. *Let $\mathcal{C}_j \subset U_E$ be a connected component, and let $\mathcal{S}$ be the trial tensor defined on $\partial\mathcal{C}_j$ satisfying* (6.63), (6.64). *Then there exists an extension $\hat{\mathcal{S}} \in L^2(\mathcal{C}_j)$ in $\mathcal{C}_j$ such that* $\operatorname{div}\hat{\mathcal{S}} = 0$ *in $\mathcal{C}_j$, $\hat{\mathcal{S}} = \mathcal{S}$ on $\partial\mathcal{C}_j$ and*

$$\int_{\mathcal{C}_j} \hat{\mathcal{S}} \cdot \hat{\mathcal{S}}\, ds \le c\delta^{-1}$$

with c independent of δ.

This proposition is proved using the same techniques as those in section 5.2.[7]

6.2.5. Estimates of the dual dissipation functional in the gaps. In our proof of the main theorem of the paper (see section 6.2.6), we use the following estimate.

PROPOSITION 6.8. *Let* $\mathbf{u}$ *be defined by* (6.6) *and let $\mathcal{S}$ be the trial tensor defined by* (6.17) *and* (6.23)–(6.25). *Then*

$$W^*_{U_\Pi}(\mathcal{S}) = W_{U_\Pi}(\mathbf{u}) + O(\delta^{-1/2}). \tag{6.65}$$

Proof. By (6.8), we have $W^*_{U_\Pi} = \int_{\partial\Omega\cap\overline{U}_\Pi} \mathbf{g}\cdot\mathcal{S}\mathbf{n} ds - \int_{U_\Pi} F(\mathcal{S}) d\mathbf{x}$. Let us then estimate first the boundary integral. Fix a gap Π and consider $\int_{\partial\Pi} \mathcal{S}\mathbf{n}\cdot\mathbf{u}\, ds$. Integrating by parts using the definition of $\mathcal{S}$, the incompressibility of $\mathbf{u}$, and (6.18), we get

$$\int_{\partial\Pi} \mathcal{S}\mathbf{n}\cdot\mathbf{u}\, ds = 2W_\Pi(\mathbf{u}) - \int_\Pi \mathcal{K}\cdot\mathcal{E}(\mathbf{u}) d\mathbf{x}. \tag{6.66}$$

The integral $\int_\Pi \mathcal{K}\cdot\mathcal{E}(\mathbf{u}) d\mathbf{x}$ is estimated using the explicit expressions of $\mathcal{K}$ and $\mathcal{E}(\mathbf{u})$ and then by applying Lemma 6.1. These calculations, already carried out in section

[7] See also Appendix C in the preprint version of this article, available at http://www.math.wsu.edu/math/faculty/panchenko/welcome.html.

6.2.2 show that the integral in the right-hand side of (6.66) is $O(\delta^{-1/2})$. Summing up over all gaps, we have

$$\int_{\partial U_\Pi} \mathcal{S}\mathbf{n}\cdot\mathbf{u}\,ds = 2W_{U_\Pi}(\mathbf{u}) + O(\delta^{-1/2}). \tag{6.67}$$

Since ∂U_Π is a union of circular arcs, lateral segments that belong to the gap boundaries, and a set $(\partial\Omega\cap\bar{U}_\Pi)\subset\partial\Omega$, we write

$$\begin{aligned}\int_{\partial\Omega\cap\overline{U}_\Pi} \mathbf{g}\cdot\mathcal{S}\mathbf{n}\,ds \;&= \int_{\partial U_\Pi}\mathbf{u}\cdot\mathcal{S}\mathbf{n}\,ds\\ &\quad -\sum_{j=1}^{N}\mathbf{T}^{(j)}\cdot\int_{\partial D^{(j)}\cap\overline{U}_\Pi}\mathcal{S}\mathbf{n}\,ds\\ &\quad -\int_{\partial U_\Pi\setminus(\cup_j\partial D^{(j)})}\mathbf{u}\cdot\mathcal{S}\mathbf{n}\,ds + O(\delta^{-1/2}).\end{aligned} \tag{6.68}$$

Here, we use the same technique that gave (6.53) from (6.52). By Proposition 6.4, the sum of the second and third terms in the right-hand side of (6.68) is $O(\delta^{-1/2})$. Hence,

$$\int_{\partial\Omega\cap\overline{U}_\Pi}\mathbf{g}\cdot\mathcal{S}\mathbf{n}\,ds = \int_{\partial U_\Pi}\mathbf{u}\cdot\mathcal{S}\mathbf{n}\,ds + O(\delta^{-1/2}). \tag{6.69}$$

Combining (6.67) and (6.69), we obtain

$$\int_{\partial\Omega\cap\overline{U}_\Pi}\mathbf{g}\cdot\mathcal{S}\mathbf{n}\,ds = 2W_{U_\Pi}(\mathbf{u}) + O(\delta^{-1/2}). \tag{6.70}$$

Finally, to estimate the second integral in the definition of $W^*_{U_\Pi}$, we apply section 6.14 and sum up over all gaps:

$$\int_{U_\Pi} F(\mathcal{S})d\mathbf{x} = W_{U_\Pi}(\mathbf{u}) + O(\delta^{-1/2}). \tag{6.71}$$

The estimate (6.65) follows from (6.70) and (6.71). □

We remark here that construction of the lower bound which accounts for rotations (with the error term $O(1)$) requires developing more sophisticated techniques even for periodic densely packed arrays, and this will be addressed elsewhere.

6.2.6. The main theorems. The trial field for the upper bound (6.7) is constructed by patching up the local approximate solutions (6.6), which depend on the translational particle velocities $\mathbf{T}^{(i)}, i=1,\dots,N$, minimizing the quadratic functional

$$\begin{aligned}Q = \sum_{i=1}^{N}\sum_{\substack{j\in\mathcal{N}_i\\ j<i}}&\left\{\left[\frac{3\pi\mu}{4}\left(\frac{a}{\delta^{ij}}\right)^{\frac{3}{2}} + \frac{27\pi\mu}{10}\sqrt{\frac{a}{\delta^{ij}}}\right][(\mathbf{T}^{(i)}-\mathbf{T}^{(j)})\cdot\mathbf{q}^{ij}]^2\right.\\ &\left.+\frac{\pi\mu}{2}\sqrt{\frac{a}{\delta^{ij}}}[(\mathbf{T}^{(i)}-\mathbf{T}^{(j)})\cdot\mathbf{p}^{ij}]^2\right\}\\ &+\sum_{i\in\mathcal{B}}\left\{\left[\frac{3\pi\mu}{4}\left(\frac{2a}{\delta^{i}}\right)^{\frac{3}{2}} + \frac{27\pi\mu}{10}\sqrt{\frac{2a}{\delta^{i}}}\right][(\mathbf{T}^{(i)}-\mathbf{g})\cdot\mathbf{q}^{i}]^2\right.\\ &\left.+\frac{\pi\mu}{2}\sqrt{\frac{2a}{\delta^{i}}}[(\mathbf{T}^{(i)}-\mathbf{g})\cdot\mathbf{p}^{i}]^2\right\}.\end{aligned} \tag{6.72}$$

In this section we introduce the corresponding dissipation rate

$$E_2 = \min_{\substack{\mathbf{T}^{(i)} \\ i=1,\dots,N}} Q = Q(\mathbf{T}^{(i)}_{min}, i = 1, \dots, N). \tag{6.73}$$

By Proposition 6.1, the minimizing collection of vectors $\mathbf{T}^{(i)}_{min}$ (solving the system (6.2)) is unique.

Since the error terms appearing in the construction of the lower bound are of order δ^{-1} (recall Proposition 6.7), we need to make sure that $E_2 \geq c\delta^{-3/2}$ with c independent of δ. So far, we know from the upper bound in section 6.1 that if $(\mathbf{T}^{(i)} - \mathbf{T}^{(j)})_2 \neq 0$, the local dissipation rate in each gap Π^{ij} blows up as $\delta^{-3/2}$. Otherwise, the rate of growth is at most $\delta^{-1/2}$. The vectors $\mathbf{T}^{(i)}_{min}$ are solutions of a (large) system of network equations (6.2) and, until these are solved, we cannot say whether the quantities $((\mathbf{T}^{(i)}_{min} - \mathbf{T}^{(j)}_{min}) \cdot \mathbf{q}^{ij})^2$ vanish as $\delta \to 0$. That is, we cannot determine the global rate of blow up of E_2 as $\delta \to 0$. In the scalar case of electrical conduction, it has been shown in [5] that for all connected graphs, the total energy blows up at the same rate as the energy in each gap. In the vectorial case considered here, connectivity is not sufficient to ensure the analogous property. The global rate of blow up depends on other geometrical characteristics of a connected graph (e.g., the coordination number; see [7] for details).

The functional Q depends on the interparticle distances $\delta^{ij} = \delta d^{ij}$, where the rescaled distances d^{ij} do not depend on δ, and $0 < c \leq d^{ij} \leq 1$ for all pairs of neighboring disks. To study asymptotic behavior of Q as $\delta \to 0$, we factor out the powers of δ and write

$$Q(\mathbf{T}^{(1)}, \dots, \mathbf{T}^{(N)}) = \delta^{-3/2}\widehat{Q}(\mathbf{T}^{(1)}, \dots, \mathbf{T}^{(N)}) + \delta^{-1/2}Q'(\mathbf{T}^{(1)}, \dots, \mathbf{T}^{(N)}), \tag{6.74}$$

where the coefficients of Q and Q' do not depend on δ, and

$$\widehat{Q} = \sum_{i=1}^{N} \sum_{\substack{j \in \mathcal{N}_i \\ j < i}} A^{ij}[(\mathbf{T}^{(i)} - \mathbf{T}^{(j)}) \cdot \mathbf{q}^{ij}]^2 + \sum_{i \in \mathcal{B}} A^i[(\mathbf{T}^{(i)} - \mathbf{g}) \cdot \mathbf{q}^i]^2 \tag{6.75}$$

with

$$A^{ij} = \frac{3\pi\mu}{4}\left(\frac{a}{d^{ij}}\right)^{\frac{3}{2}}, \qquad A^i = \frac{3\pi\mu}{4}\left(\frac{2a}{d^i}\right)^{\frac{3}{2}}. \tag{6.76}$$

Note that our boundary conditions (2.9) correspond to $\mathbf{g} = \pm\mathbf{e}_2$ on $\partial\Omega^{\pm}$ and by Definition 3.3, $\mathbf{q}^i = \pm\mathbf{e}_2$ on $\partial\Omega^{\pm}$. We keep the general notation in (6.75) because it may be applied to more general boundary conditions. In this section we use the rescaled dissipation rate

$$\widehat{E} = \min_{\substack{\mathbf{T}^{(i)} \\ i=1,\dots,N}} \widehat{Q}, \tag{6.77}$$

which does not depend on δ, and the corresponding minimizers $\widehat{\mathbf{T}^{(i)}}, i = 1, \dots, N$. From (6.72) and (6.74),

$$\delta^{-3/2}\widehat{E} \leq \delta^{-3/2}\widehat{Q}(\mathbf{T}^{(i)}_{min}) \leq E_2 = Q(\mathbf{T}^{(i)}_{min}) \leq Q(\widehat{\mathbf{T}^{(i)}}) = \delta^{-3/2}\widehat{E} + \delta^{-1/2}Q'(\widehat{\mathbf{T}^{(i)}}). \tag{6.78}$$

Since Q' and $\widehat{\mathbf{T}^{(i)}}$ are independent of δ, (6.78) and (6.73) yield $\delta^{-3/2}\widehat{E} \le E_2 \le \delta^{-3/2}\widehat{E} + O(\delta^{-1/2})$ and thus

$$E_2 = \delta^{-3/2}\widehat{E} + O(\delta^{-1/2}). \tag{6.79}$$

Therefore, the inequality

$$\widehat{E} > 0 \tag{6.80}$$

would imply $E_2 = O(\delta^{-3/2})$ as $\delta \to 0$, and the leading term of the asymptotics of E_2 would be determined by minimizing the δ-independent functional $\widehat{Q}$.

In this paper, we consider a mathematical model for uniformly closely packed suspensions. For geometrical arrays of particles which correspond to such suspensions, the inequality (6.80) does hold. The detailed investigation of geometric properties of arrays for which (6.80) does or does not hold under various external boundary conditions is a subject of a separate investigation carried out in [7]. Here, we describe only one sufficient condition for validity of (6.80), discuss its physical relevance, and present an example which illustrates this condition.

Uniform, closely packed geometries can be modeled by the so-called densely packed quasi-triangular graphs. Roughly speaking, these are graphs such that each particle in the corresponding array has six neighbors, and the interparticle distances are uniformly small. More precisely, a quasi-triangular graph Γ is defined as follows. We start with a graph Γ' in Ω such that the interior vertices of Γ' are points of the triangular periodic lattice. Then Γ is obtained by perturbing the locations of the vertices of Γ' in such a way that if two vertices were neighbors, they would remain neighbors. Moreover, a vertex of Γ' is connected to $\partial\Omega$ if and only if the corresponding vertex of Γ is connected to $\partial\Omega$. More precisely, let Γ denote a network graph, and let Γ' be a graph corresponding to a periodic triangular lattice restricted to Ω. We also define K and K' to be (topological) complexes associated with Γ and Γ', respectively. We say that the graph Γ is quasi-triangular if K and K' are combinatorially equivalent. (The definition of combinatorial equivalence can be found, for instance, in [26, p. 4].)

To define the close packing condition for such graphs, recall that the interior vertices of Γ' are the centers of disks of radius a and that the corresponding periodic lattice is closely packed if the interparticle distance $\delta = l - 2a \ll 1$, where l denotes the length of an interior edge. For a densely packed quasi-triangular graph, we require that

$$\max_{ij} \delta^{ij} = \max_{ij}(l^{ij} - 2a) \ll 1, \tag{6.81}$$

where the maximum is taken over all pairs of neighbors and l^{ij} is the length of the corresponding interior edge of Γ.

In [7], we prove that (6.80) holds for a quasi-triangular graph under the close packing condition. An example of a network satisfying (6.80) is presented in the appendix.

We now formulate the main theorems.

THEOREM 6.2. *Let E be the dissipation rate* (2.20) *equal to the effective viscosity $\langle\mu\rangle$ up to a constant normalizing factor (see* (2.19)*). Then as $\delta \to 0$,*

$$E \le E_2 + O(1) \tag{6.82}$$

and

$$E_2 + O(\delta^{-1}) \le E, \tag{6.83}$$

where E_2 is the minimum of the quadratic form (6.72).

THEOREM 6.3. *For uniform, closely packed geometries such that condition* (6.80) *holds, the rescaled effective viscosity (dissipation rate) E defined by* (2.20) *has the asymptotic representation*

$$E = \delta^{-3/2}\widehat{E} + O(\delta^{-1}) \quad \text{as } \delta \to 0, \tag{6.84}$$

where $\widehat{E}$ is a minimum of the quadratic form $\widehat{Q}$ defined by (6.75).

COROLLARY 6.1. *Let $\langle\mu\rangle$ be the effective viscosity defined by* (2.19)*, and let the conditions of Theorem* 6.3 *hold. Then*

$$\langle\mu\rangle = \frac{\widehat{E}}{\int_\Omega(\mathcal{E}(\mathbf{u}^0),\mathcal{E}(\mathbf{u}^0))d\mathbf{x}}\delta^{-3/2} + O(\delta^{-1}) \quad \text{as } \delta \to 0, \tag{6.85}$$

where $\mathbf{u}^0$ solves the Stokes equation $\Delta\mathbf{u}^0 - \nabla P^0 = 0$ in Ω with the boundary conditions (2.9) *and* (2.10).

Proof of the Theorem 6.2. Let us define the trial tensor $\mathcal{S}$ in Ω_F as follows. In each gap Π^{ij}, we use formula (6.17), and in each connected component $\mathcal{C}_l$ of $U_E = \Omega_F \setminus U_\Pi$, we let $\mathcal{S}$ be an extension from $\partial\mathcal{C}_l$, as given in Proposition 6.7. Note that, through our construction, we have ensured that $\mathcal{S} \in \mathcal{F}$ and, as such, it is an admissible trial field for the dual variational problem (2.22). Furthermore, let $\mathbf{u} \in \mathcal{U}$, defined by (6.6), be the trial function for primal variational problem (2.20).

Let us evaluate the dual functional $W^*_{\Omega_F}(\mathcal{S})$ defined in (6.8) and (6.9). Since $\mathcal{S}\mathbf{n} = 0$ on $\partial\Omega \setminus (\partial\Omega^+ \cup \partial\Omega^-)$,

$$W^*_{\Omega_F}(\mathcal{S}) = \int_{\partial\Omega^+\cup\partial\Omega^-} \mathbf{g}\cdot\mathcal{S}\mathbf{n}ds - \int_{\Omega_F} F(\mathcal{S})d\mathbf{x}.$$

First, we estimate the boundary integral. By Proposition 6.6, $|\mathcal{S}| \le c\delta^{-1/2}$ on $(\partial\Omega^+ \cup \partial\Omega^-) \setminus \partial U_\Pi$, and $\mathbf{g}$ is independent of δ. Hence,

$$\int_{\partial\Omega^+\cup\partial\Omega^-} \mathbf{g}\cdot\mathcal{S}\mathbf{n}ds = \int_{(\partial\Omega^+\cup\partial\Omega^-)\cap\partial U_\Pi} \mathbf{g}\cdot\mathcal{S}\mathbf{n}ds + O(\delta^{-1/2}). \tag{6.86}$$

Next, we use the notation from (5.11) to write

$$\int_{\Omega_F} F(\mathcal{S})d\mathbf{x} = \int_{U_\Pi} F(\mathcal{S})d\mathbf{x} + \int_{U_E} F(\mathcal{S})d\mathbf{x}. \tag{6.87}$$

The second integral in the right-hand side of (6.87) is $O(\delta^{-1})$ by Proposition 6.7. Using (6.8) with $M = U_\Pi$ and taking into account the boundary conditions (2.10), we write

$$\int_{(\partial\Omega^+\cup\partial\Omega^-)\cap\partial U_\Pi} \mathbf{g}\cdot\mathcal{S}\mathbf{n}ds - \int_{U_\Pi} F(\mathcal{S})d\mathbf{x} = W^*_{U_\Pi}(\mathcal{S}), \tag{6.88}$$

and, combining (6.88) with (6.86) and (6.87), we obtain

$$W^*_{\Omega_F}(\mathcal{S}) = W^*_{U_\Pi}(\mathcal{S}) + O(\delta^{-1}). \tag{6.89}$$

Now Proposition 6.8 and (6.1), (6.72), and (6.73) imply

$$W^*_{\Omega_F}(\mathcal{S}) = W_{U_\Pi}(\mathbf{u}) + O(\delta^{-1}) = E_2 + O(\delta^{-1}). \tag{6.90}$$

Applying the direct and dual variational principles (2.20)–(2.21), (2.22)–(2.23) with the trial fields $\mathbf{u}$ defined in (6.6) and $\mathcal{S}$ defined in the beginning of the proof, we obtain

$$W^*_{\Omega_F}(\mathcal{S}) \le E \le W_{\Omega_F}(\mathbf{u}), \tag{6.91}$$

and the estimates (6.82) and (6.83) follow. □

Proof of Theorem 6.3. The inequalities (6.82) and (6.83) imply $E = E_2 + O(\delta^{-1})$. Together with (6.79), this yields $E = \widehat{E}\delta^{-3/2} + O(\delta^{-1})$, which gives the representation (6.84) provided (6.80) holds. □

7. Summary. In this paper we obtain and rigorously justify an asymtotic formula for the effective viscosity of a suspension of closely packed solid particles in a viscous Newtonian fluid. This formula accounts for variable distances between particles which form a nonperiodic (e.g., random) array.

The rigorous justification is presented in two dimensions. It is based on a construction of matching to the leading order lower and upper bounds by means of two, dual to each other, variational principles for the effective viscosity. The key point here is the construction of the lower bound, which accounts for all pairwise interactions between neighboring particles as well as for the incompressibility condition in the fluid domain. These interactions influence each other over the entire domain, leading to considerable difficulties in the construction of the corresponding trial function.

In both three and two dimensions, we obtain formal asymptotics formulas for the effective viscosity for nonperiodic arrays of particles of different sizes. For a particular case of a periodic array when identical particles move toward each other (along the line which joins their centers), the leading term in our formulas recovers the formal asymptotics previously obtained by [15, 16]. Our formulas also contain lower order terms which take into account the rotations and movements of adjacent particles in directions orthogonal to the axis of their centers. In our formal asymptotic analysis, we develop the corresponding generalization of the lubrication approximation.

While the previously obtained asymptotic formulas [15, 16, 27] capture the dependence of the effective viscosity on the volume fraction in a periodic array of closely packed particles, the network approximation proposed in this work also accounts for other geometrical characteristics such as variable distances between particles and the coordination number (the number of neighboring particles).

Appendix. Proof of estimate (6.80). Here we prove that (6.80) holds for the spring network corresponding to the graph in Figure 10.

PROPOSITION 7.1. *Let* $\widehat{Q}$ *be the rescaled dissipation rate* (6.75) *corresponding to the network in Figure* 10. *Then* $\min \widehat{Q} > 0$.

Proof. The functional $\widehat{Q}$ is of the form

$$\begin{aligned}\widehat{Q} = {} & A^{12}((\mathbf{T}^{(1)} - \mathbf{T}^{(2)}) \cdot \mathbf{q}^{12})^2 + A^{13}((\mathbf{T}^{(1)} - \mathbf{T}^{(3)}) \cdot \mathbf{q}^{13})^2 + A^{23}((\mathbf{T}^{(2)} - \mathbf{T}^{(3)}) \cdot \mathbf{q}^{23})^2 \\ & + A^{24}((\mathbf{T}^{(2)} - \mathbf{T}^{(4)}) \cdot \mathbf{q}^{24})^2 + A^{1}((\mathbf{T}^{(1)} - \tfrac{1}{2}\mathbf{e}_2) \cdot \mathbf{e}_2)^2 + A^{2}((\mathbf{T}^{(2)} - \tfrac{1}{2}\mathbf{e}_2) \cdot \mathbf{e}_2)^2 \\ & + A^{3}((\mathbf{T}^{(3)} + \tfrac{1}{2}\mathbf{e}_2) \cdot \mathbf{e}_2)^2 + A^{4}((\mathbf{T}^{(4)} + \tfrac{1}{2}\mathbf{e}_2) \cdot \mathbf{e}_2)^2,\end{aligned} \tag{7.1}$$

where $A^{ij}, A^i, i, j = 1, 2, \dots, 4$, are given by (6.76). We show that $\min_{\mathbf{T}^{(i)}, i=1,2\dots,4} \widehat{Q} > 0$. Arguing by contradiction, assume that $\min \widehat{Q} = 0$. This is possible only if the minimizing set of vectors $\mathbf{T}^{(i)}$ satisfies the system of equations

$$(\mathbf{T}^{(1)} - \mathbf{T}^{(2)}) \cdot \mathbf{q}^{12} = 0, \quad (\mathbf{T}^{(1)} - \mathbf{T}^{(3)}) \cdot \mathbf{q}^{13} = 0, \quad (\mathbf{T}^{(2)} - \mathbf{T}^{(3)}) \cdot \mathbf{q}^{23} = 0, \tag{7.2}$$

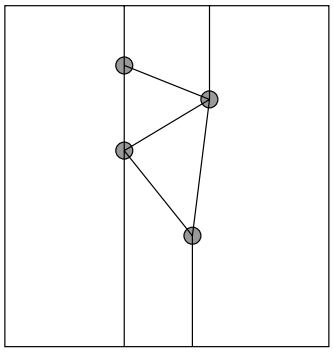

FIG. 10. *A four-disk network.*

$$(\mathbf{T}^{(2)} - \mathbf{T}^{(4)}) \cdot \mathbf{q}^{24} = 0, \qquad (\mathbf{T}^{(3)} - \mathbf{T}^{(4)}) \cdot \mathbf{q}^{34} = 0, \tag{7.3}$$

$$\mathbf{T}^{(1)} \cdot \mathbf{e_2} = \frac{1}{2}, \quad \mathbf{T}^{(2)} \cdot \mathbf{e_2} = \frac{1}{2}, \quad \mathbf{T}^{(3)} \cdot \mathbf{e_2} = -\frac{1}{2}, \quad \mathbf{T}^{(4)} \cdot \mathbf{e_2} = -\frac{1}{2}. \tag{7.4}$$

To write this system of nine equations in a more compact form $A\mathbf{z} = \mathbf{b}$, introduce a 1×8 vector of unknowns,

$$\mathbf{z} = (\mathbf{T}^{(1)}, \mathbf{T}^{(2)}, \mathbf{T}^{(3)}, \mathbf{T}^{(4)})^T.$$

The right-hand side of (7.2)–(7.4) is a 1×9 vector $\mathbf{b}$ that has the entries $b_i = 0, i = 1, 2, \ldots, 5$, $b_6 = b_7 = \frac{1}{2}$, $b_8 = b_9 = -\frac{1}{2}$. Performing (partial) Gaussian elimination on the transpose of the augmented matrix $(A \mid \mathbf{b})^T$, we find that it is similar to the matrix

$$C = \begin{pmatrix} \hat{\mathbf{q}}^{12} & \hat{\mathbf{q}}^{13} & \hat{\mathbf{o}} & \hat{\mathbf{o}} & \hat{\mathbf{o}} & \hat{\mathbf{e}}_2 & \hat{\mathbf{o}} & \hat{\mathbf{o}} & \hat{\mathbf{o}} \\ \hat{\mathbf{o}} & \hat{\mathbf{q}}^{13} & \hat{\mathbf{q}}^{23} & \hat{\mathbf{q}}^{24} & \hat{\mathbf{o}} & \hat{\mathbf{e}}_2 & \hat{\mathbf{e}}_2 & \hat{\mathbf{o}} & \hat{\mathbf{o}} \\ \hat{\mathbf{o}} & \hat{\mathbf{o}} & \hat{\mathbf{o}} & \hat{\mathbf{q}}^{24} & \hat{\mathbf{q}}^{34} & \hat{\mathbf{e}}_2 & \hat{\mathbf{e}}_2 & \hat{\mathbf{e}}_2 & \hat{\mathbf{o}} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 \end{pmatrix}. \tag{7.5}$$

Here

$$\begin{pmatrix} \hat{\mathbf{q}}^{12} & \hat{\mathbf{q}}^{13} & \hat{\mathbf{o}} & \hat{\mathbf{o}} & \hat{\mathbf{o}} & \hat{\mathbf{e}}_2 & \hat{\mathbf{o}} & \hat{\mathbf{o}} & \hat{\mathbf{o}} \end{pmatrix}$$

is a shorthand notation for two rows:

$$\mathbf{c}_1 = \begin{pmatrix} q_1^{12} & q_1^{13} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

and

$$\mathbf{c}_2 = \begin{pmatrix} q_2^{12} & q_2^{13} & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix},$$

and, similarly, the other two boldfaced rows in (7.5) are the shorthand notation for the four rows $\mathbf{c}_3, \ldots, \mathbf{c}_6$. The last two rows of C in (7.5) are denoted by $\mathbf{c}_7, \mathbf{c}_8$.

To show that $\mathbf{c}_j$, $j = 1, 2, \ldots, 8$, are linearly independent, argue by contradiction. When $\mathbf{c}_j$ are linearly dependent, there exist λ_j, $j = 1, 2, \ldots, 8$, not all zero, such that

$$\sum_{j=1}^{8} \lambda_j \mathbf{c}_j = 0. \tag{7.6}$$

Let us take the sixth and seventh columns in the vector equation (7.6). Then (7.6) yields the equations

$$\lambda_2 + \lambda_4 + \lambda_6 + \lambda_7 = 0 \quad \text{and} \quad \lambda_4 + \lambda_6 + \lambda_7 = 0. \tag{7.7}$$

Thus $\lambda_2 = 0$. Next, consider the first column in (7.6) to obtain $\lambda_1 q_1^{12} + \lambda_2 q_2^{12} = 0$. Note that q_1^{12} cannot be zero since the edge e^{12} connects the boundary vertices 1, 2, and thus cannot be vertical. Hence $\lambda_1 = 0$. Next, consider columns 2 and 3 in (7.6). Since $\lambda_1 = \lambda_2 = 0$, from (7.6) we obtain two equations for λ_3, λ_4:

$$\begin{aligned} \lambda_3 q_1^{13} + \lambda_4 q_2^{13} &= 0, \\ \lambda_3 q_1^{23} + \lambda_4 q_2^{23} &= 0. \end{aligned} \tag{7.8}$$

Since $\mathbf{q}^{13}, \mathbf{q}^{23}$ are linearly independent, $\lambda_3 = \lambda_4 = 0$.

Consider columns four and five in (7.6). Since λ_3, λ_4 are zero, we obtain two equations for λ_5, λ_6:

$$\begin{aligned} \lambda_5 q_1^{24} + \lambda_4 q_2^{24} &= 0, \\ \lambda_3 q_1^{34} + \lambda_4 q_2^{34} &= 0. \end{aligned} \tag{7.9}$$

Linear independence of $\mathbf{q}^{24}, \mathbf{q}^{34}$ implies $\lambda_5 = \lambda_6 = 0$. Returning to (7.7), we see that $\lambda_7 = 0$. Finally, considering column 9, we obtain the equation for λ_8,

$$\lambda_7 - \lambda_8 = 0, \tag{7.10}$$

which yields $\lambda_8 = 0$. Thus, all λ_j must be zero, and we arrive at a contradiction, which yields $\operatorname{rank}(A \mid \mathbf{b}) = 8$. Applying the same Gaussian elimination to A^T, we see that $\operatorname{rank}(A) = 7$ and thus $\operatorname{rank}(A \mid \mathbf{b}) > \operatorname{rank}(A)$. This means that the system (7.2)–(7.4) has no solutions, and the minimum of $\widehat{Q}$ must be positive. ∎

Remark 7.1. A quasi-triangular structure of the graph is sufficient for positivity of $\widehat{Q}$. The proof of the Proposition 7.1 shows that if a graph contains a triangulated path (see Figure 10), then $\widehat{Q} > 0$ for the external boundary conditions (2.9), (2.10). We now explain heuristically why triangulization ensures positivity of $\widehat{Q}$. Start from vertices $1, 2$ in Figure 10. They are connected by the nonvertical edge e^{12}, which implies $\lambda_1 = 0$. We next add vertex 3 and observe that it is connected to vertices $1, 2$ by noncollinear edges e^{12}, e^{23}, which are adjacent sides of a triangle 123. The noncollinearity of these edges implies $\lambda_3 = \lambda_4 = 0$. Next, add vertex 4 to obtain triangle 234 and, as before, noncollinearity of the edges e^{34}, e^{42} implies $\lambda_5 = \lambda_6 = 0$. Finally, since e^{34} is nonvertical, $\lambda_7 = \lambda_8 = 0$. This argument also admits a straightforward generalization to a triangulated path of n vertices.

REFERENCES

[1] G. Allaire, *Homogenization of the Navier-Stokes equations in open sets perforated with tiny holes.* I. *Abstract framework, a volume distribution of holes*, Arch. Ration. Mech. Anal., 113 (1990), pp. 209–259.

[2] G. Allaire, *Homogenization of the Navier-Stokes equations in open sets perforated with tiny holes.* II. *Noncritical sizes of the holes for a volume distribution and a surface distribution of holes*, Arch. Ration. Mech. Anal., 113 (1990), pp. 261–298.

[3] G. K. Batchelor and J. T. Green, *The determination of the bulk stress in a suspension of spherical particles to order c^2*, J. Fluid Mech., 56 (1972), pp. 401–427.

[4] L. Berlyand and E. Khruslov, *Homogenized non-Newtonian viscoelastic rheology of a suspension of interacting particles in a viscous Newtonian fluid*, SIAM J. Appl. Math., 64 (2004), pp. 1002–1034.

[5] L. Berlyand and A. Kolpakov, *Network approximation in the limit of small interparticle distance of the effective properties of a high-contrast random dispersed composite*, Arch. Ration. Mech. Anal., 159 (2001), pp. 179–227.

[6] L. Berlyand and A. Novikov, *Error of the network approximation for densely packed composites with irregular geometry*, SIAM J. Math. Anal., 34 (2002), pp. 385–408.

[7] L. Berlyand and A. Panchenko, *Geometric Instability in the Asymptotics of the Effective Viscosity for Highly Packed Suspensions of Rigid Particles*, preprint.

[8] L. Borcea, *Asymptotic analysis of quasi-static transport in high contrast conductive media*, SIAM J. Appl. Math., 59 (1998), pp. 597–635.

[9] L. Borcea, J. G. Berryman, and G. Papanicolaou, *Matching pursuit for imaging high contrast conductivity*, Inverse Problems, 15 (1999), pp. 811–849.

[10] L. Borcea and G. Papanicolaou, *Network approximation for transport properties of high contrast materials*, SIAM J. Appl. Math., 58 (1998), pp. 501–539.

[11] L. Borcea and G. Papanicolaou, *Low frequency electromagnetic fields in high contrast media*, in Surveys on Solution Methods for Inverse Problems, D. Colton, H. W. Engl, A. Louis, J. R. McLaughlin, and W. Rundell, eds., Springer, New York, 2000, pp. 195–233.

[12] J. F. Brady, *The rheological behavior of concentrated colloidal suspensions*, J. Chem. Phys., 99 (1993), pp. 567–581.

[13] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. I, John Wiley, New York, 1953.

[14] A. Einstein, *Eine neue Bestimmung der Moleküldimensionen*, Ann. Phys., 19 (1906), p. 289, and 34 (1906), p. 591.

[15] N. A. Frankel and A. Akrivos, *On the viscosity of a concentrated suspension of solid spheres*, Chem. Engrg. Sci., 22 (1967), pp. 847–853.

[16] A. L. Graham, *On the viscosity of suspension of solid spheres*, Appl. Sci. Res., 37 (1981), pp. 275–286.

[17] B. I. Halperin, *Remarks on percolation and transport in networks with a wide range of bond strengths*, Phys. D, 38 (1989), pp. 179–183.

[18] L. Hörmander, *Analysis of Linear Partial Differential Operators*, Vol. 1, Springer, New York, 1983.

[19] D. J. Jeffrey and A. Acrivos, *The rheological properties of suspensions of rigid particles*, American Institute of Chemical Engineering, 22 (1976), pp. 417–432.

[20] J. B. Keller, *Conductivity of a medium containing a dense array of perfectly conducting spheres or cylinders or nonconducting cylinders*, J. Appl. Phys., 43 (1963), pp. 991–993.

[21] J. Koplik, *Creeping flow in two-dimensional networks*, J. Fluid Mech., 119 (1982), pp. 219–247.

[22] S. M. Kozlov, *Geometric aspects of averaging*, Russian Math. Surveys, 44 (1989), pp. 91–144.

[23] L. D. Landau and E. M. Lifshitz, *Fluid Mechanics*, Pergamon Press, New York, 1987.

[24] R. Larsen, *The Structure and Rheology of Complex Fluids*, Oxford University Press, New York, 1999.

[25] T. Levy and E. Sanchez-Palencia, *Suspension of solid particles in a Newtonian fluid*, J. Non-Newtonian Fluid Mech., 13 (1983), pp. 63–78.

[26] E. E. Moise, *Geometric Topology in Dimensions* 2 *and* 3, Springer, New York, 1977.

[27] K. C. Nunan and J. B. Keller, *Effective viscosity of a periodic suspension*, J. Fluid Mech., 142 (1984), pp. 269–287.

[28] L. M. Schwartz, J. R. Banavar, and B. I. Halperin, *Biased-diffusion calculations of effective transport in inhomogeneous continuum systems*, Phys. Rev. B, 3 (1989), pp. 9155–9161.

[29] C. B. Shah and Y. C. Yortsos, *The permeability of strongly disordered systems*, Phys.

Fluids, 8 (1996), pp. 280–282.
[30] C. A. Shook and M. C. Rocko, *Slurry Flow. Principles and Practice*, Butterworth-Heinemann, New York, 1991.
[31] A. Sierou and J. F. Brady, *Accelerated Stokesian dynamic simulations*, J. Fluid Mech., 448 (2001), pp. 115–146.
[32] B. Veytsman, J. Morrison, and A. Scaroni, *The packing viscosity of concentrated polydisperse coal/water slurries*, Energy Fuels, 12 (1998), pp. 1031–1039.

SIAM J. MATH. ANAL.
Vol. 36, No. 5, pp. 1629–1651

DYNAMICAL SCALING IN SMOLUCHOWSKI'S COAGULATION EQUATIONS: UNIFORM CONVERGENCE*

GOVIND MENON† AND ROBERT L. PEGO‡

Abstract. We consider the approach to self-similarity (or dynamical scaling) in Smoluchowski's coagulation equations for the solvable kernels $K(x,y) = 2, x+y$ and xy. We prove the uniform convergence of densities to the self-similar solution with exponential tails under the regularity hypothesis that a suitable moment have an integrable Fourier transform. For the discrete equations we prove uniform convergence under optimal moment hypotheses. Our results are completely analogous to classical local convergence theorems for the normal law in probability theory. The proofs rely on the Fourier inversion formula and the solution by the method of characteristics for the Laplace transform.

Key words. coagulation equations, dynamic scaling, self-similar solutions, complex characteristics

AMS subject classifications. 70F99, 82C28, 45M10, 35L65, 35Q99

DOI. 10.1137/S0036141003430263

1. Introduction.

Smoluchowski's coagulation equation

$$\partial_t n(t,x) = \frac{1}{2}\int_0^x K(x-y,y)n(t,x-y)n(t,y)dy - \int_0^\infty K(x,y)n(t,x)n(t,y)dy \tag{1.1}$$

is a widely studied mean-field model for cluster growth [4, 8, 17]. We study the evolution of $n(t,x)$, the number of clusters of mass x per unit volume at time t, which coalesce by binary collisions with a symmetric rate kernel $K(x,y)$. Equation (1.1) has been used as a model of cluster growth in a surprisingly diverse range of fields such as physical chemistry, astrophysics, and population dynamics (see [4] for a review of applications). In addition, over the past few years a rich mathematical theory has been developed for these equations. Aldous [1] provides an excellent introduction.

Many kernels in applications are homogeneous; that is, $K(\alpha x, \alpha y) = \alpha^\gamma K(x,y)$, $x, y, \alpha > 0$, for some exponent γ [4]. A mathematical problem of scientific interest is to study self-similar or dynamical scaling behavior for homogeneous kernels. There are no general mathematical results for this problem despite an extensive scientific literature (especially formal asymptotics and numerics [12, 13, 18]). It is known that γ plays a crucial role. On physical grounds, we expect solutions to (1.1) to conserve the total mass $\int_0^\infty xn(t,x)dx$. When $K(x,y) \le 1+x+y$ (corresponding to $0 \le \gamma \le 1$), mass-conserving solutions exist globally in time under suitable moment hypotheses on initial data [5]. It is then typical in applications to assert that the solutions approach "scaling form" [13, 18], but there is no rigorous mathematical justification for this in general.

*Received by the editors June 24, 2003; accepted for publication (in revised form) July 23, 2004; published electronically April 29, 2005. This material is based upon work supported by the National Science Foundation under grants DMS 00-72609 and DMS 03-05985.
http://www.siam.org/journals/sima/36-5/43026.html
†Department of Mathematics, University of Wisconsin, Madison, WI 53706. Current address: Division of Applied Mathematics, Brown University, Providence, RI 02912 (menon@dam.brown.edu).
‡Department of Mathematics and Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742. Current address: Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213 (rpego@cmu.edu).

For a large class of kernels satisfying $(xy)^{\gamma/2} \leq K(x,y)$ with $1 < \gamma < 2$, it is known that there is no solution that preserves mass for all time. This breakdown phenomenon is known as *gelation*. It was first demonstrated by McLeod [14] with an explicit solution for the kernel $K = xy$. A general result using only the growth of the kernel was proved probabilistically by Jeon [9] (see also [6] for a simple analytical proof). It is natural to ask whether the blow-up is self-similar, but there are no general results on this problem yet.

There are a number of results, however, for the "solvable" kernels $K = 2$, $x + y$, and xy (see [15] and references therein; also see [13]). A remarkable feature of these kernels is that the problem of dynamical scaling can be understood quite deeply by analogy with classical limit theorems in probability theory. For example, an analogue to the classical Lévy–Khintchine representation for infinitely divisible laws was proved by Bertoin [2] for eternal solutions to Smoluchowski's equation with kernel $K = x + y$. Eternal solutions are defined for all $t \in (-\infty, \infty)$, meaning that they model coagulation processes "infinitely divisible" under Smoluchowski dynamics. Later, we proved [15] that the domains of attraction of self-similar solutions (in the sense of weak convergence of measures) can be characterized by almost power-law behavior of the tails of the initial size distribution. This is analogous to the characterization of the weak domains of attraction of the Lévy stable laws [7]. An essential component in both proofs is a simple solution formula for the Laplace transform of n that is widely known [4]. These results may be used as a basis for refined convergence theorems, as we now explain.

A general theme in probabilistic limit theorems is the interplay between moment and regularity hypotheses and the topology of convergence. In this article, we develop one aspect of this idea. Under stronger regularity hypotheses, the weak convergence results of [15] will be strengthened to obtain uniform convergence of densities using the Fourier transform. This method is classical in probability theory and is used to prove uniform convergence of densities in the central limit theorem [7, Theorem XV.5.2]. Feller's argument in [7] is simple and robust, and our main contribution is to show that it extends naturally to Smoluchowski's equation. The key new idea is to use the method of characteristics in the right half of the complex plane to obtain strong decay estimates on the Laplace transform. A broader contribution of this work and [15] is to show that the analytical methods used to prove classical limit theorems in probability apply to a wider range of problems involving scaling phenomenon for integral equations of convolution type.

Let us briefly connect our results to previous work: the only uniform convergence theorems in the literature are that of Kreer and Penrose for the kernel $K = 2$ [11] and closely connected work of da Costa [3]. In this article, for $K = 2$ and $x + y$ we present theorems on uniform convergence to the self-similar solutions with exponential tails for the continuous and discrete Smoluchowski equations. For $K = xy$, we prove uniform convergence of densities to self-similar form as t approaches the gelation time T_{gel}. For $K = 2$, we strengthen the result of Kreer and Penrose and simplify the proof. Their decay hypothesis on the initial data ($n_0(x) \leq Ce^{-ax}$) is weakened to an (almost) optimal moment hypothesis, and their regularity hypothesis ($n_0 \in C^2$) is weakened to a little bit more than continuity. For $K = x+y$ the convergence theorem is new. Study of the kernel $K = xy$ is reduced to $K = x+y$ by a well-known change of variables [4]. Uniform convergence to the self-similar solutions with "fat" or "heavy" tails is a more delicate issue, which will not be considered here.

Our uniform convergence theorems may be stated in a unified manner as follows

for the continuous Smoluchowski equations with kernels $K(x,y) = 2$, $x+y$, and xy, corresponding to $\gamma = 0, 1, 2$, respectively. Presuming the γth and $(\gamma+1)$st moments are finite, we may scale x and n so both moments are initially 1. For the multiplicative kernel this ensures that the gelation time $T_{\rm gel} = 1$. Let $T_\gamma = \infty$ for $\gamma = 0, 1$ and $T_\gamma = T_{\rm gel} = 1$ for $\gamma = 2$. The self-similar solutions with exponential tails are explicitly given by [15]

$$n(t,x) = \frac{m_\gamma(t)}{\lambda_\gamma(t)^{\gamma+1}}\,\hat{n}_{*,\gamma}\left(\frac{x}{\lambda_\gamma(t)}\right), \tag{1.2}$$

where, for $\hat{x} \geq 0$,

$$\hat{n}_{*,0}(\hat{x}) = e^{-\hat{x}}, \qquad \hat{x}\hat{n}_{*,1}(\hat{x}) = \hat{x}^2\hat{n}_{*,2}(\hat{x}) = \frac{1}{\sqrt{2\pi}}\hat{x}^{-1/2}e^{-\hat{x}/2}, \tag{1.3}$$

and

$$m_0(t) = t^{-1}, \quad m_1(t) = 1, \quad m_2(t) = (1-t)^{-1}, \tag{1.4}$$

$$\lambda_0(t) = t, \quad \lambda_1(t) = e^{2t}, \quad \lambda_2(t) = (1-t)^{-2}. \tag{1.5}$$

Our sufficient conditions for uniform convergence to these self-similar solutions for the continuous Smoluchowski equations are summarized by the following result.

THEOREM 1.1. *Let $n_0 \geq 0$, $\int_0^\infty x^\gamma n_0(x)dx = \int_0^\infty x^{1+\gamma}n_0(x)dx = 1$. Assume that the Fourier transform of $x^{1+\gamma}n_0$ is integrable, and let $n(t,x)$ be the solution to Smoluchowski's equation with initial data $n_0(x)$ and $K = 2$, $x+y$ or xy, for $\gamma = 0, 1$, or 2. Then the rescaled solution*

$$\hat{n}(t,\hat{x}) = \frac{\lambda_\gamma(t)^{1+\gamma}}{m_\gamma(t)}n(t,\hat{x}\lambda_\gamma(t))$$

satisfies

$$\lim_{t\to T_\gamma}\sup_{\hat{x}>0}\hat{x}^{1+\gamma}\,|\hat{n}(t,\hat{x}) - \hat{n}_{*,\gamma}(\hat{x})| = 0.$$

It has been traditional to treat the discrete Smoluchowski equations separately from the continuous equations. Yet, within the framework of measure valued solutions [15, 16], the discrete Smoluchowski equations simply correspond to the special case of a lattice distribution, a measure valued solution supported on the lattice $h\mathbb{N}$ and taking the form $\nu_t = \sum_{l=1}^\infty n_l(t)\delta_{hl}(x)$, where $\delta_{hl}(x)$ is a Dirac delta at hl. If h is maximal we call ν_t a lattice measure with *span* h. The coefficients n_l satisfy the discrete Smoluchowski equations

$$\partial_t n_l(t) = \frac{1}{2}\sum_{j=1}^{l-1}\kappa_{l-j,j}n_{l-j}(t)n_j(t) - \sum_{j=1}^{\infty}\kappa_{l,j}n_l(t)n_j(t), \tag{1.6}$$

where $\kappa_{l,j} = K(lh, jh)$. Physically, this case is of importance, since some mass aggregation processes (e.g., polymerization) have a fundamental unit of mass (e.g., a monomer). The uniform convergence theorems for the continuous Smoluchowski equations have a natural extension to this case.

THEOREM 1.2. *Let $\nu_0 \geq 0$ be a lattice measure with span h such that $\int_0^\infty x^\gamma\nu_0(dx) = \int_0^\infty x^{1+\gamma}\nu_0(dx) = 1$. Then with*

$$\hat{l} = \frac{lh}{\lambda_\gamma(t)}, \quad \hat{n}_l(t) = \frac{1}{h}\frac{\lambda_\gamma(t)^{1+\gamma}}{m_\gamma(t)}n_l(t),$$

we have

$$\lim_{t\to T_\gamma} \sup_{l\in\mathbb{N}} \hat{l}^{1+\gamma} \left|\hat{n}_l(t) - \hat{n}_{*,\gamma}(\hat{l})\right| = 0.$$

Let us comment on the hypotheses in Theorems 1.1 and 1.2. The moment hypotheses in both theorems are essentially the same. $\int_0^\infty x^\gamma \nu_0(dx) = 1$ is the natural hypothesis for existence and uniqueness of solutions [15]. The other moment condition $\int_0^\infty x^{1+\gamma}\nu_0(dx) = 1$ is of a different character. It implies that n_0 or ν_0 is in the weak domain of attraction of the self-similar solution with exponential tail, under a rescaling $n(t,x) \longrightarrow \hat{n}(\hat{t},\hat{x})$ that fixes both moments

$$\int_0^\infty \hat{x}^\gamma \hat{n}(\hat{t},\hat{x})d\hat{x} = \int_0^\infty \hat{x}^{\gamma+1}\hat{n}(\hat{t},\hat{x})d\hat{x} = 1 \quad \text{for all } \hat{t} \geq 0.$$

The hypothesis that the $(\gamma+1)$st moment is finite is almost optimal. The weak domain of attraction under a broader class of rescalings is a bit bigger, as it allows for a weak divergence $\int_0^y x^{1+\gamma}\nu_0(dx) \sim L(y)$ as $y \to \infty$ for a slowly varying function $L(y)$ [15]. Thus, Theorem 1.2 shows that within the class of lattice measures, the weak convergence of measures almost implies uniform convergence of the coefficients.

Theorem 1.1 requires an additional hypothesis on integrability of a suitable Fourier transform. This is a regularity hypothesis that is the analogue of the hypothesis for uniform convergence to the normal law used by Feller [7]. One may heuristically understand the role of regularity as follows. Equation (1.1) is hyperbolic and discontinuities in the initial data persist for all finite times. On the other hand, the self-similar solutions in (1.3) are analytic. Thus, one expects that some regularity on the initial data is necessary to obtain uniform convergence to a self-similar solution. Loosely speaking, regularity of the initial data $n_0(x)$ translates into a decay hypothesis on its Fourier transform. We need only the weak decay implied by integrability.

We do not know if this assumption is optimal, or if it may be weakened further. We briefly comment on this issue here; it will not be considered in the rest of the paper. The space of functions with integrable Fourier transforms is of great interest in harmonic analysis. Precisely, for $f \in L^1(\mathbb{R})$, let F be its Fourier transform. Then the space

$$A(\mathbb{R}) = \{f \in L^1(\mathbb{R}) | F \in L^1(\mathbb{R})\}$$

is a closed subalgebra of $L^1(\mathbb{R})$ known as the Wiener algebra [10]. Integrability of F implies that f is continuous. But it also implies more. It is known that functions in $A(\mathbb{R})$ possess some delicate regularity properties. For example, a function in $A(\mathbb{R})$ has a logarithmic modulus of continuity in a neighborhood where it is monotonic. It is definitely not obvious whether this regularity is truly necessary to obtain uniform convergence. If $v_0(ik) = \int_0^\infty e^{-ikx}x^{1+\gamma}n_0(x)dx$ is integrable, it also follows that $v_0 \in H^1(\mathbb{R}) \cap A(\mathbb{R})$, since v_0 is the boundary limit of an analytic function (the Laplace transform of $x^{1+\gamma}n_0$). Here H^1 denotes the classical Hardy space. This is turn means that v_0 has some hidden regularity and integrability properties. It is worth remarking that the precise characterization of $A(\mathbb{R})$ remains an outstanding open problem in harmonic analysis (though several sufficient conditions are known; see [10]).

2. Uniform convergence of densities for the constant kernel $K = 2$.

2.1. Evolution of the Laplace transform. Let $\mathbb{C}_+ = \{z \in \mathbb{C} \mid \operatorname{Re} z > 0\}$ and $\bar{\mathbb{C}}_+ = \{z \in \mathbb{C} \mid \operatorname{Re} z \geq 0\}$. We let

$$N(t,z) = \int_0^\infty e^{-zx} n(t,x)\,dx, \quad z \in \bar{\mathbb{C}}_+,$$

denote the Laplace transform of the number density n. We take the Laplace transform of (1.1) with $K = 2$, and its limit as $z \to 0$, to see that $N(t,z)$ solves

$$\partial_t N = N^2 - 2N(t,0)N, \quad \partial_t N(t,0) = -N(t,0)^2. \tag{2.1}$$

Without loss of generality, we may suppose that the initial time $t = 1$. We will always assume that the initial data is normalized such that

$$\int_0^\infty n(1,x)\,dx = \int_0^\infty x n(1,x)\,dx = 1. \tag{2.2}$$

If the initial number of clusters, $\int_0^\infty n(1,x)dx$, and the mass, $\int_0^\infty xn(1,x)dx$, are finite, we may always assume that (2.2) holds after rescaling x and n. We solve the second equation in (2.1) to see that the total number of clusters decreases according to

$$\int_0^\infty n(t,x)\,dx = N(t,0) = t^{-1}, \quad t \geq 1. \tag{2.3}$$

We hold z fixed and integrate (2.1) in t to obtain the solution

$$N(t,z) = \frac{1}{t}\frac{N(1,z)}{t(1-N(1,z)) + N(1,z)}. \tag{2.4}$$

The evolution preserves mass. Indeed, if we differentiate (2.4) with respect to z, we find

$$\int_0^\infty xn(t,x)\,dx = -\partial_z N(t,0) = -\partial_z N(1,0) = \int_0^\infty xn(1,x)\,dx = 1. \tag{2.5}$$

2.2. Approach to self-similarity. A special case of the weak convergence result of [15], also given by Leyvraz [13], is obtained as follows: Observe that for each fixed $s \in \bar{\mathbb{C}}_+$, equations (2.3), (2.4), and (2.5) imply

$$tN(t,st^{-1}) = \frac{N(1,st^{-1})}{t(1-N(1,st^{-1})) + N(1,st^{-1})} \underset{t\to\infty}{\longrightarrow} \frac{1}{1+s}. \tag{2.6}$$

It is classical that the pointwise convergence of Laplace transforms is equivalent to weak convergence of measures [7, Theorem XIII.1.2a]. Thus, (2.6) implies that rescaled solutions to Smoluchowski's equations converge weakly. Let us be more precise about the rescaling. We define the similarity variables

$$\tau = \log t, \quad \hat{x} = \frac{x}{t} = e^{-\tau}x, \quad s = tz = e^\tau z \tag{2.7}$$

and the rescaled number distribution

$$\hat{n}(\tau,\hat{x}) = e^{2\tau} n(e^\tau, e^\tau \hat{x}) = t^2 n(t,x). \tag{2.8}$$

Observe that this rescaling preserves *both* total number and mass, that is,

$$\int_0^\infty \hat{n}(\tau,\hat{x})d\hat{x} = \int_0^\infty \hat{x}\hat{n}(\tau,\hat{x})d\hat{x} = 1, \quad \tau \geq 0. \tag{2.9}$$

We denote the Laplace transform of $\hat{n}(\tau,\hat{x})$ by

$$u(\tau,s) = \int_0^\infty e^{-s\hat{x}}\hat{n}(\tau,\hat{x})\,d\hat{x} = e^\tau N(e^\tau, se^{-\tau}) = tN(t,z). \tag{2.10}$$

In these variables, the pointwise convergence of (2.6) takes the simple form

$$\lim_{\tau\to\infty} u(\tau,s) = \frac{1}{1+s} =: u_{*,0}(s), \quad s \in \bar{\mathbb{C}}_+, \tag{2.11}$$

where $u_{*,0}(s)$ denotes the Laplace transform of

$$\hat{n}_{*,0}(\hat{x}) = e^{-\hat{x}}, \quad \hat{x} \geq 0, \tag{2.12}$$

the profile for the self-similar solution in (1.2). Now, (2.11) is equivalent to

$$\hat{n}(\tau,\hat{x})\,d\hat{x} \to \hat{n}_{*,0}(\hat{x})\,d\hat{x}$$

as $\tau \to \infty$, in the sense of weak convergence of measures.

Our goal is to strengthen this to uniform convergence in both continuous and discrete cases, under appropriate hypotheses on initial data. For the continuous Smoluwchowski equation (1.1) we prove the following theorem.

THEOREM 2.1. *Let $n(1,x) \geq 0$, $\int_0^\infty n(1,x)\,dx = \int_0^\infty xn(1,x)\,dx = 1$. Assume that the Fourier transform of $xn(1,x)$ is integrable. Then in terms of the rescaling in* (2.7)–(2.8) *we have*

$$\lim_{\tau\to\infty} \sup_{\hat{x}>0} \hat{x}|\hat{n}(\tau,\hat{x}) - \hat{n}_{*,0}(\hat{x})| = 0, \tag{2.13}$$

where $\hat{n}_{*,0}(\hat{x}) = e^{-\hat{x}}$ is the similarity profile in (2.12).

The proof of this theorem extends to treat uniform convergence of coefficients for solutions of the discrete equations (1.6) under only the hypothesis that the zeroth and first moments are finite; see Theorem 2.2 below.

Observe that we prove uniform convergence of the weighted densities $\hat{x}\hat{n}(\tau,\hat{x})$. The reason can be ascribed to use of the Fourier–Laplace inversion formula. We cannot apply the inversion formula directly to $u_{*,0}$ as it is not integrable on the imaginary axis ($|u_{*,0}(ik)| \sim |k|^{-1}$ as $|k| \to \infty$). The slow decay of the Fourier transform is caused by the jump discontinuity at $x = 0$, since $\hat{n}_{*,0}(x) = 0$ for $x < 0$. In order to gain a uniform convergence result, we smooth this discontinuity and consider the mass density $\hat{x}\hat{n}$. Its Laplace transform we denote by

$$v(\tau,s) = -\partial_s u(\tau,s) = \int_0^\infty e^{-s\hat{x}}\hat{x}\hat{n}(\tau,\hat{x})\,d\hat{x}. \tag{2.14}$$

Differentiating (2.11), we obtain a corresponding self-similar profile, with

$$v_{*,0}(s) := \frac{1}{(1+s)^2}, \qquad |v_{*,0}(ik)| = \frac{1}{1+k^2},\ k \in \mathbb{R}. \tag{2.15}$$

2.3. Evolution on characteristics. The explicit solution for $u(\tau, s)$ and $v(\tau, s)$ can be obtained directly by substituting (2.10) into (2.4). But we rederive the solution to make explicit the geometric idea underlying the proof of Theorem 2.1. The same ideas underlie the proof of Theorem 3.1 for the additive kernel and are more easily understood here. We use the change of variables (2.7) and (2.10) in (2.1), and the conservation of moments in (2.9), to obtain the equation of evolution for u:

$$\partial_\tau u + s\partial_s u = -u(1-u). \tag{2.16}$$

The solution of (2.16) may be described by the method of characteristics. A characteristic curve $s(\tau; \tau_0, s_0)$ is the solution to

$$\frac{ds}{d\tau} = s, \quad s(\tau; \tau_0, s_0) = s_0 \in \bar{\mathbb{C}}_+. \tag{2.17}$$

Explicitly,

$$s(\tau; \tau_0, s_0) = e^{\tau - \tau_0} s_0. \tag{2.18}$$

Equation (2.17) is an autonomous differential equation in $\bar{\mathbb{C}}_+$ and may be thought of geometrically. For fixed $s_0 \in \bar{\mathbb{C}}_+$ the trajectory of the characteristic curve $s(\tau; \tau_0, s_0), \tau \in \mathbb{R}$, is a ray in $\bar{\mathbb{C}}_+$ emanating from the origin. In particular, the imaginary axis is invariant under the flow of (2.17). Equation (2.18) shows that the characteristics expand uniformly outward at the rate e^τ. Along characteristics we have

$$\frac{du}{d\tau} = -u(1-u), \tag{2.19}$$

which may be integrated to obtain the solution

$$u(\tau, s) = \frac{u(\tau_0, s_0)e^{-(\tau-\tau_0)}}{1 - u(\tau_0, s_0)(1 - e^{-(\tau-\tau_0)})}. \tag{2.20}$$

We need to estimate the decay of the derivative $v = -\partial_s u$. Differentiating (2.16), we see that on characteristics the derivative solves

$$\frac{dv}{d\tau} = -2(1-u)v. \tag{2.21}$$

We integrate (2.21) using (2.20) to find

$$v(\tau, s) = \frac{v(\tau_0, s_0)e^{-2(\tau-\tau_0)}}{\big(1 - u(\tau_0, s_0)(1 - e^{-(\tau-\tau_0)})\big)^2}. \tag{2.22}$$

For $\tau \geq \tau_0$ we may take absolute values in (2.20) and (2.22) to obtain the decay estimates

$$|u(\tau, s)| \leq \frac{|u(\tau_0, s_0)|e^{-(\tau-\tau_0)}}{1 - |u(\tau_0, s_0)|(1 - e^{-(\tau-\tau_0)})} \tag{2.23}$$

and

$$|v(\tau, s)| \leq \frac{|v(\tau_0, s_0)|e^{-2(\tau-\tau_0)}}{\big(1 - |u(\tau_0, s_0)|(1 - e^{-(\tau-\tau_0)})\big)^2} \leq \frac{|v(\tau_0, s_0)|e^{-2(\tau-\tau_0)}}{\big(1 - |u(\tau_0, s_0)|\big)^2}. \tag{2.24}$$

2.4. Proof of Theorem 2.1. 1. We use the Fourier–Laplace inversion formula

$$\hat{x}(\hat{n}(\tau,\hat{x}) - \hat{n}_{*,0}(\hat{x})) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{ik\hat{x}}\left(v(\tau,ik) - v_{*,0}(ik)\right)\,dk. \tag{2.25}$$

Thus, in order to prove (2.13) it suffices to show

$$\lim_{\tau\to\infty}\int_{\mathbb{R}} |v(\tau,ik) - v_{*,0}(ik)|\,dk = 0. \tag{2.26}$$

2. Let $\varepsilon \in (0,\frac{1}{2})$ and put $R = \varepsilon^{-1}$. We will prove (2.26) by estimating the integral separately in three regions: $|k| \le R$, $R \le |k| \le Re^{\tau-T}$, and $Re^{\tau-T} \le |k|$ for $\tau \ge T$, where $T > 0$ will be chosen sufficiently large, depending on ε and the initial data v_0. This is essentially the same decomposition used in the proof of uniform convergence in the central limit theorem by Feller [7, Theorem XV.5.2]. The main new idea here is the use of the decay estimates (2.24) and the method of characteristics in the regions where $R \le |k|$.

3. $|k| \le R$: Recall that the pointwise convergence of Laplace transforms (2.11) is equivalent to $\hat{n}(\tau,\hat{x})\,d\hat{x} \to \hat{n}_{*,0}(\hat{x})\,d\hat{x}$ in the sense of weak convergence of measures. Combined with (2.9) this also implies that the mass measures $\hat{x}\hat{n}(\tau,\hat{x})\,d\hat{x}$ converge weakly to $\hat{x}\hat{n}_{*,0}(\hat{x})\,d\hat{x}$ as $\tau \to \infty$. But this implies $v(\tau,ik)$ converges to $v_{*,0}(ik)$ uniformly for $|k| \le R$ [7, Theorem XV.3.2]. Therefore,

$$\lim_{\tau\to\infty}\int_{-R}^{R} |v(\tau,ik) - v_{*,0}(ik)|\,dk = 0. \tag{2.27}$$

4. It remains to consider $|k| \ge R$. It is sufficient to consider only $k \ge R$, since $|v(\tau,ik)| = |v(\tau,-ik)|$. We will control $v(\tau,ik)$ and $v_{*,0}$ separately:

$$\int_R^\infty |v(\tau,ik) - v_{*,0}(ik)|\,dk \le \int_R^\infty |v(\tau,ik)|\,dk + \int_R^\infty |v_{*,0}(ik)|\,dk.$$

But $|v_{*,0}(ik)| = (1+|k|^2)^{-1}$ by (2.15), so that

$$\int_R^\infty |v_{*,0}(ik)|\,dk \le R^{-1} = \varepsilon.$$

In the rest of the proof we estimate $\int_R^\infty |v(\tau,ik)|dk$.

5. Since $u(\tau,ik) \to u_{*,0}(ik)$ and $v(\tau,ik) \to v_{*,0}$ as $\tau \to \infty$ for each real k, using (2.11) and (2.15) we may choose $T > 0$ such that

$$\sup_{\tau\ge T} |u(\tau,iR)| \le R^{-1} = \varepsilon, \quad \sup_{\tau\ge T} |v(\tau,iR)| \le R^{-2}. \tag{2.28}$$

6. $R \le k \le Re^{\tau-T}$: The control obtained from (2.28) propagates outward along characteristics as τ increases. Precisely, whenever $\tau \ge T$, for any k such that $R \le k \le Re^{\tau-T}$ we have $ik = e^{\tau-\tau_0}iR$, where $\tau_0 \ge T$. By (2.18) this means that $ik = s(\tau;\tau_0,s_0)$, with $s_0 = iR$. Then the decay estimate (2.24) and the boundary control (2.28) imply

$$|v(\tau,ik)| \le \frac{|v(\tau_0,iR)|e^{-2(\tau-\tau_0)}}{\left(1-|u(\tau_0,iR)|\right)^2} \le \frac{1}{(1-\varepsilon)^2}R^{-2}\left(\frac{R}{k}\right)^2 \le 4k^{-2}. \tag{2.29}$$

Integrating this estimate we obtain

$$\int_R^{Re^{\tau-T}} |v(\tau,ik)|\,dk \le \int_R^\infty 4k^{-2}\,dk = 4R^{-1} = 4\varepsilon.$$

7. $Re^{\tau-T} \le k$: For brevity, let $\tilde{R} = Re^{-T}$. With $u_0(s) := u(0,s)$, $v_0(s) := v(0,s)$, we use (2.24) and (2.18) with $\tau_0 = 0$ to obtain

$$\int_{\tilde{R}e^\tau}^\infty |v(\tau,ik)|\,dk \le e^{-2\tau}\int_{\tilde{R}e^\tau}^\infty \frac{|v_0(ike^{-\tau})|}{\left(1-|u_0(ike^{-\tau})|\right)^2}dk$$
$$= e^{-\tau}\int_{\tilde{R}}^\infty \frac{|v_0(ik')|}{\left(1-|u_0(ik')|\right)^2}dk' \le \left(\sup_{|k'|\ge\tilde{R}} \frac{1}{\left(1-|u_0(ik')|\right)^2}\right) e^{-\tau}\|v_0\|_{L^1},$$

where $k' = ke^{-\tau}$. Since $|u_0(ik')| < 1$ for $k' \ne 0$ and $u_0(ik') \to 0$ as $k \to \infty$ by the Riemann–Lebesgue lemma, we have $\sup_{|k'|\ge\tilde{R}}(1-|u_0(ik')|)^{-2} < \infty$.

8. Putting together the estimates we have obtained, it follows that for τ sufficiently large, the integral in (2.26) is less than 12ε. This completes the proof.

2.5. The discrete Smoluchowski equations. We consider measure solutions of the form $\nu_t = \sum_{l=1}^\infty n_l(t)\delta_{hl}(x)$, where $\delta_{hl}(x)$ denotes a Dirac mass at hl. To avoid redundancy, we always assume that h is the *span* of the lattice, that is, the maximal $h > 0$ so that all initial clusters, and thus clusters at any time $t > 0$, are concentrated on $h\mathbb{N}$. We will call ν_t a lattice measure with span h. Notice that if the initial number of clusters and the mass are finite, by rescaling n_l and h we may assume that $\int_0^\infty \nu_1(dx) = \int_0^\infty x\nu_1(dx) = 1$. Under these conditions, the weak convergence theorem of [15] asserts that $\lim_{t\to\infty} tN(t,s/t) = u_{*,0}(s)$. We show that this theorem may be strengthened by use of Fourier series. The Fourier transform of ν_t is the Fourier series

$$N(t,ik) = \sum_{l\in\mathbb{N}} n_l(t)e^{-ilhk}, \qquad k \in \mathbb{R},$$

which has minimal period $2\pi/h$. Thus $n_l(t) = (h/2\pi)\int_{-\pi/h}^{\pi/h} e^{ilhk}N(t,ik)\,dk$, or

$$t^2 n_l(t) = \frac{h}{2\pi}\int_{-\pi e^\tau/h}^{\pi e^\tau/h} \exp(ilhke^{-\tau})u(\tau,ik)\,dk, \tag{2.30}$$

in similarity variables from (2.10). We integrate by parts and let

$$\hat{l} = lhe^{-\tau} = lht^{-1}, \quad \hat{n}_l(t) = h^{-1}t^2 n_l(t) \tag{2.31}$$

to obtain

$$\hat{l}\hat{n}_l(t) = tln_l(t) = \frac{1}{2\pi}\int_{-\pi e^\tau/h}^{\pi e^\tau/h} e^{i\hat{l}k}v(\tau,ik)\,dk. \tag{2.32}$$

As in Theorem 2.1, we expect the right-hand side to converge to $\hat{l}\hat{n}_{*,0}(\hat{l})$ as $\tau \to \infty$, indeed uniformly for $\hat{l} \in ht^{-1}\mathbb{N}$.

THEOREM 2.2. *Let $\nu_1 \ge 0$ be a lattice measure with span h such that $\int_0^\infty \nu_1(dx) = \int_0^\infty x\nu_1(dx) = 1$. Then with the scaling* (2.31) *we have*

$$\lim_{t\to\infty}\sup_{l\in\mathbb{N}} \hat{l}\left|\hat{n}_l(t) - \hat{n}_{*,0}(\hat{l})\right| = 0. \tag{2.33}$$

Proof. By (2.32) and the continuous Fourier inversion formulas, it suffices to show that

$$\lim_{\tau\to\infty}\sup_{\hat{l}\geq 0}\left|\int_{-\pi e^{\tau}/h}^{\pi e^{\tau}/h} e^{i\hat{l}k}v(\tau,ik)\,dk-\int_{\mathbb{R}} e^{i\hat{l}k}v_{*,0}(ik)\,dk\right|=0.$$

As earlier, it suffices to consider $k>0$. The integrals

$$\int_{-R}^{R}|v(\tau,ik)-v_{*,0}(ik)|\,dk,\quad \int_{R}^{\tilde{R}e^{\tau}}|v(\tau,ik)|\,dk,\quad \int_{R}^{\infty}|v_{*,0}(ik)|\,dk,$$

with $\tilde{R}=Re^{-T}$, are controlled exactly as in the proof of Theorem 2.1. It only remains to estimate the integral of $|v(\tau,ik)|$ over the region $\tilde{R}e^{\tau}<k<\pi e^{\tau}/h$. We assume that $\pi/h>\tilde{R}$, for otherwise there is nothing to prove. But then by the formula (2.18), the uniform decay estimate (2.24), and the change of variables $k'=ke^{-\tau}$, we have

$$\int_{\tilde{R}e^{\tau}}^{\pi e^{\tau}/h}|v(\tau,ik)|\,dk\leq e^{-\tau}\int_{\tilde{R}}^{\pi/h}\frac{|v_0(ik')|}{|1-u_0(ik')(1-e^{-\tau})|^2}dk'.$$

Since the domain of integration is finite, it suffices to show that the integrand is uniformly bounded in time. Since $|v_0(ik)|\leq 1$, it is only necessary to control the denominator. But $u_0(ik)=\sum_{l\in\mathbb{N}}n_l(0)e^{-ilkh}$ with $n_l(0)\geq 0$. Therefore, $|u_0(ik)|\leq 1$, and [7, Lemma XV.1.4] yields that

$$u_0(ik)=1\quad\text{if and only if}\quad k=\frac{2\pi m}{h},\qquad m\in\mathbb{Z}.$$

In particular, we have the strict inequality

$$\min_{k\in[\tilde{R},\frac{\pi}{h}]}|1-u_0(ik)|\geq\delta>0.$$

Therefore,

$$\left|1-u_0(ik)(1-e^{-\tau})\right|\geq|1-u_0(ik)|-|u_0(ik)|e^{-\tau}\geq\delta-e^{-\tau}\geq\frac{\delta}{2}$$

for sufficiently large τ. Thus,

$$\int_{\tilde{R}e^{\tau}}^{\pi e^{\tau}/h}|v(\tau,ik)|\,dk\leq\frac{2\pi}{\delta h}e^{-\tau}.\qquad\square$$

3. Uniform convergence of densities for the additive kernel.

3.1. Rescaling and approach to self-similarity. In this section we prove the analogues of Theorems 2.1 and 2.2 for the additive kernel. The essential geometric ideas of the proof are similar to the previous section. However, the trajectories of the characteristic curves $s(t;t_0,s_0)$ in the complex plane are no longer rays, and the proofs require more careful analysis. As earlier, we will work with the explicit solution formula for an appropriate Laplace transform. For $z\in\bar{\mathbb{C}}_+$ we define

$$\Phi(t,z)=\int_0^{\infty}\left(1-e^{-zx}\right)n(t,x)dx. \tag{3.1}$$

We observe that $1-e^{-zx} = zx+O(z^2x^2)$ as $x \to 0$. We use Φ instead of the standard Laplace transform of n because the latter may not be well defined: for example, the similarity profile $\hat{n}_{*,1}$ in (1.3) satisfies $\hat{n}_{*,1}(x) \sim Cx^{-3/2}$ as $x \to 0$. More generally, one needs the initial data to have only a finite first moment for existence and uniqueness of a solution to (1.1) in the case of the additive kernel [15]. A deeper reason for this choice of variables (and notation) is probabilistic: (3.1) is the Lévy–Khintchine formula for the Laplace exponent of a subordinator with no drift [2]. We will always assume that the initial data n_0 satisfies the moment conditions

$$\int_0^\infty x n_0(x)dx = 1, \quad \int_0^\infty x^2 n_0(x)dx = 1. \tag{3.2}$$

We substitute (3.1) in (1.1) and use (3.2) to see that $\Phi(t,z)$ solves the equation

$$\partial_t \Phi - \Phi \partial_z \Phi = -\Phi, \quad \Phi(0,z) = \int_0^\infty (1-e^{-zx}) n_0(x)dx. \tag{3.3}$$

As shown in [15] by the method of characteristics, (3.3) has a unique solution for $z > 0$, $t > 0$ which is analytic with derivative $\partial_z\Phi$ completely monotone in z and satisfying $\partial_z\Phi(t,0) = 1$ for all t. For each $t > 0$ then, $\partial_z\Phi(t,\cdot)$ is the Laplace transform of a probability measure, so its domain contains $\bar{\mathbb{C}}_+$ and (3.3) holds by analytic continuation for $z \in \mathbb{C}_+$, $t > 0$.

In contrast with (2.4), it is not obvious that a suitable rescaling will lead to convergence to self-similar form. This point is discussed in [15, sect.7], and we refer the reader to that article for motivation for the following change of variables. We define the similarity variables

$$\hat{x} = xe^{-2t}, \quad s = ze^{2t} \tag{3.4}$$

and the rescaled number density

$$\hat{n}(t,\hat{x}) = e^{4t} n(t, \hat{x}e^{2t}) = e^{4t} n(t,x). \tag{3.5}$$

We also define the rescaled Laplace transforms

$$\varphi(t,s) = e^{2t}\Phi(t,e^{-2t}s) = \int_0^\infty (1-e^{-s\hat{x}})\hat{n}(t,\hat{x})d\hat{x}. \tag{3.6}$$

Part of the motivation for the rescaling (3.4) and (3.5) is that this choice preserves *both* moment conditions in (3.2). That is, we have

$$\int_0^\infty \hat{x}\hat{n}(t,\hat{x})d\hat{x} = \int_0^\infty \hat{x}^2\hat{n}(t,\hat{x})d\hat{x} = 1, \quad t \geq 0. \tag{3.7}$$

This should be compared with (2.9) for the constant kernel. The mass measure plays the same role here as the number measure did for $K = 2$. Thus, we denote its Laplace transform by the same letter, and let

$$u(t,s) = \partial_s\varphi(t,s) = \int_0^\infty e^{-s\hat{x}}\hat{x}\hat{n}(t,\hat{x})d\hat{x}. \tag{3.8}$$

By Theorem 7.1 in [15] (also see [13, Appendix G]), the assumptions in (3.2) imply that the rescaled mass measures converge to the similarity profile, with

$$\hat{x}\hat{n}(t,\hat{x})d\hat{x} \to \hat{x}\hat{n}_{*,1}(\hat{x})d\hat{x} = \frac{1}{\sqrt{2\pi}}\hat{x}^{-1/2}e^{-\hat{x}/2}d\hat{x}, \quad t \to \infty, \tag{3.9}$$

in the sense of weak convergence of measures. It then follows from [7, Theorem XIII.1.2] that (3.9) is equivalent to

$$\lim_{t\to\infty} u(t,s) = \frac{1}{\sqrt{1+2s}} =: u_{*,1}(s), \quad s \in \bar{\mathbb{C}}_+. \tag{3.10}$$

Our goal is to strengthen (3.9) to uniform convergence of densities for (1.1) and uniform convergence of coefficients for (1.6). For the continuous Smoluchowski equations we prove the following theorem.

THEOREM 3.1. *Suppose* $n_0(x) \geq 0$, $\int_0^\infty x n_0(x)dx = \int_0^\infty x^2 n_0(x)dx = 1$. *Suppose also that the Fourier transform of* $x^2 n_0$ *is integrable. Then in terms of the rescaling* (3.4)–(3.5) *we have*

$$\lim_{t\to\infty} \sup_{\hat{x}>0} \hat{x}^2 |\hat{n}(t,\hat{x}) - \hat{n}_{*,1}(\hat{x})| = 0, \tag{3.11}$$

where $\hat{n}_{*,1}(\hat{x})$ *is the similarity profile defined in* (1.3).

Once Theorem 3.1 is established, it is relatively straightforward to obtain the analogous result for the discrete Smoluchowski equations; see Theorem 3.6 below. Thus, most of our effort is devoted to Theorem 3.1.

Observe that we prove uniform convergence of the weighted density $\hat{x}^2\hat{n}(t,\hat{x})$. As in the previous section, this is because Theorem 3.1 is proved using the Fourier–Laplace inversion formula. Since $|u_{*,1}(ik)| \sim |k|^{-1/2}$ as $|k| \to \infty$, $u_{*,1}$ is not integrable on the imaginary axis. This divergence is due to the fact that $\hat{n}_{*,1}(\hat{x}) = 0$ for $\hat{x} < 0$ and $\hat{x}\hat{n}_{*,1}(\hat{x}) \sim C\hat{x}^{-1/2}$ as $\hat{x} \to 0^+$. As earlier, we resolve the situation by considering the transform of the next moment. Let

$$v(t,s) = -\partial_s u(t,s) = \int_0^\infty e^{-s\hat{x}} \hat{x}^2 \hat{n}(t,\hat{x}) d\hat{x}, \quad s \in \bar{\mathbb{C}}_+. \tag{3.12}$$

We integrate and differentiate (3.10) to obtain

$$\varphi_{*,1}(s) = \sqrt{1+2s} - 1, \quad v_{*,1}(s) = (1+2s)^{-3/2}, \quad s \in \bar{\mathbb{C}}_+. \tag{3.13}$$

3.2. Characteristics and estimates. The equations of evolution for φ and u are

$$\partial_t \varphi + (2s - \varphi)\partial_s \varphi = \varphi, \tag{3.14}$$

$$\partial_t u + (2s - \varphi)\partial_s u = -u(1-u). \tag{3.15}$$

In what follows, we first derive solution formulas to (3.14) by the method of characteristics. We then show that the solution map for the characteristic equation is never degenerate and that characteristics flow out of the right half into the left half of the complex plane as t increases. For most parts of our analysis, it will suffice to study characteristics in the right half plane only. But for one part, we need to study characteristics that start in the right half plane but move into the left half plane.

We use the notation $s(t; t_0, s_0)$ to denote the solution to

$$\frac{ds}{dt} = 2s - \varphi, \quad s(t_0; t_0, s_0) = s_0. \tag{3.16}$$

Along the characteristic curve $s(t; t_0, s_0)$, we have

$$\frac{d\varphi}{dt} = \varphi \quad \text{and} \quad \frac{du}{dt} = -u(1-u). \tag{3.17}$$

We integrate (3.17) to obtain

$$\varphi(t,s) = e^{t-t_0}\varphi(t_0,s_0), \quad u(t,s) = \frac{u(t_0,s_0)e^{-(t-t_0)}}{1-u(t_0,s_0)(1-e^{-(t-t_0)})}. \tag{3.18}$$

We now substitute for $\varphi(t,s)$ from (3.18) in (3.16) and integrate to obtain the explicit solution

$$e^{-2(t-t_0)}s(t;t_0,s_0) = s_0 - \varphi(t_0,s_0)(1-e^{-(t-t_0)}). \tag{3.19}$$

This equation can also be rewritten in two other useful forms, namely

$$e^{-2(t-t_0)}\,(s-\varphi(t,s)) = (s_0-\varphi(t_0,s_0)) \tag{3.20}$$

and

$$\frac{\varphi(t,s)}{s} = \frac{(\varphi(t_0,s_0)/s_0)e^{-(t-t_0)}}{1-(\varphi(t_0,s_0)/s_0)(1-e^{-(t-t_0)})}. \tag{3.21}$$

The method of characteristics also yields an explicit solution for $v(t,s)$. We differentiate (3.15) to obtain

$$\frac{dv}{dt} = -3(1-u)v. \tag{3.22}$$

We substitute for u from (3.18) and integrate (3.22) to obtain

$$v(t,s) = \frac{v(t_0,s_0)e^{-3(t-t_0)}}{\big(1-u(t_0,s_0)(1-e^{-(t-t_0)})\big)^3}. \tag{3.23}$$

Let $\varphi_0(s) := \varphi(0,s)$, and similarly $u_0(s) := u(0,s)$, $v_0(s) := v(0,s)$. Since $u = \partial_s\varphi$ and $\varphi(t,0) = 0$, the moment conditions (3.2) and the identity $\varphi_0(s)/s = \int_0^1 u_0(\tau s)\,d\tau$ imply

$$|u_0(s)| \le 1, \quad |v_0(s)| \le 1, \quad |\varphi_0(s)| \le |s|, \quad s \in \bar{\mathbb{C}}_+. \tag{3.24}$$

These inequalities are strict for $s \neq 0$ because $xn_0(x)\,dx$ is not a lattice measure [7, Lemma XV.1.4]. Taking $t_0 = 0$ at first, for $t \ge t_0$ we take absolute values in (3.18) and (3.23) to see that $|u|$ and $|v|$ decay along characteristics according to

$$|u(t,s)| \le \frac{|u(t_0,s_0)|e^{-(t-t_0)}}{1-|u(t_0,s_0)|(1-e^{-(t-t_0)})}, \tag{3.25}$$

$$|v(t,s)| \le \frac{|v(t_0,s_0)|e^{-3(t-t_0)}}{(1-|u(t_0,s_0)|)^3}. \tag{3.26}$$

From (3.25) and the fact that $|u_0(s_0)| < 1$ for $s_0 \neq 0$, and a similar estimate using (3.21) and $|\varphi_0(s_0)/s_0| < 1$, it follows that

$$|u(t,s)| < 1, \quad |\varphi(t,s)/s| < 1, \quad t \ge 0,\ s \neq 0. \tag{3.27}$$

Then (3.25) and (3.26) hold also for any $t_0 \ge 0$ if $t \ge t_0$.

Let us also note the uniform outward growth of characteristics implied by (3.27). Using (3.27) together with (3.16) we obtain

$$|s| \le \frac{d|s|}{dt} \le 3|s|. \tag{3.28}$$

Thus, $|s_0|e^{(t-t_0)} \le |s| \le e^{3(t-t_0)}|s_0|$. We will refine this crude estimate in the proof of Theorem 3.1, but we note here that $|s(t;t_0,s_0)|$ is a strictly increasing function of t.

In addition to the decay along characteristics, we will need the following uniform Riemann–Lebesgue lemma. Let $C_R = \{s \in \bar{\mathbb{C}}_+ \mid |s| = R\}$ denote the semicircle of radius R in the right half plane.

LEMMA 3.2. *Let $g(x) \in L^1(0,\infty)$ and $G(s) = \int_0^\infty e^{-sx} g(x)dx$. Then*

$$\lim_{R\to\infty} \sup_{s\in C_R} |G(s)| = 0. \tag{3.29}$$

Proof. Let $\varepsilon > 0$. We choose a step function $g_\varepsilon = \sum_{k=1}^K c_k 1_{[a_k,b_k]}$ so that $\|g - g_\varepsilon\|_{L^1} < \varepsilon$. But then, $\|e^{-sx}(g-g_\varepsilon)\|_{L^1} < \varepsilon$. Therefore, for $s \in \bar{\mathbb{C}}_+$,

$$|G(s)| \le \varepsilon + \left|\int_0^\infty e^{-sx} g_\varepsilon(x)dx\right| = \varepsilon + \left|\sum_{k=1}^K c_k \int_{a_k}^{b_k} e^{-sx}dx\right| \le \varepsilon + \frac{C_\varepsilon}{|s|}. \qquad \square$$

We apply this lemma and (3.7) to $g(\hat{x}) = \hat{x}^j \hat{n}(t,\hat{x})$ for $j = 1,2$ to infer that for every $t \ge 0$, as $|s| \to \infty$ with $\operatorname{Re} s \ge 0$, we have

$$|u(t,s)| \to 0, \quad |v(t,s)| \to 0, \quad \left|\frac{\varphi(t,s)}{s}\right| \to 0. \tag{3.30}$$

3.3. Geometry of the characteristic map in the complex plane. In this subsection, we study the solution formula (3.19). Our goal is to delineate some key properties of the map $s_0 \mapsto s(t;t_0,s_0)$ for $t, t_0 \ge 0$.

Let $\mathbb{C}_+$ denote the open right half plane. We let Ω_t denote the image of $\mathbb{C}_+$ under the map $s_0 \mapsto s(t;0,s_0)$, and let Γ_t denote the image of the imaginary axis under the same map. We aim to prove the following.

LEMMA 3.3.

(i) *For any $t > 0$, Γ_t is a C^2 curve that passes through the origin but otherwise lies in the open left half plane. On Γ_t, $\operatorname{Re} s$ is a C^2 function of $\operatorname{Im} s$.*

(ii) *Ω_t is the component of the complex plane to the right of Γ_t. Consequently $\Gamma_t = \partial\Omega_t$ and $\Omega_t \supset \bar{\mathbb{C}}_+ \setminus \{0\}$.*

(iii) *Whenever $t_1 \ge t_0 \ge 0$, the map $s_0 \mapsto s_1 = s(t_1;t_0,s_0)$ is one to one from $\bar{\Omega}_{t_0}$ onto $\bar{\Omega}_{t_1}$. It is C^2 on $\bar{\Omega}_{t_0}$ and analytic in Ω_{t_0}. The inverse map is given by $s_1 \mapsto s_0 = s(t_0;t_1,s_1)$ and is C^2 on $\bar{\Omega}_{t_1}$ and analytic in Ω_{t_1}.*

(iv) *Whenever $t_1 \ge 0$ and $s_1 \in \bar{\mathbb{C}}_+$, the backward characteristic curve $s(t_0;t_1,s_1)$, $t_0 \in [0,t_1]$, lies in $\bar{\mathbb{C}}_+$.*

Proof. We first establish part (iii), taking $t_0 = 0$ at first. Since $x^2 n_0$ is integrable, $v_0(s)$ is continuous in $\bar{\mathbb{C}}_+$ and analytic for $\operatorname{Re} s > 0$. It follows by a standard dominated convergence argument that u_0 is C^1 and φ_0 is C^2 in $\bar{\mathbb{C}}_+$, and these functions are analytic in $\mathbb{C}_+$. From (3.19) we see that the map $s_0 \mapsto s(t;0,s_0)$ is analytic in $\mathbb{C}_+$ and C^2 on $\bar{\mathbb{C}}_+$ (meaning derivatives up to second order extend continuously to $\bar{\mathbb{C}}_+$).

We next claim that this map is one to one. The proof relies on the fact that φ_0 is contractive, with

$$|\varphi_0(\tilde{s}_0) - \varphi_0(s_0)| \le |\tilde{s}_0 - s_0|, \quad \tilde{s}_0, s_0, \in \bar{\mathbb{C}}_+. \tag{3.31}$$

This holds because $|\partial_s \varphi_0(s)| \le 1$ for $s \in \bar{\mathbb{C}}_+$ as an immediate consequence of (3.7) and (3.8). Now suppose $s(t;0,\tilde{s}_0) = s(t;0,s_0)$, where $\tilde{s}_0, s_0 \in \bar{\mathbb{C}}_+$. Then (3.19) implies

$$\tilde{s}_0 - s_0 = \left(1 - e^{-t}\right)\left(\varphi_0(\tilde{s}_0) - \varphi_0(s_0)\right).$$

From this and (3.31) we infer $|\tilde{s}_0 - s_0| \le (1 - e^{-t})|\tilde{s}_0 - s_0|$, whence $\tilde{s}_0 = s_0$. So $s_0 \mapsto s(t;0,s_0)$ is one to one.

We observe that the derivative of this map is uniformly bounded away from zero. Indeed, (3.19) and (3.24) yield

$$\left|\frac{ds}{ds_0}\right| \ge e^{2t}\left(1 - |u_0(s_0)|(1 - e^{-t})\right) \ge e^t.$$

It follows by the inverse function theorem that Ω_t is an open set, and by continuity the image of $\bar{\mathbb{C}}_+$ is $\bar{\Omega}_t$. The inverse map from $\bar{\Omega}_t$ to $\bar{\mathbb{C}}_+$ is analytic in Ω_t, and C^2 on $\bar{\Omega}_t$.

For $t_1 > 0$, the inverse of the map $s_0 \mapsto s_1 = s(t_1;0,s_0)$ may be obtained by solving the characteristic equation in (3.16) backward from time t_1 to $t_0 = 0$, so that we have $s_0 = s(0;t_1,s_1)$. Now whenever $t_1 \ge t_0 \ge 0$ in general, we may follow any characteristic curve back from a point in $\bar{\Omega}_{t_1}$ at time t_1 to a point in $\bar{\mathbb{C}}_+$ at time 0 and then forward to a point in $\bar{\Omega}_{t_0}$ at time t_0. This means that $s(t_1;t_0,s_0) = s(t_1;0,s(0;t_0,s_0))$. Part (iii) of the lemma now follows from the properties established in the case $t_0 = 0$.

Next we prove part (i). For $t > 0$, Γ_t is the image of the map $k \mapsto s(t;0,ik) = e^{2t}(ik - \varphi_0(ik)(1-e^{-t}))$, $k \in \mathbb{R}$, and this is a C^2 function of k. We have $s(t;0,0) = 0$, but $\operatorname{Re} s < 0$ for $k \ne 0$. This is so because $\operatorname{Re} s$ and $\operatorname{Re}\varphi_0(ik)$ have opposite signs, and

$$\operatorname{Re}\varphi_0(ik) = \int_0^\infty (1 - \cos kx) n_0(x)\,dx > 0, \quad k \ne 0,$$

since n_0 is continuous. Finally, we find that

$$\operatorname{Im}\frac{d}{dk} s(t;0,ik) \ge e^{2t}(1 - |u_0(ik)|(1 - e^{-t})) > 0$$

using (3.24). Hence $\operatorname{Re} s$ is a function of $\operatorname{Im} s$ on Γ_t.

Now we establish part (ii). By (3.30) we have that as $|s_0| \to \infty$ with $s_0 \in \bar{\mathbb{C}}_+$, $|\varphi_0(s_0)/s_0| \to 0$, so $s = s_0 e^{2t}(1 + o(1))$ by (3.19). Let $s_1 \in \mathbb{C}$ lie to the right of Γ_t, and put $f(s_0) = s(t;0,s_0) - s_1$. It follows by applying the argument principle to large semicircles that the analytic function f has a single zero at some point $s_0 \in \mathbb{C}_+$. Indeed, $\arg f(re^{i\theta}) \to \theta$ as $r \to \infty$ for $-\frac{\pi}{2} \le \theta \le \frac{\pi}{2}$, and as k goes from ∞ to $-\infty$, $f(ik)$ does not cross the positive real axis, so $\arg f(ik)$ changes from $\frac{\pi}{2}$ to $\frac{3\pi}{2}$. Thus, f maps a large semicircle to a curve that winds exactly once about 0. Hence $s_1 \in \Omega_t$.

Finally, part (iv) follows by a change of variables, replacing $t - t_0$ by t, and applying parts (i)–(iii). □

3.4. Proof of Theorem 3.1. 1. By the Fourier–Laplace inversion formula, it suffices to prove

$$\lim_{t\to\infty} \sup_{x>0} \left| \int_{\mathbb{R}} e^{ikx}\left[v(t,ik) - v_{*,1}(ik)\right] dk \right| = 0. \tag{3.32}$$

2. Let $\varepsilon \in (0, \frac{1}{8})$, and put $R = \frac{1}{2}\varepsilon^{-2}$. We will prove (3.32) by estimating the integral for $t \geq T$ separately in three regions: $|k| \leq R$, $R \leq |k| \leq \tilde{R}e^{2t}$, and $\tilde{R}e^{2t} \leq |k|$, where $\tilde{R} = Re^{-2T}$ and T depends only on ε and the initial data v_0. This is the same decomposition used in the proof of Theorem 2.1, and convergence in the region $|k| \leq R$ will follow as before. However, estimates for $|k| \geq R$ are more subtle and use the analyticity and geometry of the characteristic map.

3. $|k| \leq R$: Theorem 7.1 in [15] implies that $\hat{x}\hat{n}(\tau, \hat{x})\, d\hat{x} \to \hat{x}\hat{n}_{*,0}(\hat{x})\, d\hat{x}$ in the sense of weak convergence of measures. Combined with (3.7) this also implies that the measures $\hat{x}^2\hat{n}(\tau, \hat{x})\, d\hat{x}$ converge weakly to $\hat{x}^2\hat{n}_{*,1}(\hat{x})\, d\hat{x}$ as $t \to \infty$. But this implies $v(t, ik)$ converges to $v_{*,1}(ik)$ uniformly on compact subsets of $\bar{\mathbb{C}}_+$, and in particular on compact subsets of the imaginary axis [7, Theorem XV.3.2]. Thus,

$$\lim_{t\to\infty} \int_{-R}^{R} |v(t, ik) - v_{*,1}(ik)|\, dk = 0. \tag{3.33}$$

4. $|k| \geq R$: It is sufficient to consider only $k \geq R$, since $|v(t, ik)| = |v(t, -ik)|$. We will control $v(t, ik)$ and $v_{*,1}$ separately:

$$\int_R^\infty |v(t, ik) - v_{*,1}(ik)|\, dk \leq \int_R^\infty |v(t, ik)|\, dk + \int_R^\infty |v_{*,1}(ik)|\, dk.$$

But $|v_{*,1}(ik)| \leq (2k)^{-3/2}$ by (3.13). Thus,

$$\int_R^\infty |v_{*,1}(ik)|\, dk \leq \int_R^\infty (2k)^{-3/2}\, dk = (2R)^{-1/2} = \varepsilon.$$

5. In the rest of the proof we estimate $\int_R^\infty |v(t, ik)| dk$. In order to aid the reader, we state the main estimates as two distinct lemmas.

LEMMA 3.4. *Let $\varepsilon \in (0, \frac{1}{8})$. There exists $T > 0$ depending on ε and the initial data, and a universal constant C, such that if $t \geq T$ then*

$$\int_R^{Re^{2(t-T)}} |v(t, ik)|\, dk \leq C\varepsilon. \tag{3.34}$$

LEMMA 3.5. *Let $\tilde{R} > 0$. There exists $\tilde{C}$ depending on $\tilde{R}$ and the initial data such that for all $t \geq 0$ we have*

$$\int_{\tilde{R}e^{2t}}^\infty |v(t, ik)|\, dk \leq \tilde{C}e^{-t}. \tag{3.35}$$

6. We now prove (3.32). We choose T as in Lemma 3.4, and then $\tilde{R} = Re^{-2T}$ in Lemma 3.5. Choose $T_* \geq T$ such that for $t \geq T_*$

$$\int_{-R}^{R} |v(t, ik) - v_{*,1}(ik)|\, dk < \varepsilon, \qquad \tilde{C}e^{-t} \leq \tilde{C}e^{-T_*} < \varepsilon.$$

Thus, for $t \geq T_*$ we have

$$\begin{aligned}
&\int_{\mathbb{R}} |v(t, ik) - v_{*,1}(ik)|\, dk \leq \int_{-R}^{R} |v(t, ik) - v_{*,1}(ik)|\, dk \\
&+2\left(\int_R^\infty |v_{*,1}(ik)|\, dk + \int_R^{\tilde{R}e^{2t}} |v(t, ik)|\, dk + \int_{\tilde{R}e^{2t}}^\infty |v(t, ik)|\, dk \right) \\
&\leq \varepsilon + 2\left(\varepsilon + C\varepsilon + \varepsilon\right).
\end{aligned}$$

Since $\varepsilon \in (0, \frac{1}{8})$ may be chosen arbitrarily small, this completes the proof.

3.5. Proof of Lemma 3.4. In this subsection we will always suppose $s \in \bar{\mathbb{C}}_+$. In a manner similar to step 6 of the proof of Theorem 2.1, the idea is to get estimates on the semicircle $C_R := \{s \in \bar{\mathbb{C}}_+ \mid |s| = R\}$ valid for large time and propagate these estimates outward along characteristics. We first use (3.10) and (3.13) to obtain the following estimates for $s \in \bar{\mathbb{C}}_+$:

$$|\varphi_{*,1}(s)| < |2s|^{1/2}, \quad |u_{*,1}(s)| < |2s|^{-1/2}, \quad |v_{*,1}(s)| < |2s|^{-3/2}. \tag{3.36}$$

Next, we use the uniform convergence on compact sets and (3.36) to see that there exists T_0 (depending on ε and the initial data) such that for all $s_0 \in C_R$ and $t_0 \geq T_0$ we have

$$|\varphi(t_0, s_0)/s_0| \leq 2(2R)^{-1/2} = 2\varepsilon \leq 1/4, \tag{3.37}$$

$$|u(t_0, s_0)| \leq (2R)^{-1/2} = \varepsilon, \tag{3.38}$$

$$|v(t_0, s_0)| \leq (2R)^{-3/2} = \varepsilon^3. \tag{3.39}$$

We first extend (3.37) to a larger domain in s.

Claim 1. There exists $T_1 \geq T_0$ such that

$$\left|\frac{\varphi(t,s)}{s}\right| \leq 1/3, \quad t \geq T_1,\ s \in \bar{\mathbb{C}}_+,\ |s| \geq R. \tag{3.40}$$

Proof of Claim 1. Observe that by using (3.27) and (3.30) in (3.21), we have

$$a := \sup\{|\varphi(T_0, s)/s| \mid s \in \bar{\mathbb{C}}_+,\ |s| \geq R\} < 1.$$

Fix $t_1 \geq T_0$, $s_1 \in \bar{\mathbb{C}}_+$ with $|s_1| \geq R$. Either the characteristic curve $s(t; t_1, s_1)$ that passes through s_1 at time t_1 intersects C_R at some time $t_0 \in [T_0, t_1]$, or not. If so, then $s_1 = s(t_1; t_0, s_0)$ for some $s_0 \in C_R$, and (3.21) and (3.37) directly yield

$$\left|\frac{\varphi(t_1, s_1)}{s_1}\right| \leq \frac{1/4}{1 - 1/4} = \frac{1}{3}.$$

If not, then $|s(t; t_1, s_1)| > R$ for all $t \in [T_0, t_1]$, by continuity and the fact that $s(t; t_1, s_1) \in \bar{\mathbb{C}}_+$ for all $t \in [0, t_1]$ by part (iv) of Lemma 3.3. Then taking $t_0 = T_0$, $s_0 = s(T_0; t_1, s_1)$ in (3.21) yields

$$\left|\frac{\varphi(t_1, s_1)}{s_1}\right| \leq \frac{ae^{-(t_1 - T_0)}}{1 - a} \leq \frac{1}{3},$$

provided $t_1 \geq T_1$ with T_1 sufficiently large. This proves the claim.

Claim 2. Let $T = T_1 + \frac{1}{2}\ln 2$. Suppose $t_1 \geq T$ and $R \leq |s_1| \leq Re^{2(t_1 - T)}$. Then the characteristic curve $s(t; t_1, s_1)$ that passes through s_1 at time t_1 intersects C_R at some time $t_0 \in [T_1, t_1]$.

Proof of Claim 2. Suppose the claim were false. Then the continuity of $|s(t; t_1, s_1)|$ and part (iv) of Lemma 3.3 imply $R < |s(t_0; t_1, s_1)|$ for all $t_0 \in [T_1, t_1]$. But now, by (3.20) with $s_0 = s(t_0; t_1, s_1)$ we have

$$s_0\left(1 - \frac{\varphi(t_0, s_0)}{s_0}\right) = e^{-2(t_1 - t_0)} s_1 \left(1 - \frac{\varphi(t_1, s_1)}{s_1}\right). \tag{3.41}$$

We take $t_0 = T_1$ and apply (3.40) and the hypothesis $|s_1| \leq Re^{2(t_1 - T)} = \frac{1}{2}Re^{2(t_1 - T_1)}$ to deduce

$$R < |s_0| \leq |s_1| e^{-2(t_1 - T_1)} \frac{1 + 1/3}{1 - 1/3} \leq R,$$

a contradiction. This proves the claim.

We now apply these claims to propagate the decay estimate (3.39). From Claim 2, for any $t = t_1 \geq T$, $R \leq k \leq Re^{2(t-T)}$, with $s_1 = ik$, we obtain $t_0 \in [T_1, t]$ and $s_0 \in C_R$ and substitute (3.20), (3.39), and (3.40) in the decay estimate (3.26) to obtain

$$\begin{aligned}|v(t,ik)| &\leq \frac{|v(t_0,s_0)|}{(1-|u(t_0,s_0)|)^3}\left|\frac{s_0-\varphi(t_0,s_0)}{ik-\varphi(t,ik)}\right|^{3/2}\\ &\leq (1-\varepsilon)^{-3}\ |v(t_0,s_0)|\left|\frac{2s_0}{k}\right|^{3/2}\\ &\leq (1-\varepsilon)^{-3}(2R)^{-3/2}\left(\frac{2R}{k}\right)^{3/2} = (1-\varepsilon)^{-3}k^{-3/2}.\end{aligned}$$

Therefore,

$$\int_R^{Re^{2(t-T)}} |v(t,ik)|\,dk \leq (1-\varepsilon)^{-3}\int_R^\infty k^{-3/2}\,dk = \frac{2R^{-1/2}}{(1-\varepsilon)^3} \leq C\varepsilon, \tag{3.42}$$

with $C = 2(8/7)^3 2^{1/2}$. This completes the proof of Lemma 3.4.

3.6. Proof of Lemma 3.5. We consider the initial time $t_0 = 0$ and the following special case of (3.19):

$$s = s(t;0,s_0) = e^{2t}\left[s_0 - \varphi_0(s_0)(1-e^{-t})\right]. \tag{3.43}$$

For any $t \geq 0$, the map $s_0 \mapsto s(t;0,s_0)$ is analytic for $\mathrm{Re}(s_0) > 0$, and

$$\frac{ds}{ds_0} = e^{2t}\left(1 - u_0(s_0)(1-e^{-t})\right), \qquad u_0(s_0) = u(0,s_0). \tag{3.44}$$

Recall that Ω_t denotes the image of $\mathbb{C}_+$ under $s_0 \mapsto s(t;0,s_0)$, and Γ_t denotes the image of the imaginary axis; we let Γ_{-t} denote its preimage. As was observed in Lemma 3.3, Γ_t is a graph over the imaginary axis, contained in the left half plane.

We will use the analyticity of $v(t,s)$ in Ω_t and contour deformation. For large finite $R_2 < \infty$, consider the domain $ABCD$ shown in Figure 3.1. The path AB is chosen so that $A'B'$ is a straight line. CD is parallel to the real axis and lies in Ω_t since Γ_t is a graph over the imaginary axis. Then by Cauchy's theorem,

$$\begin{aligned}\int_{\tilde{R}e^{2t}}^{R_2} e^{ikx}v(t,ik)\,dk &= \int_{BC} e^{ikx}v(t,ik)\,dk\\ &= \int_{DA} e^{sx}v(t,s)\,ds + \int_{AB} e^{sx}v(t,s)\,ds + \int_{CD} e^{sx}v(t,s)\,ds.\end{aligned}$$

Let σ denote $\mathrm{Re}\,s$. Since $\sigma < 0$ in Ω_t for $s \in CD$ we see that the last integral is estimated by

$$\left|\int_{CD} e^{sx}v(t,s)\,ds\right| \leq \sup_{s\in CD}|v(t,s)|\int_{-\infty}^0 e^{\sigma x}d\sigma = \frac{\sup_{s\in CD}|v(t,s)|}{x}.$$

By the decay estimate (3.26) we have

$$\sup_{s\in CD}|v(t,s)| \leq \sup_{s_1\in CD}\frac{|v_0(s_0)|e^{-3t}}{(1-|u_0(s_0)|)^3}, \quad s_0 = s(0;t,s_1).$$

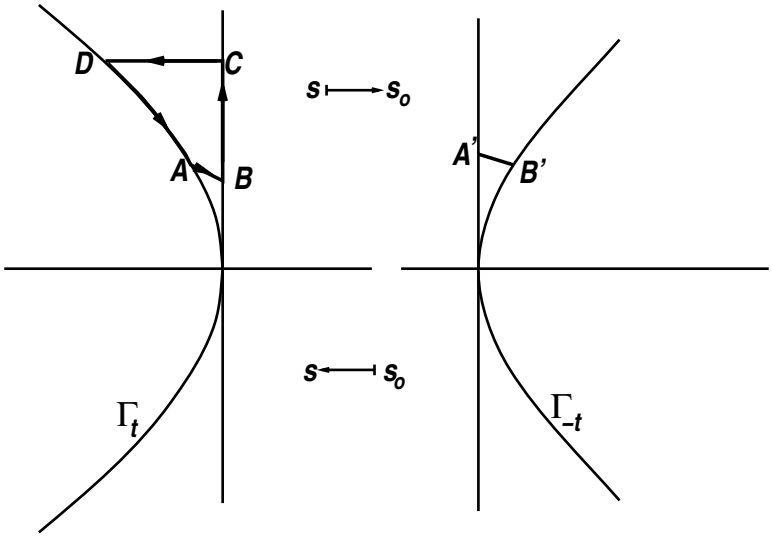

FIG. 3.1. *The s-plane is on the left, the s_0-plane on the right. Ω_t is the region to the right of Γ_t. $A = s(t;0,i\tilde{R})$, $B = i\tilde{R}e^{2t}$, $C = iR_2$, $\mathrm{Im}(D) = R_2$, $A' = i\tilde{R}$, $B' = s(0;t,i\tilde{R}e^{2t})$.*

It follows from (3.30) and the fact that $|s_0| = |s_1|e^{-2t}(1+o(1)) \to \infty$ as $R_2 \to \infty$ that $\sup_{s_1 \in CD} |v_0(s_0)| \to 0$. We thus let $R_2 \to \infty$ to conclude that

$$\int_{\tilde{R}e^{2t}}^{\infty} e^{ikx} v(t,ik)\, dk = \int_{\Gamma_{t,A}} e^{sx} v(t,s)\, ds + \int_{AB} e^{sx} v(t,s)\, ds, \tag{3.45}$$

where $\Gamma_{t,A}$ denotes the path from ∞ to A on Γ_t. Notice that (3.45) holds independent of x.

The virtue of deforming the contour is that the integrals may now be estimated by changing variables from s to s_0. We use the solution formula (3.23) together with the change of variables $s = s(t;0,ik)$ and (3.44) to obtain

$$\int_{\Gamma_{t,A}} e^{sx} v(t,s)\, ds = ie^{-t} \int_{\tilde{R}}^{\infty} e^{s(t;0,ik)x} \frac{v_0(ik)}{\left(1 - u_0(ik)(1-e^{-t})\right)^2}\, dk.$$

Since $\mathrm{Re}\, s(t;0,ik) \le 0$ and $\sup_{|k| \ge \tilde{R}} |u_0(ik)| < 1$, this yields the estimate

$$\left| \int_{\Gamma_{t,A}} e^{sx} v(t,s)\, ds \right| \le C_1 e^{-t} \|v_0\|_{L^1}. \tag{3.46}$$

Similarly, we have by (3.23) and (3.44)

$$\begin{aligned} \left| \int_{AB} e^{sx} v(t,s)\, ds \right| &= e^{-t} \left| \int_{A'B'} e^{s(t;0,s_0)x} \frac{v_0(s_0)}{\left(1 - u_0(s_0)(1-e^{-t})\right)^2}\, ds_0 \right| \\ &\le e^{-t} |A'B'| \sup_{s_0 \in A'B'} \left| 1 - u_0(s_0)(1-e^{-t}) \right|^{-2}. \end{aligned}$$

The point $A' = i\tilde{R}$ is independent of t. It also follows from (3.43) that $B' = s(0;t,i\tilde{R}e^{2t})$ converges to the point $s_0 \in \bar{\mathbb{C}}_+$ that solves $i\tilde{R} = s_0 - \varphi_0(s_0)$. Thus,

we have the exponential decay estimate

$$\left| \int_{AB} e^{sx} v(t,s)\, ds \right| \leq C_2 e^{-t}. \tag{3.47}$$

The constants C_i in (3.46) and (3.47) depend only on $\tilde{R}$ and the initial data u_0. To be explicit, we set $\tilde{C} = C_1 \|v_0\|_{L^1} + C_2$. This completes the proof.

3.7. The discrete Smoluchowski equations. We now use the proof of Theorem 3.1 to obtain a uniform convergence theorem for the discrete Smoluchowski equations with additive kernel. The proof is simpler and we do not need the contour deformation argument.

Let $\nu_t = \sum_{l=1}^{\infty} n_l(t)\delta_{hl}(x)$ denote a measure-valued solution to (1.1). We first adapt the rescaling (3.4) and (3.5) to similarity variables. Let

$$\hat{l} = lhe^{-2t}, \quad \hat{n}_l(t) = h^{-1}e^{4t} n_l(t). \tag{3.48}$$

Then the discrete Fourier inversion formula analogous to (2.32) is

$$\hat{l}^2 \hat{n}_l(t) = \frac{1}{2\pi} \int_{-\pi e^{2t}/h}^{\pi e^{2t}/h} e^{i\hat{l}k} v(t, ik)\, dk. \tag{3.49}$$

THEOREM 3.6. *Let $\nu_0 \geq 0$ be a lattice measure with span h such that $\int_0^\infty x\nu_0(dx) = \int_0^\infty x^2 \nu_0(dx) = 1$. Then with the scaling* (3.48) *we have*

$$\lim_{t\to\infty} \sup_{l\in\mathbb{N}} \hat{l}^2 \left| \hat{n}_l(t) - \hat{n}_{*,1}(\hat{l}) \right| = 0.$$

Proof. By (3.49) and the continuous Fourier inversion formulas it suffices to show that

$$\lim_{t\to\infty} \sup_{\hat{l}\geq 0} \left| \int_{-\pi e^{2t}/h}^{\pi e^{2t}/h} e^{i\hat{l}hk} v(t,ik)\, dk - \int_{\mathbb{R}} e^{i\hat{l}hk} v_{*,1}(ik)\, dk \right| = 0. \tag{3.50}$$

Let $\varepsilon \in (0, \frac{1}{8})$ and choose $R = \frac{1}{2}\varepsilon^{-2}$. The integrals over $[-R, R]$ and $R < |k| < \tilde{R}e^{2t}$ with $\tilde{R} = e^{-2T}$ are controlled as in the proof of Theorem 3.1, and it only remains to control the integral of $|v(t,ik)|$ over $\tilde{R}e^{2t} < k < \pi e^{2t}/h$. This is considerably simpler than in the previous proof. We use the solution formula (3.23) and change variables via $ik = s(t;0,s_0)$, then use (3.44) to obtain

$$\int_{\tilde{R}e^{2t}}^{\pi e^{2t}/h} e^{ikx} v(t,ik)\, dk = ie^{-t} \int_{\Gamma_{-t}(\tilde{R},\pi/h)} \frac{e^{xs(t;0,s_0)} v_0(s_0)}{\left(1 - u_0(s_0)(1-e^{-t})\right)^2}\, ds_0.$$

Here $\Gamma_{-t}(\tilde{R}, \pi/h)$ denotes the segment along the curve Γ_{-t} from $s(0;t,i\tilde{R}e^{2t})$ to $s(0;t,i\pi e^{2t}/h)$. The formula (3.19) shows that $\Gamma_{-t}(\tilde{R},\pi/h)$ converges to a compact C^2 curve defined implicitly by $ik = s_0 - \varphi_0(s_0), \tilde{R} \leq k \leq \pi/h$. Thus, for $t \geq T$ we have

$$e^{-t} \left| \int_{\Gamma_{-t}(\tilde{R},\pi/h)} \frac{e^{xs(t;0,s_0)} v_0(s_0)}{\left(1 - u_0(s_0)(1-e^{-t})\right)^2}\, ds_0 \right| \leq C(T, \tilde{R}, u_0, v_0) e^{-t}.$$

Thus, this term is less than ε for all t large enough. ∎

4. Self-similar gelation for the multiplicative kernel. For $K = xy$, McLeod solved the coagulation equation explicitly for monodisperse initial data and showed that a mass-conserving solution failed to exist for $t > 1$. The second moment satisfies $m_2(t) = (1-t)^{-1}$. The divergence of the second moment indicates that breakdown is associated with an explosive flux of mass toward large clusters. A rescaled limit of McLeod's solution is the following self-similar solution for $K = xy$ [1]:

$$n(t,x) = \frac{1}{\sqrt{2\pi}} x^{-5/2} e^{-(1-t)^2 x/2}, \quad x > 0, \quad t < 1. \tag{4.1}$$

Evidently this solution has infinite mass (first moment). This should not be thought unnatural, however, since it was shown in [15] that (1.1) has a unique weak solution for any initial distribution with finite second moment.

The problem of solving Smoluchowski's equation with multiplicative kernel can be reduced to that for the additive kernel by a change of variables [4]. Let us briefly review this. In unscaled variables we define

$$\Psi(t,z) = \int_0^\infty (1 - e^{-zx}) x n(t,x)\, dx. \tag{4.2}$$

Then Ψ solves the inviscid Burgers equation:

$$\partial_t \Psi - \Psi \partial_z \Psi = 0, \tag{4.3}$$

with initial data

$$\Psi_0(z) = \int_0^\infty (1 - e^{-zx}) x n_0(x)\, dx. \tag{4.4}$$

The gelation time for initial data with finite second moment is $T_{\text{gel}} = (\int_0^\infty x^2 \nu_0(dx))^{-1}$, and this is exactly the time for the first intersection of characteristics [15]. We presume that the initial data is scaled to ensure

$$\int_0^\infty x^2 n_0(x)\, dx = \int_0^\infty x^3 n_0(x)\, dx = 1. \tag{4.5}$$

Then the gelation time is $T_{\text{gel}} = 1$. The connection between the additive and multiplicative kernels is that Ψ solves (4.3) with initial data Ψ_0 if and only if $\Phi(\tau, z)$ is a solution to (3.3) with the same initial data, where

$$\Psi(t,z) = e^\tau \Phi(\tau, z), \quad \text{with } \tau = \log(1-t)^{-1}. \tag{4.6}$$

For solutions $n^{\text{mul}}(t,x)$ and $n^{\text{add}}(\tau,x)$ to Smoluchowski's equation with multiplicative and additive kernels, respectively, this means that

$$x n^{\text{mul}}(t,x) = (1-t)^{-1} n^{\text{add}}(\tau, x) \tag{4.7}$$

for all $t \in (0,1)$ if and only if the same holds at $t = 0$. We thus obtain a scaling limit as $t \to T_{\text{gel}}$ directly from Theorem 3.1. The similarity variables for the multiplicative kernel are

$$\hat{x} = (1-t)^2 x, \quad \hat{n}(t,\hat{x}) = \frac{n(t, \hat{x}(1-t)^{-2})}{(1-t)^5} = \frac{n(t,x)}{(1-t)^5}, \tag{4.8}$$

and the self-similar profile is

$$\hat{n}_{*,2}(\hat{x}) = \frac{1}{\sqrt{2\pi\hat{x}^5}} e^{-\hat{x}/2}. \tag{4.9}$$

THEOREM 4.1. *Suppose* $n_0(x) \geq 0$, $\int_0^\infty x^2 n_0(x)\,dx = \int_0^\infty x^3 n_0(x)\,dx = 1$. *Suppose also that the Fourier transform of* $x^3 n_0$ *is integrable. Then in terms of the rescaling* (4.8) *we have*

$$\lim_{t\to 1} \sup_{\hat{x}>0} \hat{x}^3 |\hat{n}(t,\hat{x}) - \hat{n}_{*,2}(\hat{x})| = 0, \tag{4.10}$$

where $\hat{n}_{*,2}(\hat{x})$ *is the self-similar density in* (4.9).

Notice that (4.8) is *not* a mass-preserving rescaling; indeed, the rescaled mass diverges:

$$\int_0^\infty \hat{x}\hat{n}(t,\hat{x})d\hat{x} = \frac{1}{1-t}\int_0^\infty xn(t,x)\,dx = \frac{1}{1-t} \to \infty.$$

Instead, (4.8) preserves the second moment:

$$\int_0^\infty \hat{x}^2\hat{n}(t,\hat{x})d\hat{x} = (1-t)\int_0^\infty x^2 n(t,x)\,dx = 1, \quad t \in [0,1).$$

The explanation is that the scaling in (4.8) is designed to capture the behavior of the distribution of large clusters as t approaches T_{gel}—the average cluster size is $(1-t)^{-1}$. Correspondingly, the mass of the self-similar solution is infinite.

Theorem 3.6 may be similarly adapted to $K = xy$. In the discrete case, the correspondence (4.7) between solutions of Smoluchowski's equations with multiplicative and additive kernels becomes

$$hln_l^{\text{mul}}(t) = (1-t)^{-1} n_l^{\text{add}}(\log(1-t)^{-1}) \tag{4.11}$$

We introduce similarity variables via

$$\hat{l} = lh(1-t)^2, \quad \hat{n}_l(t) = h^{-1}(1-t)^{-5} n_l(t). \tag{4.12}$$

Then directly from Theorem 3.6 we obtain the following.

THEOREM 4.2. *Let* $\nu_0 \geq 0$ *be a lattice measure with span* h *such that* $\int_0^\infty x^2\nu_0(dx) = \int_0^\infty x^3\nu_0(dx) = 1$. *Then with the rescaling* (4.12) *we have*

$$\lim_{t\to 1} \sup_{l\in\mathbb{N}} \hat{l}^3 \left|\hat{n}_l(t) - \hat{n}_{*,2}(\hat{l})\right| = 0. \tag{4.13}$$

Acknowledgments. The authors are grateful to an anonymous referee for suggestions that greatly improve the presentation and its accuracy. The authors thank the Max Planck Institute for Mathematics in the Sciences, Leipzig, for hospitality during part of this work. G.M. thanks Timo Seppäläinen for his help during early stages of this work.

REFERENCES

[1] D. J. ALDOUS, *Deterministic and stochastic models for coalescence (aggregation and coagulation): A review of the mean-field theory for probabilists*, Bernoulli, 5 (1999), pp. 3–48.

[2] J. Bertoin, *Eternal solutions to Smoluchowski's coagulation equation with additive kernel and their probabilistic interpretations*, Ann. Appl. Probab., 12 (2002), pp. 547–564.
[3] F. P. Da Costa, *On the dynamic scaling behaviour of solutions to the discrete Smoluchowski equations*, Proc. Edinburgh Math. Soc. (2), 39 (1996), pp. 547–559.
[4] R. L. Drake, *A general mathematical survey of the coagulation equation*, in Topics in Current Aerosol Research, G. M. Hidy and J. R. Brock, eds., International Reviews in Aerosol Physics and Chemistry, Pergamon,Oxford, 1972, pp. 201–376.
[5] P. B. Dubovskiĭ and I. W. Stewart, *Existence, uniqueness and mass conservation for the coagulation-fragmentation equation*, Math. Methods Appl. Sci., 19 (1996), pp. 571–591.
[6] M. Escobedo, S. Mischler, and B. Perthame, *Gelation in coagulation and fragmentation models*, Comm. Math. Phys., 231 (2002), pp. 157–188.
[7] W. Feller, *An Introduction to Probability Theory and Its Applications, Vol. 2.*, 2nd ed., Wiley, New York, 1971.
[8] S. Friedlander, *Smoke, Dust and Haze: Fundamentals of Aerosol Behavior*, Wiley, New York, 1977.
[9] I. Jeon, *Existence of gelling solutions for coagulation-fragmentation equations*, Comm. Math. Phys., 194 (1998), pp. 541–567.
[10] Y. Katznelson, *An introduction to harmonic analysis*,corrected ed., Dover, New York, 1976.
[11] M. Kreer and O. Penrose, *Proof of dynamical scaling in Smoluchowski's coagulation equation with constant kernel*, J. Statist. Phys., 75 (1994), pp. 389–407.
[12] M. H. Lee, *A survey of numerical solutions to the coagulation equation*, J. Phys A: Math. Gen., 34 (2001), pp. 10219–10241.
[13] F. Leyvraz, *Scaling theory and exactly solved models in the kinetics of irreversible aggregation*, Phys., Rep., 383 (2003), pp. 95–212.
[14] J. B. McLeod, *On an infinite set of non-linear differential equations*, Quart. J. Math. Oxford Ser. (2), 13 (1962), pp. 119–128.
[15] G. Menon and R. Pego, *Approach to self-similarity in Smoluchowski's coagulation equations*, Comm. Pure Appl. Math., 57 (2004), pp. 1197–1232.
[16] J. R. Norris, *Smoluchowski's coagulation equation: uniqueness, nonuniqueness and a hydrodynamic limit for the stochastic coalescent*, Ann. Appl. Probab., 9 (1999), pp. 78–109.
[17] M. V. Smoluchowski, *Drei Vorträge über Diffusion, Brownsche Bewegung und Koagulation von Kolloidteilchen*, Physik. Z., 17 (1916), pp. 557–585.
[18] P. G. J. van Dongen and M. H. Ernst, *Scaling solutions of Smoluchowski's coagulation equation*, J. Statist. Phys., 50 (1988), pp. 295–329.

SIAM J. MATH. ANAL.
Vol. 36, No. 5, pp. 1652–1663

CONVERGENCE TO GLOBAL EQUILIBRIUM FOR A KINETIC FERMION MODEL*

LUKAS NEUMANN† AND CHRISTIAN SCHMEISER‡

Abstract. We study the long-time asymptotics of a kinetic model for fermions in a box with periodic boundary conditions. An entropy dissipation approach is used to prove decay to the global equilibrium for this nonlinear equation, that lacks dissipation in the position variable. We prove convergence at an algebraic rate depending on the smoothness of the solution. The result relies on some initial bounds and a uniform boundedness assumption for spatial derivatives of the solution.

Key words. kinetic equations, fermions, Fermi–Dirac distribution, semiconductors, H-theorem, relative entropy, entropy dissipation, long-time asymptotics

AMS subject classifications. 82C10, 35B40, 82D37, 82B10, 82C21, 35Q40

DOI. 10.1137/S0036141003436533

1. Introduction.

We investigate the initial value problem

$$\partial_t f + v \cdot \nabla_x f = Q(f), \quad f(0,x,v) = f_0(x,v), \tag{1}$$

where $f = f(t,x,v) \geq 0$ denotes a particle distribution function, depending on time $t \geq 0$, position $x \in \mathbb{T}^d$ (where $\mathbb{T}^d$ is a d-dimensional torus, i.e., a rectangular box with periodic boundary conditions), and velocity $v \in \mathbb{R}^d$.

The particles are fermions and the scattering operator Q (acting only in the v-direction) is a simple model for the interaction of the particles with a nonmoving background medium with constant temperature

$$Q(f) = \int_{\mathbb{R}^d} [M(1-f)f' - M'(1-f')f] \mathrm{d}v'. \tag{2}$$

Here $f' = f(t,x,v')$, and $M(v) = (2\pi)^{-d/2} e^{-|v|^2/2}$ is the normalized Maxwellian. The factors $(1-f)$ and $(1-f')$ take into account the Pauli exclusion principle. The values of the distribution function have to respect the bounds $0 \leq f \leq 1$.

A kinetic equation with the same scattering operator, but also including acceleration of the particles by a given electric field, has been considered in [12]. Existence and uniqueness for initial value problems with $x \in \mathbb{R}^d$ (nonperiodic) have been proven and the macroscopic (diffusion) limit has been carried out.

A fermion Boltzmann equation modeling elastic particle-particle collisions has been studied by Dolbeault [8]. More elaborate models than (2) for the scattering of fermions due to a background medium or a different species of particles have been considered in the modeling of charge transport in semiconductors (see [10], [11], [4]). Existence and uniqueness and/or macroscopic limits are the subject of these studies.

*Received by the editors October 16, 2003; accepted for publication (in revised form) June 4, 2004; published electronically April 29, 2005. This work has been supported by the Austrian Science Fund (grants W008 and P16174-N05) and by the EU financed HYKE network (contact no. HPRN-CT-2002-00282).

http://www.siam.org/journals/sima/36-5/43653.html

†Institut für Mathematik, Universität Wien, Nordbergstraße 15, 1090 Wien, Austria (Lukas.Neumann@univie.ac.at).

‡Institut für Analysis und Scientific Computing, TU Wien, Wiedner Hauptstraße 8-10, 1040 Wien, Austria, and Johann Radon Institute for Computational and Applied Mathematics, Altenbergerstraße 69, 4040 Linz, Austria (Christian.Schmeiser@tuwien.ac.at).

Here we are interested in the long-time behavior of solutions of (1). It is characterized by two properties: conservation of total mass and entropy dissipation. The first is a consequence of the conservation property $\int_{\mathbb{R}^d} Q(f)\mathrm{d}v = 0$ of the scattering operator and of the periodic boundary conditions in position space:

$$\int_{\mathbb{T}^d}\int_{\mathbb{R}^d} f(t,x,v)\mathrm{d}v\mathrm{d}x = \int_{\mathbb{T}^d}\int_{\mathbb{R}^d} f_0(x,v)\mathrm{d}v\mathrm{d}x.$$

Before stating the entropy dissipation property, we write the collision operator in the form

$$Q(f) = \int_{\mathbb{R}^d} MM'(1-f)(1-f')(F'-F)\mathrm{d}v'$$

with

$$F = \frac{f}{M(1-f)}\,.$$

Then, by the antisymmetry of the integrand with respect to v and v', it is easily shown that

$$\begin{aligned}\int_{\mathbb{R}^d} Q(f)\chi(F)\mathrm{d}v&\\ &= -\frac{1}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} MM'(1-f)(1-f')(F-F')(\chi(F)-\chi(F'))\mathrm{d}v'\mathrm{d}v \le 0\end{aligned} \tag{3}$$

for arbitrary increasing functions χ. As a consequence, if an entropy is defined by $H_\infty = \int_{\mathbb{T}^d}\int_{\mathbb{R}^d} S_\infty(f,v)\mathrm{d}v\mathrm{d}x$ with

$$\frac{\partial S_\infty}{\partial f} = \chi\left(\frac{f}{M(v)(1-f)}\right),$$

then the entropy dissipation equality

$$\frac{\mathrm{d}H_\infty}{\mathrm{d}t} = \int_{\mathbb{T}^d}\int_{\mathbb{R}^d} Q(f)\chi(F)\mathrm{d}v\mathrm{d}x \tag{4}$$

follows. If χ is chosen strictly increasing, then (by (3)) the entropy dissipation rate on the right-hand side of (4) vanishes only if the (local) equilibrium condition $F = \kappa(t,x)$ is satisfied and the distribution function is the Fermi–Dirac distribution:

$$f(t,x,v) = f_l(t,x,v) = \frac{\kappa(t,x)M(v)}{1+\kappa(t,x)M(v)}\,. \tag{5}$$

Note that in this statement, $\kappa(t,x)$ can be chosen arbitrarily. In the following, however, we shall denote by f_l defined by (5) the local equilibrium distribution associated with an arbitrary (nonequilibrium) distribution f, where κ is chosen by fixing the position density

$$\int_{\mathbb{R}^d} f_l(t,x,v)\mathrm{d}v = \rho(t,x) := \int_{\mathbb{R}^d} f(t,x,v)\mathrm{d}v. \tag{6}$$

Continuing our argument, we expect (because of the entropy dissipation equation) that for large times f approaches an equilibrium distribution, thus making the right-hand side of the transport equation (1) vanish. The left-hand side then vanishes only for constant κ. Thus, we expect f to converge to the global equilibrium

$$f_\infty(v) = \frac{\kappa_\infty M(v)}{1 + \kappa_\infty M(v)},$$

where κ_∞ is determined by mass conservation

$$|\mathbb{T}^d| \int_{\mathbb{R}^d} f_\infty(v)\mathrm{d}v = \int_{\mathbb{T}^d} \int_{\mathbb{R}^d} f_0(x,v)\mathrm{d}v\mathrm{d}x.$$

From (4) a weak version of this statement can be proven (see, e.g., [5], [1]).

Recently, Desvillettes and Villani have developed a strategy for proving strong convergence to equilibrium for nonhomogeneous (in position) kinetic equations. It includes quantitative estimates on convergence rates. They have applied their approach to linear equations with a Fokker–Planck scattering operator and a confining potential [6] as well as, in a monumental work [7], to the Boltzmann equation of gas dynamics. Linear models have also been considered in [9] and [2].

In this work, the method of Desvillettes and Villani is applied to (1) and (2). Its main point is to overcome the following difficulty: the right-hand side of the entropy dissipation equation vanishes when the distribution is in local equilibrium. Thus, the entropy might stop decaying without f having reached the global equilibrium f_∞. As an input, the method requires certain bounds (uniform in time) on the distribution function and on its derivatives with respect to position. Whereas these could be proved for the linear problems in [6], [9], and [2], they have to be assumed for the Boltzmann equation. For the nonlinear problem (1), methods from [12] can be used to prove the propagation of bounds for the initial conditions in terms of Fermi–Dirac distributions. The boundedness of x-derivatives—at least in the perturbative setting—has recently been proved in [3] for a class of problems including (1) and (2).

In the following section, the boundedness result is proved, the method is outlined, and the main result is stated. The detailed (rather involved) computations and estimates are collected in section 3.

2. Preliminaries and main result.

THEOREM 2.1. *Assume there exist constants $\kappa_-, \kappa_+ > 0$ such that*

$$f_-(v) \le f_0(x,v) \le f_+(v) \qquad \text{with } f_\pm(v) = \frac{\kappa_\pm M(v)}{1 + \kappa_\pm M(v)}$$

for all $x \in \mathbb{T}^d$ and $v \in \mathbb{R}^d$. Then there is a unique solution $f(t,x,v)$ of (1) and (2) satisfying the same bounds

$$f_-(v) \le f(t,x,v) \le f_+(v) \tag{7}$$

for all $t > 0$, $x \in \mathbb{T}^d$ and $v \in \mathbb{R}^d$.

Proof. For details we refer the reader to [12]. We only outline the proof of the bounds (7). The existence proof is based on a fixed point iteration on the set $\mathcal{V}$ of distribution functions satisfying (7). For $f \in \mathcal{V}$, we define the next iterate g by solving

$$\partial_t g + v \cdot \nabla_x g = M \int_{\mathbb{R}^d} f' \mathrm{d}v' - g \int_{\mathbb{R}^d} (Mf' + M'(1 - f'))\, \mathrm{d}v',$$

$$g(0,x,v) = f_0(x,v).$$

The difference $r = g - f_-$ satisfies

$$r(0,x,v) = f_0(x,v) - f_-(v) \geq 0$$

and

$$\begin{aligned}&\partial_t r + v\cdot\nabla_x r + r\int_{\mathbb{R}^d}(Mf' + M'(1-f'))\mathrm{d}v'\\ &= (1-f_-)M\int_{\mathbb{R}^d} f'\mathrm{d}v' - f_-\int_{\mathbb{R}^d} M'(1-f')\mathrm{d}v' \geq Q(f_-) = 0.\end{aligned}$$

Nonnegativity of r and, thus, the lower bound $g \geq f_-$ follows. Analogously, $g \leq f_+$ and, therefore, $g \in \mathcal{V}$ is shown. □

Note that this ensures that $(1-f)$ is bounded away from zero, which we will make use of frequently.

In the following, relative entropies will be used for measuring the distance between distributions. Some arbitrariness comes from the freedom to choose the function χ in (4). We define the relative entropy of f with respect to g by

$$H(f|g) := \int_{\mathbb{T}^d}\int_{\mathbb{R}^d} S(f,g)\mathrm{d}v\mathrm{d}x$$

with

$$S(f,g) = \int_g^f \ln\frac{z(1-g)}{g(1-z)}\mathrm{d}z = f\ln\frac{f(1-g)}{g(1-f)} + \ln\frac{1-f}{1-g}. \tag{8}$$

With this choice the relative entropy $H(f|f_\infty)$ coincides with the total entropy H_∞ defined in the introduction for $\chi(z) = \ln(z/\kappa_\infty)$. Until now this choice seems somewhat artificial, but we will further comment on it after explaining the strategy to derive the convergence result. We shall need the derivatives

$$\frac{\partial S}{\partial f} = \ln\frac{f(1-g)}{g(1-f)}, \qquad \frac{\partial^2 S}{\partial f^2} = \frac{1}{f(1-f)}. \tag{9}$$

By $S(f,f) = \frac{\partial S}{\partial f}(f,f) = 0$ and $\frac{\partial^2 S}{\partial f^2} > 0$, the relative entropy has the desired property to measure the distance between f and g. Actually, the following stronger statement is true.

LEMMA 2.2. *Let f and g satisfy* (7). *Then there exist constants $c_1, c_2 > 0$ such that*

$$c_1\|f-g\|_M^2 \leq H(f|g) \leq c_2\|f-g\|_M^2$$

with the weighted L^2-norm

$$\|f\|_M^2 := \int_{\mathbb{T}^d}\int_{\mathbb{R}^d}\frac{f^2}{M}\,\mathrm{d}v\mathrm{d}x.$$

Proof. By (9) and the mean value theorem,

$$S(f,g) = \frac{(f-g)^2}{2\phi(1-\phi)}$$

with ϕ lying between f and g. In particular, ϕ also satisfies (7). As a consequence

$$\frac{M}{c_2} \le 2\phi(1-\phi) \le \frac{M}{c_1}$$

holds with appropriate constants c_1, c_2, completing the proof. ∎

A second important property is what we would call a nonlinear version of the Pythagorean theorem.

LEMMA 2.3. *The relative entropy is additive with respect to the local equilibrium,*

$$H(f|f_l) + H(f_l|f_\infty) = H(f|f_\infty)\,.$$

Proof. A straightforward computation gives

$$H(f|f_l) + H(f_l|f_\infty) = H(f|f_\infty) + \int_{\mathbb{T}^d}\int_{\mathbb{R}^d} (f - f_l)\mathrm{d}v\, \ln\frac{\kappa_\infty}{\kappa}\,\mathrm{d}x.$$

The integral with respect to velocity vanishes because of (6). ∎

Also since

$$\frac{\partial S}{\partial f}(f, f_\infty) = \ln\frac{f}{\kappa_\infty M(1-f)}\,,$$

we can use (4) and (3) with $\chi(z) = \ln(z/\kappa_\infty)$ to obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}H(f|f_\infty) = -\frac{1}{2}\int_{\mathbb{T}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} MM'(1-f)(1-f')(F-F')\ln\frac{F}{F'}\,\mathrm{d}v'\mathrm{d}v\mathrm{d}x. \tag{10}$$

As mentioned above, the right-hand side vanishes when f is a Fermi–Dirac distribution (such that F is independent of v). The basic idea of the Desvillettes–Villani approach is to show that in such a situation the entropy dissipation cannot remain zero as long as $f \neq f_\infty$. This is done by estimating the entropy dissipation in terms of the relative entropy of f with respect to the local equilibrium f_l and by deriving a second-order differential inequality for $H(f|f_l)$:

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t}H(f|f_\infty) \le -c_3 H(f|f_l), \\ &\frac{\mathrm{d}^2}{\mathrm{d}t^2}H(f|f_l) \ge c_4 H(f|f_\infty) - c_5 H(f|f_l)^{1-1/n}. \end{aligned} \tag{11}$$

It is the main contribution of this work to prove that these inequalities hold for appropriate $c_3, c_4, c_5 > 0$ and a positive integer n. This will be done in the following section. The proof requires the estimates from Theorem 2.1 and additional smoothness assumptions on the solution.

A result from [6] for systems of differential inequalities of the form (11) can then be used to get the following convergence theorem.

THEOREM 2.4. *Let the assumptions of Theorem* 2.1 *hold and let the solution* f *of* (1) *satisfy*

$$\left\| \frac{\partial^{k_1+\cdots+k_d} f}{\partial x_1^{k_1}\cdots\partial x_d^{k_d}}(t,\cdot,\cdot)\right\|_M \le c_6, \qquad \forall\, k_1+\cdots+k_d \le n \quad \textit{and} \quad \forall\, t>0, \tag{12}$$

for a constant c_6 and a positive integer n. Then there exists a constant $c_7 > 0$ such that

$$H(f|f_\infty) \le c_7 t^{1-n}.$$

Remark 1. For the Fokker–Planck collision operator, a smoothness result like (12) was proven in [6] even for nonsmooth initial conditions by exploiting a hypoellipticity property. Here, one can hope only for propagation of regularity as in [9], assuming smoothness of the initial data. This question is dealt with in [3], where (12) is derived from the corresponding bound for the initial data if the initial data is close to the equilibrium in a suitable sense.

Remark 2. In principle we have the freedom to choose in (8) any entropy of the type

$$S_\chi(f,g) = \int_g^f \chi\left(\frac{z(1-g)}{g(1-z)}\right) \mathrm{d}z$$

with an arbitrary monotone increasing function χ. Since the biggest difficulty is to deduce the second inequality in the system (11), we choose the relative entropy such that the expression for the derivative $\frac{\mathrm{d}}{\mathrm{d}t}H(f|f_l)$ becomes as simple as possible (see (14) below), leading to the choice $\chi = \ln(\cdot/\kappa_\infty)$, which corresponds to (8).

COROLLARY 2.5. *With the assumptions of Theorem* 2.4 *there exists $c_8 > 0$, such that*

$$\|f(t,\cdot,\cdot) - f_\infty\|_{L^1(\mathbb{T}^d\times\mathbb{R}^d)} \le c_8 t^{(1-n)/2}.$$

Proof. The Cauchy–Schwarz inequality implies

$$\|g\|_{L^1(\mathbb{T}^d\times\mathbb{R}^d)} = \int_{\mathbb{T}^d}\int_{\mathbb{R}^d} \frac{|g|}{\sqrt{M}}\sqrt{M}\,\mathrm{d}v\mathrm{d}x \le \sqrt{|\mathbb{T}^d|}\,\|g\|_M.$$

The result now follows from Lemma 2.2 and Theorem 2.4. ∎

3. Derivation of differential inequalities. Throughout this section c will be a positive real constant that may change from line to line.

LEMMA 3.1. *Let the assumptions of Theorem* 2.4 *hold. Then there is a constant $c_3 > 0$ such that*

$$\frac{\mathrm{d}}{\mathrm{d}t}H(f|f_\infty) \le -c_3 H(f|f_l).$$

Proof. We have to estimate the entropy production (10). Note that by the mean value theorem,

$$\ln\frac{F}{F'} = \frac{F-F'}{\Phi}$$

holds, with Φ between F and F'. Also, by (7), we have $\kappa_- \le \Phi \le \kappa_+$. This gives

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}H(f|f_\infty) &\le -\frac{1}{2\kappa_+}\int_{\mathbb{T}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} MM'(1-f)(1-f')(F-F')^2\mathrm{d}v'\mathrm{d}v\mathrm{d}x\\ &\le -\frac{1}{2\kappa_+}\int_{\mathbb{T}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} MM'(1-f)(1-f')(1-f_l)(1-f_l')\\ &\qquad\times(F-F_l-(F'-F_l'))^2\mathrm{d}v'\mathrm{d}v\mathrm{d}x,\end{aligned}$$

where we have used $F_l = \frac{f_l}{M(1-f_l)} = \kappa = F_l'$. Expanding the square and using

$$F - F_l = \frac{f - f_l}{M(1-f)(1-f_l)}, \tag{13}$$

gives

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}H(f|f_\infty) \leq &-\frac{1}{\kappa_+}\int_{\mathbb{T}^d}\int_{\mathbb{R}^d} M(1-f)(1-f_l)\mathrm{d}v\int_{\mathbb{R}^d}\frac{(f-f_l)^2}{M(1-f)(1-f_l)}\mathrm{d}v\mathrm{d}x\\ &+\frac{1}{\kappa_+}\int_{\mathbb{T}^d}\left(\int_{\mathbb{R}^d}(f-f_l)\mathrm{d}v\right)^2\mathrm{d}x.\end{aligned}$$

The last term vanishes by the requirement (6) on the local equilibrium. From (7),

$$0 < 1 - f_+(0) \leq 1 - f, 1 - f_l \leq 1$$

follows and, therefore,

$$\frac{\mathrm{d}}{\mathrm{d}t}H(f|f_\infty) \leq -c\|f - f_l\|_M^2 .$$

An application of Lemma 2.2 completes the proof. ∎

Now we shall prove the second inequality in (11). A straightforward computation gives

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}H(f|f_l) &= \int_{\mathbb{T}^d}\int_{\mathbb{R}^d}\left(\partial_t f\ln\frac{F}{F_l} - \partial_t f_l\frac{f-f_l}{f_l(1-f_l)}\right)\mathrm{d}v\mathrm{d}x\\ &= \int_{\mathbb{T}^d}\int_{\mathbb{R}^d}\left(-v\cdot\nabla_x f\ln F + v\cdot\nabla_x f\ln\kappa + Q(f)\ln F - \frac{\partial_t\kappa}{\kappa}(f-f_l)\right)\mathrm{d}v\mathrm{d}x.\end{aligned}$$

The first term on the right-hand side vanishes by the divergence theorem (with respect to the x-variable) and the last one by (6), leaving

$$\frac{\mathrm{d}}{\mathrm{d}t}H(f|f_l) = \int_{\mathbb{T}^d}\nabla_x\cdot J\,\ln\kappa\,\mathrm{d}x + \int_{\mathbb{T}^d}\int_{\mathbb{R}^d}Q(f)\ln F\,\mathrm{d}v\mathrm{d}x = A + B \tag{14}$$

with the flux density $J = \int_{\mathbb{R}^d} vf\,\mathrm{d}v$ (which vanishes for $f = f_l$).

For the computation of the time derivative of A we need the momentum balance equation

$$\partial_t J + \nabla_x\cdot P = \int_{\mathbb{R}^d} vQ(f)\mathrm{d}v,$$

where we shall split the pressure tensor into a local equilibrium part and a remainder:

$$P = \int_{\mathbb{R}^d} v\otimes vf\,\mathrm{d}v = \int_{\mathbb{R}^d} v\otimes vf_l\,\mathrm{d}v + \int_{\mathbb{R}^d} v\otimes v(f-f_l)\,\mathrm{d}v = P_l + \tilde{P}.$$

Differentiating (6) with respect to time, and the continuity equation $\partial_t\rho + \nabla_x\cdot J = 0$ lead to

$$\frac{\partial_t\kappa}{\kappa} = \frac{-\nabla_x\cdot J}{\int_{\mathbb{R}^d} f_l(1-f_l)\mathrm{d}v}.$$

With these preparations we obtain

$$\begin{aligned}\frac{\mathrm{d}A}{\mathrm{d}t} &= \int_{\mathbb{T}^d} \frac{\nabla_x \kappa}{\kappa} \cdot (\nabla_x \cdot P_l)\mathrm{d}x + \int_{\mathbb{T}^d} \frac{\nabla_x \kappa}{\kappa} \cdot (\nabla_x \cdot \tilde{P})\mathrm{d}x \\ &\quad - \int_{\mathbb{T}^d} \frac{\nabla_x \kappa}{\kappa} \cdot \int_{\mathbb{R}^d} vQ(f)\mathrm{d}v\,\mathrm{d}x - \int_{\mathbb{T}^d} \frac{(\nabla_x \cdot J)^2}{\int_{\mathbb{R}^d} f_l(1-f_l)\mathrm{d}v}\,\mathrm{d}x. \end{aligned} \tag{15}$$

Note that for $f = f_l$ all terms on the right-hand side except the first vanish. This term is responsible for moving f out of local equilibrium as long as it is not in global equilibrium. For estimating it we need

$$\nabla_x \cdot P_l = \int_{\mathbb{R}^d} v \otimes v\, f_l(1-f_l)\mathrm{d}v\, \frac{\nabla_x \kappa}{\kappa}. \tag{16}$$

The integral is an isotropic tensor which is positive definite since f_l satisfies (7):

$$\int_{\mathbb{R}^d} v \otimes v\, f_l(1-f_l)\mathrm{d}v \geq \mathrm{Id} \int_{\mathbb{R}^d} v_i^2 f_-(1-f_+)\mathrm{d}v. \tag{17}$$

The first term on the right-hand side of (15) can, thus, be estimated by

$$\int_{\mathbb{T}^d} \frac{\nabla_x \kappa}{\kappa} \cdot (\nabla_x \cdot P_l)\mathrm{d}x \geq c\|\nabla_x \kappa\|^2_{L^2(\mathbb{T}^d)}.$$

Now we estimate the remaining three terms in (15) one by one. First,

$$\begin{aligned}\left|\int_{\mathbb{T}^d} \frac{\nabla_x \kappa}{\kappa} \cdot (\nabla_x \cdot \tilde{P})\mathrm{d}x\right| &\leq c \int_{\mathbb{T}^d} |\nabla_x \kappa| \int_{\mathbb{R}^d} |v|^2 |\nabla_x (f - f_l)|\mathrm{d}v\,\mathrm{d}x \\ &\leq c \int_{\mathbb{T}^d} |\nabla_x \kappa| \sqrt{\int_{\mathbb{R}^d} |v|^4 M\,\mathrm{d}v \int_{\mathbb{R}^d} \frac{|\nabla_x (f-f_l)|^2}{M}\mathrm{d}v}\,\mathrm{d}x \\ &\leq c\|\nabla_x \kappa\|_{L^2(\mathbb{T}^d)} \|\nabla_x (f - f_l)\|_M. \end{aligned}$$

Second (similarly),

$$\left|\int_{\mathbb{T}^d} \frac{\nabla_x \kappa}{\kappa} \cdot \int_{\mathbb{R}^d} vQ(f)\mathrm{d}v\,\mathrm{d}x\right| \leq c\|\nabla_x \kappa\|_{L^2(\mathbb{T}^d)} \|Q(f)\|_M.$$

Now we need to quantify the behavior of $Q(f)$ near f_l.

LEMMA 3.2. *Q is a bounded operator and moreover*

$$\|Q(f)\|_M \leq c\|f - f_l\|_M.$$

Proof.

$$\|Q(f)\|^2_M \leq c \int_{\mathbb{T}^d} \int_{\mathbb{R}^d} M \left(\int_{\mathbb{R}^d} M' |F - F'|\mathrm{d}v'\right)^2 \mathrm{d}v\,\mathrm{d}x.$$

The integral in parentheses can be estimated by

$$\begin{aligned}\int_{\mathbb{R}^d} M'|F - F'|\mathrm{d}v' &\leq |F - F_l| + \int_{\mathbb{R}^d} M'|F' - F_l'|\mathrm{d}v' \\ &\leq |F - F_l| + \sqrt{\int_{\mathbb{R}^d} M(F - F_l)^2 \mathrm{d}v}. \end{aligned}$$

Estimating the square of the sum by the sum of the squares we get

$$\|Q(f)\|_M^2 \le c\int_{\mathbb{T}^d}\int_{\mathbb{R}^d} M(F-F_l)^2 \mathrm{d}v\,\mathrm{d}x.$$

Now the result of the lemma follows from (13) and the boundedness of $(1-f)$ away from zero. ∎

Second, continued,

$$\left|\int_{\mathbb{T}^d} \frac{\nabla_x\kappa}{\kappa}\cdot\int_{\mathbb{R}^d} vQ(f)\mathrm{d}v\,\mathrm{d}x\right| \le c\|\nabla_x\kappa\|_{L^2(\mathbb{T}^d)}\|f-f_l\|_M.$$

Third (last term in (15)),

$$\begin{aligned}\int_{\mathbb{T}^d}\frac{(\nabla_x\cdot J)^2}{\int_{\mathbb{R}^d} f_l(1-f_l)\mathrm{d}v}\,\mathrm{d}x &\le c\int_{\mathbb{T}^d}\left(\int_{\mathbb{R}^d} v\cdot\nabla_x(f-f_l)\mathrm{d}v\right)^2\mathrm{d}x\\ &\le c\int_{\mathbb{T}^d}\int_{\mathbb{R}^d}|v|^2M\mathrm{d}v\int_{\mathbb{R}^d}\frac{|\nabla_x(f-f_l)|^2}{M}\mathrm{d}v\,\mathrm{d}x = c\|\nabla_x(f-f_l)\|^2.\end{aligned}$$

Collecting our results so far, we have proved

$$\begin{aligned}\frac{\mathrm{d}A}{\mathrm{d}t} \ge c\|\nabla_x\kappa\|^2_{L^2(\mathbb{T}^d)} - \tilde{c}\big(&\|\nabla_x\kappa\|_{L^2(\mathbb{T}^d)}\|\nabla_x(f-f_l)\|_M\\ &+\|\nabla_x\kappa\|_{L^2(\mathbb{T}^d)}\|f-f_l\|_M + \|\nabla_x(f-f_l)\|_M^2\big),\end{aligned}$$

implying

$$\frac{\mathrm{d}A}{\mathrm{d}t} \ge c\|\nabla_x\kappa\|^2_{L^2(\mathbb{T}^d)} - \tilde{c}\big(\|f-f_l\|_M^2 + \|\nabla_x(f-f_l)\|_M^2\big). \tag{18}$$

The $\nabla_x\kappa$ term drives the solution out of local equilibria because it remains nonzero as long as κ is different from the constant κ_∞. A Poincaré-type estimate will help us to describe this by means of relative entropy.

LEMMA 3.3. $\|\nabla_x\kappa\|^2_{L^2(\mathbb{T}^d)} \ge cH(f_l|f_\infty)$ *with* $c>0$.

Proof. Since $\frac{\mathrm{d}\rho}{\mathrm{d}\kappa} = \frac{1}{\kappa}\int_{\mathbb{R}^d} f_l(1-f_l)\mathrm{d}v$ is bounded from above and away from zero (by (7)),

$$\|\nabla_x\kappa\|^2_{L^2(\mathbb{T}^d)} \ge c\|\nabla_x\rho\|^2_{L^2(\mathbb{T}^d)}$$

with $c>0$. Introducing $\rho_\infty = \int_{\mathbb{R}^d} f_\infty\mathrm{d}v$ and noting that $\int_{\mathbb{T}^d}(\rho-\rho_\infty)\mathrm{d}x = 0$, a Poincaré estimate gives

$$\|\nabla_x\kappa\|^2_{L^2(\mathbb{T}^d)} \ge c\|\rho-\rho_\infty\|^2_{L^2(\mathbb{T}^d)}\mathrm{d}x$$

with a possibly different, but still positive constant c. On the other hand,

$$|\rho-\rho_\infty| \ge c|\kappa-\kappa_\infty| = c\frac{|f_l-f_\infty|}{M(1-f_l)(1-f_\infty)} \ge c\frac{|f_l-f_\infty|}{M},$$

where the first inequality follows again from the boundedness of $\frac{\mathrm{d}\rho}{\mathrm{d}\kappa}$ and the last from (7). This implies

$$\|\nabla_x\kappa\|^2_{L^2(\mathbb{T}^d)} \ge c\int_{\mathbb{T}^d}\int_{\mathbb{R}^d} M(\kappa-\kappa_\infty)^2\mathrm{d}v\mathrm{d}x \ge c\|f_l-f_\infty\|_M^2.$$

An application of Lemma 2.2 completes the proof. □

It remains to estimate the time derivative of the second term in (14). Using (3), this term can be written as

$$B = -\frac{1}{2}\int_{\mathbb{T}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} MM'(1-f)(1-f')(F-F')\ln\frac{F}{F'}\,\mathrm{d}v'\mathrm{d}v\mathrm{d}x.$$

The computation of the time derivative is facilitated by the fact that the integrand is symmetric with respect to f and f':

$$\frac{\mathrm{d}B}{\mathrm{d}t} = \int_{\mathbb{T}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} M'\frac{1-f'}{1-f}\left[M(1-f)(F-F')\ln\frac{F}{F'} - \ln\frac{F}{F'} - (F-F')\frac{1}{F}\right]\partial_t f\,\mathrm{d}v'\mathrm{d}v\mathrm{d}x.$$

The term multiplying $\partial_t f$ in the integrand can by estimated (using (7)) by $cM'|F-F'|$. As a consequence,

$$\left|\frac{\mathrm{d}B}{\mathrm{d}t}\right| \le c\int_{\mathbb{T}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} M'|F-F'|(|v|\,|\nabla_x f| + |Q(f)|)\mathrm{d}v'\mathrm{d}v\mathrm{d}x$$

holds. With $|F-F'| \le |f-f_l|/M + |f'-f_l'|/M'$, the right-hand side is bounded by the sum of four terms, which we estimate one by one. First,

$$\int_{\mathbb{T}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} M'\frac{|f-f_l|\,|Q(f)|}{M}\,\mathrm{d}v'\mathrm{d}v\mathrm{d}x \le \|f-f_l\|_M\|Q(f)\|_M.$$

Second,

$$\int_{\mathbb{T}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |f'-f_l'|\,|Q(f)|\,\mathrm{d}v'\mathrm{d}v\mathrm{d}x \le \|f-f_l\|_M\|Q(f)\|_M.$$

Third,

$$\int_{\mathbb{T}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |f'-f_l'|\,|v|\,|\nabla_x f|\,\mathrm{d}v'\mathrm{d}v\mathrm{d}x \le \int_{\mathbb{R}^d}|v|^2 M\,\mathrm{d}v\,\|f-f_l\|_M\|\nabla_x(f-f_\infty)\|_M.$$

The fourth term is the most difficult to estimate. Here we have to make a small concession on the exponent. We use $|f-f_l| \le cM$:

$$\int_{\mathbb{T}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} M'|v|\frac{|f-f_l|\,|\nabla_x f|}{M}\,\mathrm{d}v'\mathrm{d}v\mathrm{d}x \le c\int_{\mathbb{T}^d}\int_{\mathbb{R}^d}\left|\frac{f-f_l}{\sqrt{M}}\right|^{1-\varepsilon}\frac{|\nabla_x f|}{\sqrt{M}}|v|M^{\varepsilon/2}\mathrm{d}v\mathrm{d}x$$
$$\le c\left(\int_{\mathbb{R}^d}|v|^{2/\varepsilon}M\,\mathrm{d}v\right)^{\varepsilon/2}\|f-f_l\|_M^{1-\varepsilon}\|\nabla_x(f-f_\infty)\|_M.$$

Since the Maxwellian has finite moments of arbitrary order, ε can be made arbitrarily small. Collecting the four estimates, using Lemma 3.2 and the fact that $\|f-f_l\|_M$ is bounded, we have

$$\left|\frac{\mathrm{d}B}{\mathrm{d}t}\right| \le c\left(\|f-f_l\|_M^2 + \|\nabla_x(f-f_\infty)\|_M\|f-f_l\|_M^{1-\varepsilon}\right).$$

This implies together with (14), (18), and Lemma 3.3 that

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}H(f|f_l) \geq cH(f_l|f_\infty) - \tilde{c}\big(\|\nabla_x(f-f_l)\|_M^2 + \|f-f_l\|_M^2 + \|\nabla_x(f-f_\infty)\|_M\|f-f_l\|_M^{1-\varepsilon}\big). \tag{19}$$

The next step is the derivation of bounds for the norms of the gradients. The interpolation inequality

$$\|\nabla_x u\|_{L^2(\mathbb{T}^d)} \leq c\|u\|_{L^2(\mathbb{T}^d)}^{1-1/n}\|u\|_{H^n(\mathbb{T}^d)}^{1/n}$$

and the Hölder inequality imply

$$\begin{aligned}\|\nabla_x g\|_M^2 = \int_{\mathbb{R}^d}\frac{1}{M}\|\nabla_x g\|_{L^2(\mathbb{T}^d)}^2\mathrm{d}v &\leq c\int_{\mathbb{R}^d}\left(\frac{1}{M}\|g\|_{L^2(\mathbb{T}^d)}^2\right)^{1-1/n}\left(\frac{1}{M}\|g\|_{H^n(\mathbb{T}^d)}^2\right)^{1/n}\mathrm{d}v\\ &\leq c\|g\|_M^{2(1-1/n)}\left(\int_{\mathbb{R}^d}\frac{1}{M}\|g\|_{H^n(\mathbb{T}^d)}^2\mathrm{d}v\right)^{1/n}.\end{aligned}$$

By assumption (12) of Theorem 2.4 the last factor is bounded uniformly in time for $g = f - f_l$ and $g = f - f_\infty$. This gives

$$\|\nabla_x(f-f_l)\|_M^2 \leq c\|f-f_l\|_M^{2(1-1/n)}$$

and, with the Young inequality,

$$\begin{aligned}\|\nabla_x(f-f_\infty)\|_M\|f-f_l\|_M^{1-\varepsilon} &\leq c\|f-f_\infty\|_M^{1-1/n}\|f-f_l\|_M^{1-\varepsilon}\\ &\leq \delta\|f-f_\infty\|_M^2 + c_\delta\|f-f_l\|_M^{2(1-\varepsilon)n/(n+1)}.\end{aligned}$$

Now we choose $\varepsilon = n^{-2}$ (such that $(1-\varepsilon)n/(n+1) = 1-1/n$) and we use the above inequalities and Lemmas 2.2 and 2.3 in (19) to obtain the desired results

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}H(f|f_l) \geq c_4 H(f|f_\infty) - c_5 H(f|f_l)^{1-1/n}.$$

This completes the derivation of the differential inequalities (11) and, thus, the proof of Theorem 2.4.

REFERENCES

[1] N. Ben Abdallah and J. Dolbeault, *Relative entropies for the Vlasov-Poisson system in bounded domains*, C. R. Acad. Sci. Paris Sér. I Math., 330 (2000), pp. 867–872.

[2] M. J. Cáceres, J. A. Carrillo, and T. Goudon, *Equilibration rate for the linear inhomogeneous relaxation-time Boltzmann equation for charged particles*, Comm. Partial Differential Equations, 28 (2003), pp. 969–989.

[3] C. Mouhot and L. Neumann, *Quantitative study of convergence to equilibrium for collisional kinetic models in the torus and application to the Boltzmann, Landau and Fokker–Planck equation*, in preparation.

[4] I. Choquet, P. Degond, and C. Schmeiser, *Energy-transport models for charge carriers involving impact ionization in semiconductors*, Transport Theory Statist. Phys., 32 (2003), pp. 99–132.

[5] L. Desvillettes, *Convergence to equilibrium in large time for Boltzmann and B.G.K. equations*, Arch. Ration. Mech. Anal., 110 (1990), pp. 73–91.

[6] L. Desvillettes and C. Villani, *On the trend to global equilibrium in spatially inhomogeneous entropy-dissipating systems: The linear Fokker-Planck equation*, Comm. Pure Appl. Math., 54 (2001), pp. 1–42.

[7] L. Desvillettes and C. Villani, *On the trend to global equilibrium for spatially inhomogeneous kinetic systems: The Boltzmann equation*, Invent. Math., 159 (2005), pp. 245–316.

[8] J. Dolbeault, *Kinetic models and quantum effects: A modified Boltzmann equation for Fermi-Dirac particles*, Arch. Ration. Mech. Anal., 127 (1994), pp. 101–131.

[9] K. Fellner, L. Neumann, and C. Schmeiser, *Convergence to global equilibrium for spatially inhomogeneous kinetic models of non-micro-reversible processes*, Monatsh. Math., 141 (2004), pp. 289–299.

[10] F. Golse and F. Poupaud, *Limite fluide des eq́uations de Boltzmann des semi-conducteurs pour une statistique de Fermi-Dirac*, Asymptot. Anal., 6 (1992), pp. 135–160.

[11] P. Markowich, F. Poupaud, and C. Schmeiser, *Diffusion approximation of nonlinear electron-phonon collision mechanisms*, M2AN Math. Model Numer. Anal., 29 (1995), pp. 857–869.

[12] F. Poupaud and C. Schmeiser, *Charge transport in semiconductors with degeneracy effects*, Math. Methods Appl. Sci., 14 (1991), pp. 301–318.

SIAM J. MATH. ANAL.
Vol. 36, No. 5, pp. 1664–1688

INTERMEDIATE MODELS IN NONLINEAR OPTICS*

THIERRY COLIN†, GÉRARD GALLICE‡, AND KAREN LAURIOUX‡

Abstract. In this paper, new models are derived for laser propagation in a nonlinear medium. These models are intermediate between nonlinear Maxwell systems and nonlinear Schrödinger equations and are exact in linear cases. We prove rigorous error estimates for a generic class of systems. In the last section, we perform numerical tests in order to investigate the numerical effectivity of the bounds given by the theorem. We compare for a particular nonlinear system the exact solutions and the approximate solutions given by our new model. It is shown that the new models behave as predicted by the theorem but are even better in some cases.

Key words. nonlinear optics, WKB expansions, nonlinear hyperbolic systems, Schrödinger equations

AMS subject classifications. 35L60, 35B40, 35C20

DOI. 10.1137/S0036141003423065

1. Introduction.

1.1. Motivations. The aim of this paper is to propose new models for the simulation of the propagation of laser pulses in a nonlinear medium. The wavelength associated with a pulse is usually near the micrometer (10^{-6} m) while the length of the pulse can be of order 100 micrometers for ultrashort pulses (10^{-4} m) or of the order of the meter. We are concerned with propagation on distances of order of the millimeter (for crystals) or of hundred of meters (for propagation in gas). From the temporal point of view, the frequency of a pulse is $10^{15}\,\mathrm{s}^{-1}$, its duration can be of the order of the picoseconds (10^{-12} s) or of 10 nanoseconds (10^{-8} s). The duration of propagation can be 10^{-11} s for crystals or 10^{-6} s for gas. The width of the beam can be of order of a fraction of millimeter to a few centimeters. Therefore, one has to handle three-dimensional processes involving several orders of magnitude. It is not possible to propose direct simulations for all these situations. Usually, the so-called paraxial approximation or envelope approximation are used. This approximation relies on the fact that the electric field has the form of a plane wave multiplied by an envelope, namely $e^{i(kz-\omega t)}\mathcal{E}(t,x,y,z)$ where $t \geq 0$ is the time, $(x,y,z) \in \mathbb{R}^3$ are the spatial variables, k is the wave number, and ω is the frequency. With this notation, the slowly varying envelope approximation can be expressed by the following set of inequalities:

$$|\partial_t \mathcal{E}| \ll \omega|\mathcal{E}|, \qquad |\partial_x \mathcal{E}| \ll k|\mathcal{E}|, \qquad |\partial_y \mathcal{E}| \ll k|\mathcal{E}|, \qquad |\partial_z \mathcal{E}| \ll k|\mathcal{E}|.$$

Using these inequalities, one obtains approximate equations satisfied by $\mathcal{E}$. These equations can be nonlinear transport equations at the group velocity (for frequency doubling in the phase-matching case in a crystal) or nonlinear Schrödinger equations (in a Kerr medium) or Schrödinger–Bloch equations (in a gas) etc. We refer the reader to general textbook of physics (see [8], [18], for instance) for a precise physical

*Received by the editors February 17, 2003; accepted for publication (in revised form) June 25, 2004; published electronically April 29, 2005.
http://www.siam.org/journals/sima/36-5/42306.html
†MAB, Université Bordeaux 1 et CNRS UMR 5466, 351 cours de la libération, 33405 Talence Cedex, France (colin@math.u-bordeaux1.fr).
‡SIS, CEA CESTA, BP2, 33114 Le Barp, France (gallice@cea.fr).

description. Here, we will address cases where the validity of the paraxial approximation is not so clear. Physically, this can occur when the pulse goes through a diffraction web or when the pulse is "chirped" in order to have a large spectral width. We want to propose alternative intermediate models that are more precise than the usual Schrödinger-like equation but less expensive to compute numerically than the full Maxwell equations. These intermediate models are obtained in the same spirit as the long wave systems for water waves of [6] or [7]. For direct simulations on nonlinear Maxwell systems, see [5] and the references therein. See also [3] for cases with nonplanar phases. In order to introduce our notations, let us recall that a standard model for propagation of a beam in a Kerr medium is the Maxwell–Lorentz system which has the nondimensional form

$$\begin{cases} \partial_t B + \operatorname{curl} E = 0, \\ \partial_t E - \operatorname{curl} B = -\partial_t P, \\ \partial_t^2 P - \dfrac{1}{\varepsilon^2}(E - P) = \dfrac{1}{\varepsilon}|P|^2 P, \end{cases} \tag{1.1}$$

where (E, B) is the electromagnetic field and P is the polarization. Introducing $Q = \varepsilon \partial_t P$, this system becomes

$$\begin{cases} \partial_t B + \operatorname{curl} E = 0, \\ \partial_t E - \operatorname{curl} B = -\dfrac{Q}{\varepsilon}, \\ \partial_t Q - \dfrac{1}{\varepsilon}(E - P) = |P|^2 P, \\ \partial_t P - \dfrac{1}{\varepsilon} Q = 0. \end{cases} \tag{1.2}$$

For propagation in gas, one can use the two-level Maxwell–Bloch system

$$\begin{cases} \partial_t E - \operatorname{curl} B + \partial_t P = 0, \\ \partial_t B + \operatorname{curl} E = 0, \\ P = \operatorname{Re}(c_1 c_2^*) u, \end{cases} \tag{1.3}$$

where c_1 and c_2 are the complex representations of the populations in each level (c_2^* denotes the complex conjugate of c_2) and u is a fixed vector corresponding to the direction of propagation. Level 1 corresponds to the fundamental state, while level 2 corresponds to the excited state. The evolution of c_1 and c_2 is given by the following set of ordinary differential equations which is derived from the Schrödinger equation of quantum mechanics [18]:

$$\begin{cases} i\partial_t c_1 = -\dfrac{E \cdot u c_2}{\varepsilon}, \\ i\partial_t c_2 = \dfrac{1}{\varepsilon} c_2 - \dfrac{E \cdot u c_1}{\varepsilon}. \end{cases} \tag{1.4}$$

Introducing $\Lambda = c_1 c_2^*$ and $\tilde{N} = |c_1|^2 - |c_2|^2$ yields

$$\begin{cases} \partial_t \Lambda = \dfrac{i\Lambda}{\varepsilon} - \dfrac{iE \cdot u \tilde{N}}{\varepsilon}, \\ \partial_t \tilde{N} = -\dfrac{2iE \cdot u(\Lambda - \Lambda^*)}{\varepsilon}. \end{cases} \tag{1.5}$$

Let $P = \mathrm{Re}(\Lambda)$, $Q = \mathrm{Im}(\Lambda)$, and $\tilde{N} = 1 - N$, we obtain

$$\begin{cases} \partial_t P = -\dfrac{1}{\varepsilon} Q, \\ \partial_t Q = \dfrac{1}{\varepsilon} P - \dfrac{E \cdot u(1-N)}{\varepsilon}, \\ \partial_t N = -\dfrac{4E \cdot uQ}{\varepsilon}. \end{cases} \tag{1.6}$$

Now we change all the unknowns by a scaling factor $\sqrt{\varepsilon}$ and we consider a vectorial form of (1.6) without assuming that the electric filed is polarized along the unit vector u:

$$\begin{cases} \partial_t B + \mathrm{curl}\ E = 0, \\ \partial_t E - \mathrm{curl}\ B = \dfrac{Q}{\varepsilon}, \\ \partial_t Q + \dfrac{1}{\varepsilon}(E - P) = EN, \\ \partial_t P + \dfrac{1}{\varepsilon} Q = 0, \\ \partial_t N = -4E \cdot Q. \end{cases} \tag{1.7}$$

See [11] for a precise description of these models and the derivation of the nondimensional forms. See also [10] for the use of Maxwell–Bloch system in a gas. Since the solutions are expected under the form of a plane wave multiplied by an envelope, usually the initial data is taken as being equal to

$$(E, B, P, Q)(t = 0, X) = e^{i\frac{k \cdot X}{\varepsilon}} (E_0, B_0, P_0, Q_0)(X) + c.c.$$

with $X = (x, y, z)$ and $k \in \mathbb{R}^3$. The notation *c.c.* means "complex conjugate." For (1.7), one has, moreover, to take $N(t = 0, X) = 0$ since at the state of rest, all atoms are at level 1 and $c_1 = 1$ and $c_2 = 0$ which implies $N = 0$. Therefore the difficulties concerning the presence of different length scales for the propagation of the beam appears in (1.2) and (1.7) through the presence of terms of size $\frac{1}{\varepsilon}$ in the equations and also in the $e^{i\frac{k \cdot X}{\varepsilon}}$ in the initial data. These terms will create high frequencies (of order $\frac{1}{\varepsilon}$) in time. Moreover, we will need to characterize the solution on short-time scale ($O(1)$) or on long-time scale ($O(\frac{1}{\varepsilon})$), that is, on long or short distance. In order to give a synthetic presentation of these phenomena, we introduce the following general class of systems (including (1.1) and (1.3)) that has been used in several works (see [16], [15], [14], [9], ...):

$$\left(\partial_t + \sum_{j=1}^{n} A_j \partial_{x_j} + \frac{L_0}{\varepsilon} \right) u = f(u), \tag{1.8}$$

where matrices A_j are real symmetric, L_0 is skew-symmetric, f is a smooth nonlinear mapping, and

$$u(t, X) : [0, T] \times \mathbb{R}^n \to \mathbb{R}^p, \quad X = (x_1, \ldots, x_n).$$

For the sake of simplicity, in this paper we will restrict ourselves to the case where $f(u)$ is a homogeneous polynomial of degree q. Of course all of the results can be extended to more general cases.

1.2. Some classical results of nonlinear geometrical optics. We recall some tools of geometrical optics (see [14] for a more complete description). First we seek for plane wave solutions to the linear part of (3.1), that is,

$$u = Fe^{\frac{i(k\cdot X-\omega t)}{\varepsilon}}, \tag{1.9}$$

where $F \in \mathbb{C}^p$ is a constant and $k = (k_1, \dots, k_n) \in \mathbb{R}^n$. Such a plane wave is a solution to

$$\left(\partial_t + \sum_{j=1}^{n} A_j \partial_{x_j} + \frac{L_0}{\varepsilon}\right) u = 0 \tag{1.10}$$

if and only if

$$\left(-i\omega I_d + i\sum_{j=1}^{n} A_j k_j + L_0\right) F = 0, \tag{1.11}$$

where I_d denotes the identity matrix.

System (1.11) has a nontrivial solution if and only if

$$\det\left(-i\omega I_d + i\sum_{j=1}^{n} A_j k_j + L_0\right) = 0, \tag{1.12}$$

which is the dispersion relation. Note that the matrix $i\sum_{j=1}^{n} A_j k_j + L_0$ is skew-adjoint; therefore, the solutions ω are real and the solutions $i\omega$ are the eigenvalues of $i\sum_{j=1}^{n} A_j k_j + L_0$. Moreover the eigenspaces are orthogonal. We denote by $\Pi(\omega, k)$ (or simply by $\Pi(k)$ if no confusion is possible) the orthogonal projector onto $\mathrm{Ker}(-i\omega I_d + i\sum_{j=1}^{n} A_j k_j + L_0)$. We also give the following definition.

DEFINITION 1.1. *The characteristic variety $\mathcal{C}_{\mathcal{L}}$ of the operator $\mathcal{L}(\partial_t, \partial_X) = \partial_t + A(\partial_X) + L_0 := \partial_t + \sum_{j=1}^{n} A_j \partial_{x_j} + L_0$ is the set*

$$\mathcal{C}_{\mathcal{L}} = \{(\tau, \xi) \in \mathbb{R} \times \mathbb{R}^n \text{ such that } \det(-i\tau I_d + iA(\xi) + L_0) = 0\}.$$

Now, coming back to the nonlinear system (1.8), one tries to solve

$$\left(\partial_t + \sum_{j=1}^{n} A_j \partial_{x_j} + \frac{L_0}{\varepsilon}\right) u = f(u).$$

For a given $k \in \mathbb{R}^n$, we select a frequency ω. The way we solve this problem is the following one. We look for u in the form

$$u(t, X) = \mathcal{U}\left(\frac{k\cdot X - \omega t}{\varepsilon}, t, X\right),$$

where $\theta \mapsto \mathcal{U}(\theta, t, X)$ is 2π-periodic. Of course this is not enough in order to define completely function $\mathcal{U}$. We see that $\mathcal{U}$ satisfies the following singular equation:

$$\left(\partial_t + A(\partial_X) + \frac{1}{\varepsilon}\left(-\omega\partial_\theta + A(k)\partial_\theta + L_0\right)\right)\mathcal{U} = f(\mathcal{U}) \quad \text{for all } t \in [0, T], \tag{1.13}$$

for all $X \in \mathbb{R}^n$, and for $\theta = \frac{k\cdot X-\omega t}{\varepsilon}$. At this stage, the function $\mathcal{U}$ is not well defined since it satisfies (1.13) only for $\theta = \frac{k\cdot X-\omega t}{\varepsilon}$. In order to give a correct definition, we impose that $\mathcal{U}$ satisfies (1.13) for all $t \in [0,T]$, $X \in \mathbb{R}^N$, and for $\theta \in \mathbb{T}$ where $\mathbb{T}$ denotes the usual one-dimensional torus. We make the following generic hypothesis.

Hypothesis 1. (k,ω) is a regular point of $\mathcal{C}_L$ (that is, the multiplicity of the eigenvalue $\lambda_j(\xi)$ such that $\lambda_j(k) = \omega$ is constant in a neighborhood of $\xi = k$).

Hypothesis 2. $(pk, p\omega) \notin \mathcal{C}_L$ for all integer $p \le q$, where q is the degree of the nonlinearity f.

Note that Hypothesis 2 is not necessary; we could replace it by the strong finiteness hypothesis as in [12]. One then can construct an approximate solution for u as follows. Let $\mathcal{U}_0$ be the solution to

$$\begin{cases} \partial_t \mathcal{U}_0 + \omega'(k)\cdot\partial_X \mathcal{U}_0 = \Pi(k) C_1(f(\Pi(k)\mathcal{U}_0 e^{i\theta} + c.c.)), \\ \qquad\qquad \mathcal{U}_0(t=0,X) = \mathcal{U}_0(X), \end{cases} \tag{1.14}$$

where $C_q(F(\theta))$ denotes the qth Fourier coefficient of $\theta \mapsto F(\theta)$:

$$C_q(F(\theta)) = \frac{1}{2\pi}\int_0^{2\pi} F(\theta) e^{iq\theta} d\theta.$$

One then shows the following theorem.

THEOREM 1.2. *Let $u_0 \in H^s(\mathbb{R}^n)$ (for s large enough) such that $\Pi(k)u_0 = u_0$. There exists a unique $\mathcal{U}^\varepsilon(\theta,t,X)$ solution to the singular equation* (1.13) *such that $\mathcal{U}^\varepsilon(\theta,0,X) = (e^{i\theta}u_0 + c.c.)$ is defined on $[0,T]$ and*

$$|\mathcal{U}^\varepsilon(\theta,t,X) - (\mathcal{U}_0(t,X)e^{i\theta} + c.c.)|_{L^\infty_t(0,T;H^s_{\theta,X})} \le C_0\varepsilon.$$

It follows that there exists a solution $u^\varepsilon(t,X)$ to (1.8) such that $u^\varepsilon(0,X) = (e^{\frac{ik\cdot X}{\varepsilon}}u_0(X) + c.c.)$ and

$$\left| u^\varepsilon(t,X) - \left(\mathcal{U}_0(t,X)e^{i\frac{k\cdot X-\omega t}{\varepsilon}} + c.c.\right)\right|_{L^\infty([0,T]\times\mathbb{R}^n)} \le C_0\varepsilon.$$

This regime is called geometrical optics. For solutions on long-time scale of size $O(\frac{1}{\varepsilon})$, diffractive effects are important and we have to give another expansion. We look for a solution of (1.8) satisfying

$$u(t=0,X) = \varepsilon^{1/(q-1)}\left(e^{\frac{ik\cdot X}{\varepsilon}}u_0(X) + c.c.\right).$$

Let us recall that q is the order of the nonlinearity. In order to explain briefly why this scaling, $\varepsilon^{1/(q-1)}$, is relevant, let us consider the ordinary differential equation $y' = y^q$. An initial data of size $\varepsilon^{1/(q-1)}$ will lead to a solution of the same size $y(t) = \varepsilon^{1/(q-1)}z(t)$. Then the function $z(t)$ satisfies

$$z'(t) = \varepsilon^{q/(q-1)}\varepsilon^{-1/(q-1)}z^q = \varepsilon z(t)^q,$$

and $z(t)$ is therefore defined on a time interval of size $O(\frac{1}{\varepsilon})$.

The solution u is sought in the form

$$u(t,X) = \mathcal{U}\left(\frac{k\cdot X-\omega t}{\varepsilon}, X - \omega'(k)t, \varepsilon t\right),$$

where $\theta \mapsto \mathcal{U}(\theta, X, \tau)$ is defined on $\mathbb{T} \times \mathbb{R}^n \times [0, T]$. This will lead to a solution to (1.8) defined on $[0, \frac{T}{\varepsilon}]$. $\mathcal{U}$ then satisfies

$$\left(\varepsilon\partial_\tau + (-\omega'(k)\partial_X + A(\partial_X)) + \frac{1}{\varepsilon}(-\omega\partial_\theta + A(k)\partial_\theta + L_0)\right)\mathcal{U} = f(\mathcal{U}) \tag{1.15}$$

with

$$\mathcal{U}(\theta, t = 0, X) = \varepsilon^{1/(q-1)}(e^{i\theta}u_0(x) + c.c.).$$

Let $\mathcal{V}_0$ be the solution to the following nonlinear Schrödinger equation:

$$\partial_\tau \mathcal{V}_0 + i\frac{\omega''(k)}{2}(\partial_X, \partial_X)\mathcal{V}_0 = \Pi(k)C_1(f(\Pi(k)\mathcal{V}_0 e^{i\theta} + c.c.)) \tag{1.16}$$

with $\mathcal{V}_0(\tau = 0, X) = u_0(X)$.

THEOREM 1.3. *Let $u_0 \in H^s(\mathbb{R}^n)$ (for s large enough) such that $\Pi(k)u_0 = u_0$. There exists a unique $\mathcal{V}^\varepsilon(\theta, X, \tau)$ solution to the singular equation* (1.15) *such that $\mathcal{V}^\varepsilon(\theta, X, 0) = \varepsilon^{1/(q-1)}(e^{i\theta}u_0 + c.c.)$ defined on $[0, T]$ and*

$$\left|\frac{1}{\varepsilon^{1/(q-1)}}\mathcal{V}^\varepsilon(\theta, X, \tau) - (\mathcal{V}_0(\tau, X)e^{i\theta} + c.c.)\right|_{L^\infty_\tau(0,T;H^s_{\theta,X})} \leq C_1\varepsilon.$$

As before, it follows that

$$\left|\frac{1}{\varepsilon^{1/(q-1)}}u^\varepsilon(t, X) - \left(\mathcal{V}_0(\varepsilon t, X - \omega'(k)t)e^{i\frac{k\cdot X - \omega t}{\varepsilon}} + c.c.\right)\right|_{L^\infty([0,T]\times\mathbb{R}^n)} \leq C\varepsilon.$$

Of course, from the computational point of view, it is much easier to find low-frequency solutions to (1.14) or (1.16) on $[0, T] \times \mathbb{R}^n$ than oscillatory solutions of (1.8) on $[0, \frac{T}{\varepsilon}] \times \mathbb{R}^n$. Indeed the frequencies in time and space that are relevant for the solution of (1.8) are of size $O(\frac{1}{\varepsilon})$. Therefore, the time and space steps used in any numerical method have to be small compared to ε. This gives a number of points (in space) that has to be large compared to $O(\frac{1}{\varepsilon})$ and a number of time steps large compared to $O(\frac{1}{\varepsilon})$. For (1.14) or (1.16), the frequencies are $O(1)$ and the time or space steps have only to be small with respect to 1. Moreover, while (1.8) has to be solved in the diffractive regime on long-time intervals $[0, \frac{T}{\varepsilon}]$, (1.16) has to be solved on $[0, T]$ only, which decreases the number of time steps. This is why (1.14) or (1.16) are used in practical applications [17].

1.3. Limitations of the models. In some applications (ultrashort pulses), one can have to handle cases where ε is small, but not very small ($\varepsilon \sim 10^{-2}$). The error estimates given by the above results are not very precise especially when the constants C_0 and C_1 (depending at least on the H^s-norm of the initial data) are large. These constants can be large when the initial data has rapid variations and this is the case for short pulses or pulses with a quite large spectrum. This configuration arises when the laser beam propagates through a diffraction web. We give a numerical example below. Let us consider the simplified system

$$\partial_t \begin{pmatrix} u \\ v \end{pmatrix} + \partial_x \begin{pmatrix} v \\ u \end{pmatrix} + \frac{1}{\varepsilon}\begin{pmatrix} -v \\ u \end{pmatrix} = \begin{pmatrix} -(u^2 + v^2)v \\ (u^2 + v^2)u \end{pmatrix} \tag{1.17}$$

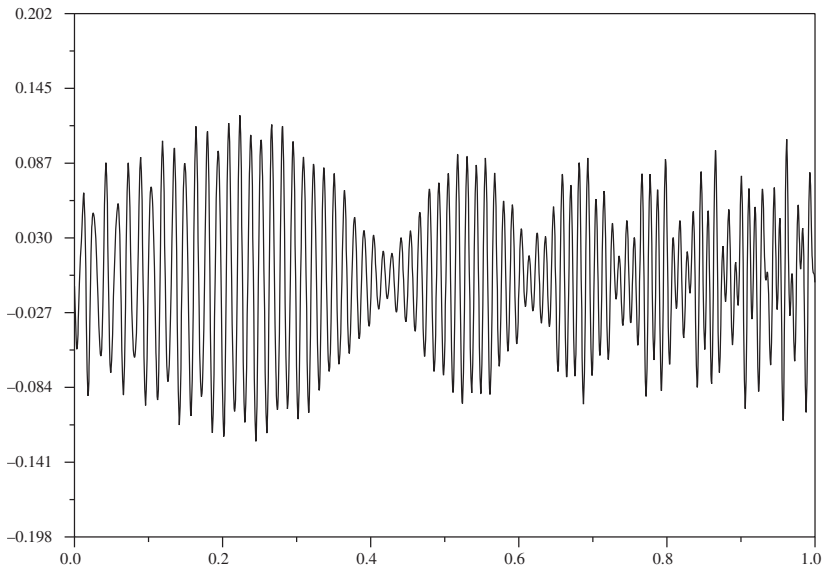

FIG. 1. *Real part of the first component of the solution of system* (1.17) *with* $\varepsilon = 0.01$ *at time* $T = 50$*, with chirped initial data, Case* 6.

with

$$\begin{pmatrix} u_0 \\ v_0 \end{pmatrix} = \varepsilon^{1/2} e^{i\frac{kx}{\varepsilon}} \begin{pmatrix} 1 \\ \frac{-ik+1}{i\omega} \end{pmatrix} a(x) + c.c., \quad x \in [0,1],$$

where $\omega = \sqrt{1+k^2}$. The function $a(x)$ is given by

$$a(x) = e^{-75(x-1/2)^2} e^{i15\cos(15x)}.$$

We make a simulation as described in the last section with $\varepsilon = 10^{-2}$ on $t \in [0, 50]$. The solution to (1.17) at time $t = 50$ is given on Figure 1 and the solution given by the nonlinear Schrödinger equation (1.16) is given on Figure 2. They have nothing in common and the relative error in L^2-norm is 1.4 as indicated in section 3.3.2. For practical use, Morice [17] has already introduced some modification of the linear Schrödinger equation in order to take into account higher-order diffraction effects. Other tentative modifications have been made by Alterman and Rauch [1], Schäfer and Wayne [19], and Barrailh and Lannes [2] for ultrashort pulses. In all these contributions, the authors obtain linear equation, because in a context of pulses with large spectrum it can be shown that the nonlinear effects are less important than usually (see [1] and [2]). Nevertheless, from the physical point of view, it is impossible to neglect nonlinear effects. We therefore need to construct new models that will be exact in the linear case, but that take into account the nonlinear effects and that are not numerically stiff.

This paper is organized as follows. In section 2, we introduce our new models and prove the main result. In section 3, we present some numerical results in order to illustrate our error bounds and also to investigate the numerical effectivity of our model.

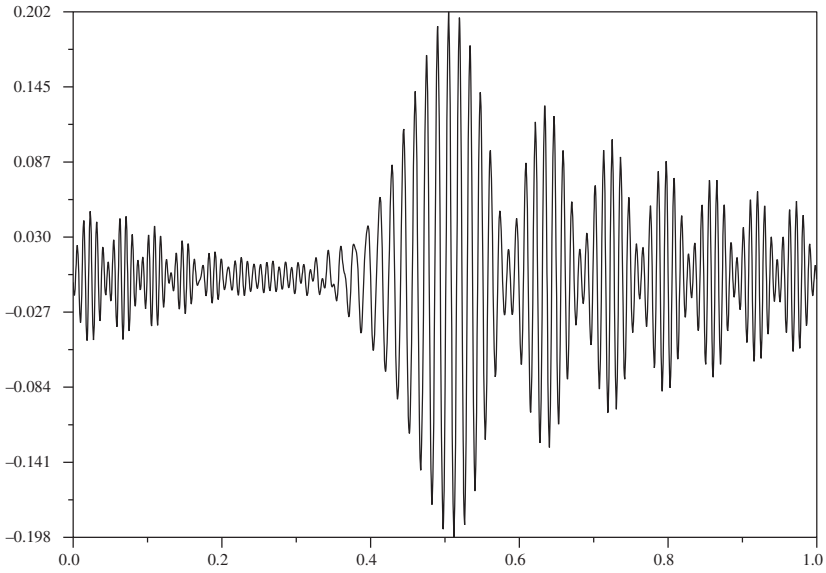

FIG. 2. *Real part of the approximate solution of the first component of system* (1.17) *with* $\varepsilon = 0.01$ *given by the nonlinear Schrödinger equation* (3.5) *at time* $T = 50$*, with chirped initial data, Case* 6.

2. New intermediate models.

2.1. Formal obtaining of the models. We restrict ourselves to geometrical optics regime. We go back to the singular equation (1.13). For $\xi \in \mathbb{R}^n$, let us introduce the following spectral decomposition of the matrix $iA(\xi) + L_0$:

$$iA(\xi) + L_0 = \sum_{j=1}^{m} i\lambda_j(\xi)\Pi_j(\xi), \tag{2.1}$$

where m denotes the number of distinct eigenvalues of $iA(\xi) + L_0$. We have implicitly used the following assumption.

Hypothesis 3. There exist m continuous functions $\xi \mapsto \lambda_j(\xi)$ defining a global parametrization of the characteristic variety C_L.

Of course the functions $\xi \mapsto \Pi_j(\xi)$ are not necessary continuous at the points ξ_0 where there exist j_1 and j_2 such that $\lambda_{j_1}(\xi_0) = \lambda_{j_2}(\xi_0)$. However, since the projector $\Pi_j(\xi)$ are orthogonal projectors, the functions $\xi \mapsto \Pi_j(\xi)$ are bounded. Let us now fix a vector $k \in \mathbb{R}^n$ and take $\omega = \lambda_{l_0}(k)$ one eigenvalue of $iA(\xi) + L_0$ for some $l_0 \in \{1, \ldots, m\}$. In order to simplify the notations, we take $l_0 = 1$. (k, ω) will be the main frequencies of the solution described in the introduction.

Hypothesis 4. There exists a neighborhood V of k such that for all $\xi \in V$ and for all integers $j \geq 2$,

$$\lambda_j(\xi) \neq \lambda_1(\xi).$$

From now on, we use the usual notations $D_\theta = \frac{\partial_\theta}{i}$ and $D_X = \frac{\partial_X}{i}$. Then, (1.13) reads

$$\left(\partial_t + \frac{1}{\varepsilon}(-i\omega D_\theta + iA(kD_\theta + \varepsilon D_X) + L_0)\right)\mathcal{U}^\varepsilon = f(\mathcal{U}^\varepsilon).$$

Using (2.1), we get

$$\left(\partial_t + \frac{1}{\varepsilon}(-i\omega D_\theta + i\lambda_j(kD_\theta + \varepsilon D_X))\right)\Pi_j(kD_\theta + \varepsilon D_X)\mathcal{U}^\varepsilon = \Pi_j(kD_\theta + \varepsilon D_X)f(\mathcal{U}^\varepsilon), \quad j = 1, \dots, m. \tag{2.2}$$

The first model that we introduce relies on the following idea: we want to obtain a model that is exact for the linear regime ($f \equiv 0$) and the best possible for the nonlinear one. Moreover, one starts with initial data that are polarized on the first eigenspace, that is,

$$\Pi_1(kD_\theta + \varepsilon D_X)\mathcal{U}^\varepsilon(t=0) = \mathcal{U}^\varepsilon(t=0).$$

We now make the following hypothesis.

Hypothesis 5. If $m \in \mathbb{Z}$, $j = 1, \dots, m$, then

$$\lambda_j(mk) = m\omega \Rightarrow j = 1 \text{ and } m = \pm 1.$$

One can modify the model obtained below if this assumption is not satisfied. In fact a generalized assumption is the following strong finiteness hypothesis introduced in [12].

Hypothesis 5′. The set $\{m \in \mathbb{Z}$ such that there exists j satisfying $\lambda_j(mk) = m\omega\}$ is finite.

However, for the sake of simplicity, we will restrict ourselves in this work to Hypothesis 5. Under Hypothesis 5, the spectrum of the solution will be mainly supported by the first sheet of the characteristic variety. That is, for all time, we will have $\Pi_1(kD_\theta + \varepsilon D_X)\mathcal{U}^\varepsilon(t) \approx \mathcal{U}^\varepsilon(t)$. We therefore introduce $\mathcal{V}^\varepsilon$ the solution to

$$\left(\partial_t + \frac{1}{\varepsilon}(-i\omega D_\theta + i\lambda_1(kD_\theta + \varepsilon D_X))\right)\Pi_1(kD_\theta + \varepsilon D_X)\mathcal{V}^\varepsilon = \Pi_1(kD_\theta + \varepsilon D_X)f(\mathcal{V}^\varepsilon), \tag{2.3}$$

and

$$\Pi_j(kD_\theta + \varepsilon D_X)\mathcal{V}^\varepsilon = 0 \quad \text{for } j \geq 2. \tag{2.4}$$

We expect $\mathcal{V}^\varepsilon$ to be a good approximation of $\mathcal{U}^\varepsilon$. For $s \in \mathbb{R}$ and $T > 0$ we denote $X_T = L^\infty(0, T; H^s(\mathbb{R}^n_X \times \mathbb{T}_\theta))$. Our first result reads as follows.

THEOREM 2.1. *Let us assume Hypotheses* 3, 4, *and* 5, *and let* $s > \frac{n+1}{2}$, $\alpha > 0$. *Let* $u_0(X) \in H^\sigma(\mathbb{R}^n)$ *(for* σ *large enough) satisfy*

$$\Pi_1(k + \varepsilon D_X)u_0(X) = u_0(X).$$

Then there exists $T > 0$ *(independent of* ε*) and there exist solution* $\mathcal{U}^\varepsilon$ *and* $\mathcal{V}^\varepsilon$, *respectively, to* (2.2), (2.3), and (2.4) *such that*

$$\mathcal{U}^\varepsilon(t=0) = \mathcal{V}^\varepsilon(t=0) = \varepsilon^\alpha(e^{i\theta}u_0 + c.c.).$$

Moreover,

$$\frac{1}{\varepsilon^\alpha}|\Pi_1(kD_\theta + \varepsilon D_X)(\mathcal{U}^\varepsilon - \mathcal{V}^\varepsilon)|_{X_T} = O(\varepsilon^{2\alpha(q-1)+1})$$

and

$$\frac{1}{\varepsilon^\alpha}|\Pi_j(kD_\theta+\varepsilon D_X)\mathcal{U}^\varepsilon|_{X_T}=O(\varepsilon^{\alpha(q-1)+1})\quad \textit{for } j\geq 2.$$

Remark 2.2. The scaling ε^α allows us to see how the error estimate evolves when the nonlinear effects decrease. Indeed, for large α, the nonlinear estimate is better than for small α. The case $\alpha=\infty$ corresponds to the linear regime and the solution is then exact.

Remark 2.3. As usual for the proofs using WKB-type method, we need a lot of regularity on the approximate solution $\mathcal{V}^\varepsilon$. Therefore, we will impose the initial data u_0 to be more regular than the space in which we want the error estimates [14].

Now, we can introduce a second model as follows. Thanks to Hypothesis 5, we expect the Fourier coefficients of order different from ± 1 of $\mathcal{V}^\varepsilon$ to be small. We therefore expect $\mathcal{V}^\varepsilon\approx\mathcal{V}_1^\varepsilon(t,X)e^{i\theta}+c.c.$. We therefore introduce the function $H^\varepsilon(t,X)$ solution to

$$\begin{aligned}\left(\partial_t+\frac{1}{\varepsilon}(-i\omega+i\lambda_1(k+\varepsilon D_X))\right)\Pi_1(k+\varepsilon D_X)H^\varepsilon\\ =\Pi_1(k+\varepsilon D_X)C_1(f(H^\varepsilon e^{i\theta}+c.c.))\end{aligned}\tag{2.5}$$

and we expect $H^\varepsilon e^{i\theta}+c.c.$ to be a good approximation of $\mathcal{V}^\varepsilon$ and hence of $\mathcal{U}^\varepsilon$. Our second result reads as follows.

THEOREM 2.4. *Under the same hypothesis for Theorem* 2.1*, there exist* T_0 *independent of* ε *such that* $T\geq T_0>0$*, a unique solution* $H^\varepsilon(t,X)\in L^\infty(0,T_0;H^s_X(\mathbb{R}^n))$ *to* (2.5) *satisfying* $H^\varepsilon(0,X)=\varepsilon^\alpha u_0(x)$*, and moreover*

$$\frac{1}{\varepsilon^\alpha}|C_1(\mathcal{V}^\varepsilon)-H^\varepsilon(t,X)|_{L^\infty(0,T_0;H^s_X(\mathbb{R}^n))}=O(\varepsilon^{2\alpha(q-1)+1})$$

and

$$\frac{1}{\varepsilon^\alpha}|\mathcal{V}^\varepsilon-(H^\varepsilon e^{i\theta}+c.c.)|_{X_T}=O(\varepsilon^{\alpha(q-1)+1}).$$

Remark 2.5. • The error estimate between $\mathcal{V}^\varepsilon$ and $H^\varepsilon e^{i\theta}+c.c.$ is of the same type as that between $\mathcal{V}^\varepsilon$ and $\mathcal{U}^\varepsilon$.

• The equation satisfied by H^ε is not stiff anymore since $\lambda_1(k)=\omega$.

2.2. Proofs of the theorems. We begin with the proof of Theorem 2.1. One first has an obvious existence result for (2.2) and (2.3).

PROPOSITION 2.6. *Let* $u_0(X)\in H^\sigma(\mathbb{R}^n)$ *(for* σ *large enough) and* $s>\frac{n+1}{2}$*. There exists* $T>0$ *(independent of* ε*) such that there exists a unique solution* $\mathcal{U}^\varepsilon$ *to* (2.2) *and there exists a unique solution* $\mathcal{V}^\varepsilon$ *to* (2.3) *satisfying*

$$\mathcal{U}^\varepsilon\in\mathcal{C}([0,T];H^s(\mathbb{R}^n_X\times\mathbb{T}_\theta)),\qquad \mathcal{V}^\varepsilon\in\mathcal{C}([0,T];H^s(\mathbb{R}^n_X\times\mathbb{T}_\theta)),$$

and

$$\mathcal{U}^\varepsilon(t=0,\theta,X)=\mathcal{V}^\varepsilon(t=0,\theta,X)=\varepsilon^\alpha(e^{i\theta}u_0(X)+c.c.).$$

Moreover, there exists C *independent of* ε *such that*

$$\frac{1}{\varepsilon^\alpha}|\mathcal{U}^\varepsilon|_{X_T}+\frac{1}{\varepsilon^\alpha}|\mathcal{V}^\varepsilon|_{X_T}\leq C.$$

This proposition is obtained by usual energy estimates. It is of course not sufficient in order to prove Theorem 2.1. Let us introduce

$$\mathcal{W}^\varepsilon = \frac{1}{\varepsilon^\alpha}(\mathcal{U}^\varepsilon - \mathcal{V}^\varepsilon), \tag{2.6}$$

and we consider the following decomposition of $\mathcal{W}^\varepsilon$:

$$\begin{aligned}\mathcal{W}^\varepsilon &:= \Pi_1(kD_\theta + \varepsilon D_X)\mathcal{W}^\varepsilon + \sum_{j=2}^{m} \Pi_j(kD_\theta + \varepsilon D_X)\mathcal{W}^\varepsilon \\ &:= \varepsilon^{2\alpha(q-1)+1}a + \sum_{j=2}^{m} \varepsilon^{\alpha(q-1)+1}b_j.\end{aligned} \tag{2.7}$$

In order to prove Theorem 2.1, it is enough to show that the functions a and b_j are bounded in $X_T = L^\infty([0,T]; H^s(\mathbb{R}^n_X \times \mathbb{T}_\theta))$. Let us now write the equations satisfied respectively by a and b_j. Let us form the difference of (2.3) from (2.2) and then apply the projector Π_j. Decomposition of (2.1) yields (using the fact that f is a homogeneous polynomial of degree q)

$$\begin{aligned}&\left(\partial_t + \frac{1}{\varepsilon}(-i\omega D_\theta + i\lambda_1(kD_\theta + \varepsilon D_X))\right) a = \frac{1}{\varepsilon^{\alpha(q-1)+1}}\Pi_1(kD_\theta + \varepsilon D_X) \\ &\cdot \left[f\left(\varepsilon^{2\alpha(q-1)+1}a + \sum_{j=2}^{m} \varepsilon^{\alpha(q-1)+1}b_j + \frac{1}{\varepsilon^\alpha}\mathcal{V}^\varepsilon\right) - f\left(\frac{1}{\varepsilon^\alpha}\mathcal{V}^\varepsilon\right)\right]\end{aligned} \tag{2.8}$$

and

$$\begin{aligned}&\left(\partial_t + \frac{1}{\varepsilon}(-i\omega D_\theta + i\lambda_j(kD_\theta + \varepsilon D_X))\right) b_j \\ &= \frac{1}{\varepsilon}\Pi_j(kD_\theta + \varepsilon D_X)\left[f\left(\varepsilon^{2\alpha(q-1)+1}a + \sum_{j=2}^{m} \varepsilon^{\alpha(q-1)+1}b_j + \frac{1}{\varepsilon^\alpha}\mathcal{V}^\varepsilon\right)\right]\end{aligned} \tag{2.9}$$

for $j = 2, \dots, m$. We start with (2.8). We first use Taylor's formula in the right-hand side of (2.8):

$$\begin{aligned}&f\left(\varepsilon^{2\alpha(q-1)+1}a + \sum_{j=2}^{m} \varepsilon^{\alpha(q-1)+1}b_j + \frac{1}{\varepsilon^\alpha}\mathcal{V}^\varepsilon\right) - f\left(\frac{1}{\varepsilon^\alpha}\mathcal{V}^\varepsilon\right) \\ &= \int_0^1 f'\left(\frac{1}{\varepsilon^\alpha}\mathcal{V}^\varepsilon + \nu\left(\varepsilon^{2\alpha(q-1)+1}a + \sum_{j=2}^{m} \varepsilon^{\alpha(q-1)+1}b_j\right)\right) \\ &\cdot \left(\varepsilon^{2\alpha(q-1)+1}a + \sum_{j=2}^{m} \varepsilon^{\alpha(q-1)+1}b_j\right) d\nu.\end{aligned}$$

f is a homogeneous polynomial of degree q since H^s is an algebra for s large enough, hence this quantity can be estimated in $H^s_{\theta,X}$-norm by

$$\Delta_1(t)= \left| f\left(\varepsilon^{2\alpha(q-1)+1}a+\sum_{j=2}^m \varepsilon^{\alpha(q-1)+1}b_j+\frac{1}{\varepsilon^\alpha}\mathcal{V}^\varepsilon\right)-f\left(\frac{1}{\varepsilon^\alpha}\mathcal{V}^\varepsilon\right)\right|_{H^s}$$

$$\leq C\varepsilon^{\alpha(q-1)+1}\left(\left|\frac{\mathcal{V}^\varepsilon}{\varepsilon^\alpha}\right|^{q-1}_{H^s}+|a|^{q-1}_{H^s}+\sum_{j=2}^m|b_j|^{q-1}_{H^s}\right)\cdot\left(|a|_{H^s}+\sum_{j=2}^m|b_j|_{H^s}\right). \tag{2.10}$$

We now use an integral formulation of (2.8)

$$a=e^{-\frac{1}{\varepsilon}(-i\omega D_\theta+i\lambda_1(kD_\theta+\varepsilon D_X))t}a(t=0)$$

$$+\int_0^t\frac{1}{\varepsilon^{\alpha(q-1)+1}}e^{-\frac{1}{\varepsilon}(-i\omega D_\theta+i\lambda_1(kD_\theta+\varepsilon D_X))(t-\tau)}\Pi_1(kD_\theta+\varepsilon D_X)$$

$$\cdot\left[f\left(\varepsilon^{2\alpha(q-1)+1}a(\tau)+\sum_{j=2}^m\varepsilon^{\alpha(q-1)+1}b_j(\tau)+\frac{1}{\varepsilon^\alpha}\mathcal{V}^\varepsilon(\tau)\right)-f\left(\frac{1}{\varepsilon^\alpha}\mathcal{V}^\varepsilon(\tau)\right)\right]d\tau,$$

and using (2.10),

$$|a|_{H^s}(t)\leq|a(0)|_{H^s}+C\int_0^t\left(\left|\frac{\mathcal{V}^\varepsilon}{\varepsilon^\alpha}\right|^{q-1}_{H^s}+|a|^{q-1}_{H^s}+\sum_{j=2}^m|b_j|^{q-1}_{H^s}\right)\cdot\left(|a|_{H^s}+\sum_{j=2}^m|b_j|_{H^s}\right)d\tau.$$

Using the fact that $\frac{1}{\varepsilon^\alpha}|\mathcal{V}^\varepsilon|_{H^s}$ is bounded, thanks to Proposition 2.6, and that $a(t=0)=0$, one gets

$$|a|_{H^s}(t)\leq C\int_0^t\left(1+|a|_{H^s}+\sum_{j=2}^m|b_j|_{H^s}\right)^q(\tau)d\tau. \tag{2.11}$$

We now deal with (2.9). The main point is to recover one power of ε with respect to the right-hand side using the "elliptic inversion" corresponding to the operator $-i\omega D_\theta+i\lambda_j(kD_\theta+\varepsilon D_X)$. We first rewrite (2.9) as follows:

$$\left(\partial_t+\frac{1}{\varepsilon}\left(-i\omega D_\theta+i\lambda_j(kD_\theta+\varepsilon D_X)\right)\right)b_j$$

$$=\frac{1}{\varepsilon}\Pi_j(kD_\theta+\varepsilon D_X)\left[f\left(\varepsilon^{2\alpha(q-1)+1}a+\sum_{j=2}^m\varepsilon^{\alpha(q-1)+1}b_j+\frac{1}{\varepsilon^\alpha}\mathcal{V}^\varepsilon\right)\right], \tag{2.12}$$

we write the nonlinear term in the form

$$f\left(\varepsilon^{2\alpha(q-1)+1}a+\sum_{j=2}^m\varepsilon^{\alpha(q-1)+1}b_j+\frac{1}{\varepsilon^\alpha}\mathcal{V}^\varepsilon\right)-f\left(\frac{\mathcal{V}^\varepsilon}{\varepsilon^\alpha}\right)+f\left(\frac{\mathcal{V}^\varepsilon}{\varepsilon^\alpha}\right).$$

An integral formula for (2.12) gives

$$
\begin{aligned}
b_j = \frac{1}{\varepsilon}\int_0^t e^{-\frac{1}{\varepsilon}(-i\omega D_\theta + i\lambda_j(kD_\theta+\varepsilon D_X))(t-\tau)}\Pi_j(kD_\theta+\varepsilon D_X)\\
\cdot\left[f\left(\varepsilon^{2\alpha(q-1)+1}a+\sum_{j=2}^m \varepsilon^{\alpha(q-1)+1}b_j+\frac{\mathcal{V}^\varepsilon}{\varepsilon^\alpha}\right)-f\left(\frac{\mathcal{V}^\varepsilon}{\varepsilon^\alpha}\right)\right](\tau)d\tau \qquad (2.13)\\
+\frac{1}{\varepsilon}\int_0^t e^{-\frac{1}{\varepsilon}(-i\omega D_\theta + i\lambda_j(kD_\theta+\varepsilon D_X))(t-\tau)}\Pi_j(kD_\theta+\varepsilon D_X)f\left(\frac{\mathcal{V}^\varepsilon}{\varepsilon^\alpha}\right)(\tau)d\tau\\
:= c_j + d_j.
\end{aligned}
$$

Obviously, one has in the same way that for the estimate of a

$$
|c_j|_{H^s}(t) \le C\int_0^t\left(1+|a|_{H^s}+\sum_{j=2}^m |b_j|_{H^s}\right)^q(\tau)d\tau. \qquad (2.14)
$$

We still have to estimate the term d_j. The idea is to perform the elliptic inversion on the nonlinear term associated with $\mathcal{V}^\varepsilon$ (that is, $f(\mathcal{V}^\varepsilon)$ which is relatively well known (at least asymptotically)). We introduce $\beta_j(D) = -\omega D_\theta + \lambda_j(kD_\theta+\varepsilon D_X)$ and the term d_j can be therefore written as

$$
d_j = \frac{1}{\varepsilon}\int_0^t \Pi_j(kD_\theta+\varepsilon D_X)e^{-\frac{i}{\varepsilon}\beta_j(D)(t-\tau)}f\left(\frac{\mathcal{V}^\varepsilon}{\varepsilon^\alpha}\right)(\tau)d\tau. \qquad (2.15)
$$

In order to use the oscillatory behavior of the exponential, we split the function $\mathcal{V}^\varepsilon$ into low-frequency and high-frequency parts

$$
\begin{aligned}
\mathcal{V}^\varepsilon &= \mathbf{1}_{\{|D_X|\le\frac{1}{\sqrt{\varepsilon}}\}}\mathcal{V}^\varepsilon + \mathbf{1}_{\{|D_X|>\frac{1}{\sqrt{\varepsilon}}\}}\mathcal{V}^\varepsilon\\
&:= \mathcal{V}_1^\varepsilon + \mathcal{V}_2^\varepsilon.
\end{aligned}
$$

Again, we write d_j as follows:

$$
\begin{aligned}
d_j = \frac{1}{\varepsilon}\int_0^t \Pi_j(kD_\theta+\varepsilon D_X)e^{-\frac{i}{\varepsilon}\beta_j(D)(t-\tau)}\left[f\left(\frac{\mathcal{V}^\varepsilon(\tau)}{\varepsilon^\alpha}\right)-f\left(\frac{\mathcal{V}_1^\varepsilon(\tau)}{\varepsilon^\alpha}\right)\right]d\tau\\
+\frac{1}{\varepsilon}\int_0^t \Pi_j(kD_\theta+\varepsilon D_X)e^{-\frac{i}{\varepsilon}\beta_j(D)(t-\tau)}f\left(\frac{\mathcal{V}_1^\varepsilon(\tau)}{\varepsilon^\alpha}\right)d\tau \qquad (2.16)\\
:= e_j + f_j.
\end{aligned}
$$

We begin by estimating e_j

$$
e_j = \frac{1}{\varepsilon}\int_0^t \Pi_j(kD_\theta+\varepsilon D_X)e^{-\frac{i}{\varepsilon}\beta_j(D)(t-\tau)}\int_0^1 f'\left(\frac{\mathcal{V}_1^\varepsilon+\alpha\mathcal{V}_2^\varepsilon}{\varepsilon^\alpha}\right)\cdot\frac{\mathcal{V}_2^\varepsilon}{\varepsilon^\alpha}\,d\alpha d\tau
$$

and

$$
|e_j|_{H^s} \le \frac{1}{\varepsilon}\int_0^t\left(\left|\frac{\mathcal{V}_1^\varepsilon}{\varepsilon^\alpha}\right|_{H^s}^{q-1}+\left|\frac{\mathcal{V}_2^\varepsilon}{\varepsilon^\alpha}\right|_{H^s}^{q-1}\right)\frac{|\mathcal{V}_2^\varepsilon|_{H^s}}{\varepsilon^\alpha}\,d\tau.
$$

Now since

$$\frac{|\mathcal{V}_i^\varepsilon|_{H^s}}{\varepsilon^\alpha} \leq \frac{|\mathcal{V}^\varepsilon|_{H^s}}{\varepsilon^\alpha}$$

for $i = 1, 2$ and thanks to Proposition 2.6, $\frac{|\mathcal{V}^\varepsilon|_{H^s}}{\varepsilon^\alpha}$ is bounded, one has

$$\left|\frac{\mathcal{V}_1^\varepsilon}{\varepsilon^\alpha}\right|_{H^s}^{q-1} + \left|\frac{\mathcal{V}_2^\varepsilon}{\varepsilon^\alpha}\right|_{H^s}^{q-1} \leq C$$

and gets

$$|e_j|_{H^s} \leq \frac{C}{\varepsilon}\int_0^t \frac{|\mathcal{V}_2^\varepsilon(\tau)|_{H^s}}{\varepsilon^\alpha}\, d\tau.$$

Moreover, for all $N \in \mathbb{N}$ and for all $s \in \mathbb{R}$

$$\left|\mathcal{V}^\varepsilon \mathbf{1}_{\{|D_X|>\frac{1}{\sqrt{\varepsilon}}\}}\right|_{H^s} \leq C\varepsilon^N\, |\mathcal{V}^\varepsilon|_{H^{s+2N}}$$

and therefore

$$|e_j|_{H^s}(t) \leq Ct. \tag{2.17}$$

We now deal with the term f_j:

$$f_j = \frac{1}{\varepsilon}\int_0^t \Pi_j(kD_\theta + \varepsilon D_X)e^{-\frac{i}{\varepsilon}\beta_j(D)(t-\sigma)} f\left(\frac{\mathcal{V}_1^\varepsilon(\sigma)}{\varepsilon^\alpha}\right).$$

Now thanks to Hypothesis 4, one can apply the following result of nonlinear geometrical optics (see [14]): there exists a regular function $F(t, X)$ (independent of ε) such that

$$\frac{\mathcal{V}^\varepsilon}{\varepsilon^\alpha} = F(t, X)e^{i\theta} + c.c. + O(\varepsilon),$$

the $O(\varepsilon)$ being for example in $L^\infty(0, T; H^s_{\theta,X}(\mathbb{R}^n \times \mathbb{T}))$-norm. Plugging this expression into the expression of f_j yields

$$\begin{aligned} f_j = \frac{1}{\varepsilon}\int_0^t \Pi_j(kD_\theta + \varepsilon D_X)e^{-\frac{i}{\varepsilon}\beta_j(D)(t-\sigma)} f\left(\mathbf{1}_{\{|D_X|\leq\frac{1}{\sqrt{\varepsilon}}\}}(F(t, X)e^{i\theta} + c.c.)\right) d\sigma \\ + tO(1) := h_j + O(t). \end{aligned}$$

Now, since f is a homogeneous polynomial of degree q,

$$f\left(\mathbf{1}_{\{|D_X|\leq\frac{1}{\sqrt{\varepsilon}}\}}(F(t, X)e^{i\theta} + c.c.)\right)$$

has the form

$$f\left(\mathbf{1}_{\{|D_X|\leq\frac{1}{\sqrt{\varepsilon}}\}}(F(t, X)e^{i\theta} + c.c.)\right) = \sum_{\beta=-q}^{q} a_\beta^\varepsilon(t, X)e^{i\beta\theta},$$

where $a_\beta^\varepsilon(t, X)$ are regular functions, bounded independently of ε in spaces like $W^{k,\infty}(0, T; H^s_X(\mathbb{R}^n))$ for k large enough. Moreover, since the a_β^ε are products of

components of $\mathbf{1}_{\{|D_X|\leq\frac{1}{\sqrt{\varepsilon}}\}}F$ and $\mathbf{1}_{\{|D_X|\leq\frac{1}{\sqrt{\varepsilon}}\}}\bar{F}$, the support of the Fourier transform of a^ε_β is included in $\{\xi||\xi|\leq\frac{q}{\sqrt{\varepsilon}}\}$. Taking the Fourier transform of h_j with respect to θ and X (denoting by $l\in\mathbb{Z}$ and $\xi\in\mathbb{R}^n$ the dual variables of θ and X) gives

$$\hat{h}_j(l,\xi,t)=\frac{1}{\varepsilon}\int_0^t \Pi_j(kl+\varepsilon\xi)e^{-\frac{i}{\varepsilon}[-l\omega+\lambda_j(kl+\varepsilon\xi)](t-\tau)}\hat{a}^\varepsilon_l(\tau,\xi)d\tau \tag{2.18}$$

for $l=-q,\dots,q$. Now thanks to Hypothesis 5, for all l, $l\omega\neq\lambda_j(kl)$ since $j>1$. Moreover, since the support of $\xi\mapsto\hat{a}^\varepsilon_l(s,\xi)$ is included in $\{\xi/\quad|\xi|\leq\frac{q}{\sqrt{\varepsilon}}\}$, it follows that there exist $\varepsilon_0>0$ and $\delta>0$ such that for all $\varepsilon\leq\varepsilon_0$, for all $l=-q$ to q, and for all $\xi\in\{\xi/\quad|\xi|\leq\frac{q}{\sqrt{\varepsilon}}\}$,

$$|-l\omega+\lambda_j(kl+\varepsilon\xi)|\geq\delta. \tag{2.19}$$

We perform an integration by parts in time on (2.18) and get

$$\begin{aligned}\hat{h}_j(l,\xi,t)=\frac{1}{\varepsilon}&\left[\frac{-i\varepsilon}{-l\omega+\lambda_j(kl+\varepsilon\xi)}e^{-\frac{i}{\varepsilon}[-l\omega+\lambda_j(kl+\varepsilon\xi)](t-\tau)}\Pi_j(kj+\varepsilon\xi)\hat{a}^\varepsilon_l(\tau,\xi)\right]_0^t\\&+\frac{1}{\varepsilon}\int_0^t\frac{i\varepsilon}{-l\omega+\lambda_j(kl+\varepsilon\xi)}e^{-\frac{i}{\varepsilon}[-l\omega+\lambda_j(kl+\varepsilon\xi)](t-\tau)}\Pi_j(kj+\varepsilon\xi)\partial_s\hat{a}^\varepsilon_l(\tau,\xi)d\tau.\end{aligned}$$

Therefore using (2.19),

$$|\hat{h}_j(l,\xi,t)|\leq\frac{1}{\delta}\left(|\hat{a}^\varepsilon_l(t,\xi)|+|\hat{a}^\varepsilon_l(0,\xi)|\right)+\frac{1}{\delta}\int_0^t|\partial_\tau\hat{a}^\varepsilon_l(\tau,\xi)|d\tau$$

for all $l=-q,\dots,q$. It follows that

$$|h_j|_{H^s}(t)\leq C(1+t). \tag{2.20}$$

One deduces that

$$|f_j|_{H^s}(t)\leq C(1+t),$$

and with (2.16) and (2.17) we get

$$|d_j|_{H^s}(t)\leq C(1+t).$$

Equality (2.13) and estimate (2.14) give together with the above control of d_j

$$|b_j|_{H^s}\leq C\int_0^t\left(1+|a|_{H^s}+\sum_{j=2}^m|b_j|_{H^s}\right)^q(\tau)d\tau+C(1+t). \tag{2.21}$$

Now we recall the estimate (2.11) of a:

$$|a|_{H^s}(t)\leq C\int_0^t\left(1+|a|_{H^s}+\sum_{j=2}^m|b_j|_{H^s}\right)^q(\tau)d\tau.$$

Introducing $y=1+|a|_{H^s}+\sum_{j=2}^m|b_j|_{H^s}$, one gets using (2.21)

$$y\leq c\int_0^t y^q(\sigma)d\sigma+C(1+t)$$

which implies that there exist $T_0>0$ and $C_0>0$ such that y is defined on $[0,T_0]$ and $|y|_{L^\infty(0,T_0)}\leq C_0$. This ends the proof of Theorem 2.1.

2.3. Proof of Theorem 2.4. We will now compare the solution $\mathcal{V}^\varepsilon$ given by (2.3) and (2.4) and H^ε given by (2.5). The proof is mainly the same as that for the previous result, we only sketch it. Introduce

$$\mathcal{V}^\varepsilon = \sum_{\beta\in\mathbb{Z}} \mathcal{V}^\varepsilon_\beta(t,X)e^{i\beta\theta}.$$

The equation satisfied by $\mathcal{V}^\varepsilon_\beta$ is

$$\left(\partial_t + \frac{1}{\varepsilon}\left(-i\omega\beta + i\lambda_1(k\beta+\varepsilon D_X)\right)\right)\Pi_1(k\beta+\varepsilon D_X)\mathcal{V}^\varepsilon_\beta = \Pi_1(k\beta+\varepsilon D_X)C_\beta\left(f(\mathcal{V}^\varepsilon)\right).$$

Introduce $X^\varepsilon = \frac{1}{\varepsilon^\alpha}[\mathcal{V}^\varepsilon_1 - H^\varepsilon]$ where H^ε is the solution to (2.5). The equation satisfied by $\frac{\mathcal{V}^\varepsilon_\beta}{\varepsilon^\alpha}$ for $\beta \neq \pm 1$ is

$$\begin{aligned}&\left(\partial_t + \frac{1}{\varepsilon}\left(-i\omega\beta + i\lambda_1(k\beta+\varepsilon D_X)\right)\right)\Pi_1(k\beta+\varepsilon D_X)\frac{\mathcal{V}^\varepsilon_\beta}{\varepsilon^\alpha}\\ &\qquad = \Pi_1(k\beta+\varepsilon D_X)\varepsilon^{\alpha(q-1)}C_\beta f\left(\frac{\mathcal{V}^\varepsilon_\beta}{\varepsilon^\alpha}\right).\end{aligned}$$

An elliptic inversion on $\frac{\mathcal{V}^\varepsilon_\beta}{\varepsilon^\alpha}$ gives an estimate of $\frac{\mathcal{V}^\varepsilon_\beta}{\varepsilon^\alpha}$ of size $\varepsilon^{\alpha(q-1)+1}$ for $\beta \neq \pm 1$. Now the equation satisfied by X^ε is

$$\begin{aligned}&\left(\partial_t + \frac{1}{\varepsilon}\left(-i\omega + i\lambda_1(k+\varepsilon D_X)\right)\right)\Pi_1(k+\varepsilon D_X)X^\varepsilon\\ &\qquad = \Pi_1(k+\varepsilon D_X)\varepsilon^{\alpha(q-1)}\left[C_1\left(f\left(\frac{\mathcal{V}^\varepsilon_\beta}{\varepsilon^\alpha}\right)\right) - C_1\left(f\left(\frac{H^\varepsilon e^{i\theta}+c.c.}{\varepsilon^\alpha}\right)\right)\right].\end{aligned} \tag{2.22}$$

Now write $\mathcal{V}^\varepsilon = \mathcal{V}^\varepsilon_1 e^{i\theta} + c.c. + \tilde{\mathcal{V}}^\varepsilon$. Then the right-hand side of (2.22) reads

$$\begin{aligned}&C_1\left(f\left(\frac{\mathcal{V}^\varepsilon_\beta}{\varepsilon^\alpha}\right)\right) - C_1\left(f\left(\frac{H^\varepsilon e^{i\theta}+c.c.}{\varepsilon^\alpha}\right)\right)\\ &\quad = C_1\left(f\left(\frac{H^\varepsilon e^{i\theta}+c.c.}{\varepsilon^\alpha} + X^\varepsilon e^{i\theta} + c.c. + \frac{\tilde{\mathcal{V}}^\varepsilon}{\varepsilon^\alpha}\right) - f\left(\frac{H^\varepsilon e^{i\theta}+c.c.}{\varepsilon^\alpha}\right)\right)\\ &\quad \approx C_1\left(f'\left(\frac{H^\varepsilon e^{i\theta}+c.c.}{\varepsilon^\alpha}\right)\left[X^\varepsilon e^{i\theta} + c.c. + \frac{\tilde{\mathcal{V}}^\varepsilon}{\varepsilon^\alpha}\right]\right).\end{aligned}$$

Integrating (2.22) in time gives

$$|X^\varepsilon|_{H^s}(t) \le \int_0^t \varepsilon^{\alpha(q-1)}C|X^\varepsilon|_{H^s}(\sigma)d\sigma + \int_0^t C\varepsilon^{\alpha(q-1)}\left|\frac{\tilde{\mathcal{V}}^\varepsilon}{\varepsilon^\alpha}\right|_{H^s} d\sigma.$$

But $\frac{\tilde{\mathcal{V}}^\varepsilon}{\varepsilon^\alpha} = O(\varepsilon^{\alpha(q-1)+1})$. It follows that

$$|X|_{L^\infty(0,T;H^s)} = O(\varepsilon^{2\alpha(q-1)+1})$$

which is the desired result.

2.4. Some extensions. Note that if $\alpha = 0$, that is, for $O(1)$ solutions the error estimate is the same as that for usual geometrical optics. The estimate is in fact better for $\alpha > 0$. Recall that ε^α is the size of the initial data and hence of the solution. But if $\alpha > 0$, then standard techniques on (1.13) ensures existence of time of size $\frac{1}{\varepsilon^{\alpha(q-1)}}$. The natural question is then to know if our estimates are valid on such time interval. The answer is affirmative and one has the following theorem.

THEOREM 2.7. *Under the same hypothesis as that for Theorem* 2.1*, there exist* $T_1 > 0$ *and* $C_1 > 0$ *(independent of* ε*) such that*

$$\frac{1}{\varepsilon^\alpha} \left|\Pi_1(kD_\theta + \varepsilon D_X)\left(\mathcal{U}^\varepsilon - \mathcal{V}^\varepsilon\right)\right|_{L^\infty(0,t;H^s_{\theta,X}(\mathbb{T}\times\mathbb{R}^n))} \leq C_1 \varepsilon^{\alpha(q-1)+1}\left(e^{C_1\varepsilon^{\alpha(q-1)}t} - 1\right)$$

and

$$\frac{1}{\varepsilon^\alpha} \left|\Pi_1(kD_\theta + \varepsilon D_X)\mathcal{U}^\varepsilon\right|_{L^\infty(0,t;H^s_{\theta,X}(\mathbb{T}\times\mathbb{R}^n))} \leq C_1 \varepsilon^{\alpha(q-1)+1} t$$

as long as $t \leq \frac{T_1}{\varepsilon^{\alpha(q-1)}}$*. Moreover,*

$$\frac{1}{\varepsilon^\alpha} |C_1(\mathcal{V}^\varepsilon(t,X,\theta)) - H^\varepsilon(t,X)|_{L^\infty(0,t;H^s_X(\mathbb{R}^n))} \leq C_1 \varepsilon^{\alpha(q-1)+1}\left(e_1^{C_1\varepsilon^{\alpha(q-1)}} - 1\right)$$

and

$$\frac{1}{\varepsilon^\alpha} |\mathcal{V}^\varepsilon(t,X,\theta) - (H^\varepsilon(t,X)e^{i\theta} + c.c.)|_{L^\infty(0,t;H^s_{\theta,X}(\mathbb{T}\times\mathbb{R}^n))} \leq C_1 \varepsilon^{\alpha(q-1)+1} t$$

as long $t \leq \frac{T_1}{\varepsilon^{\alpha(q-1)}}$.

That means that our asymptotics are uniform on long-time interval. See the next section for numerical illustrations of these results.

Remark 2.8. Suppose that for all $X \in \mathbb{R}^n$, $f(X) \cdot X = 0$, then for any solution $\mathcal{V}^\varepsilon$ to (2.3) one has

$$\int |\mathcal{V}^\varepsilon|^2(t)\, dX d\theta = \int |\mathcal{V}^\varepsilon|^2(0)\, dX d\theta$$

and for any solution H^ε to (2.5),

$$\int |H^\varepsilon|(t)\, dX = \int |H^\varepsilon|(0)\, dX.$$

That means that if the initial model is conservative, then the asymptotic one is conservative as well.

3. Some numerical results.

3.1. An example. In this section, we want to compare numerically the solutions of the different asymptotic regimes and we want to see to which extent the error estimates that we have proved in the previous section are effective. We choose to make the computations on a simplified system which is dispersive, nonlinear, and preserves the L^2-norm. This system is

$$\begin{cases} \partial_t u + \partial_x v - \dfrac{v}{\varepsilon} = -(u^2+v^2)v, \\ \partial_t v + \partial_x u + \dfrac{u}{\varepsilon} = (u^2+v^2)u. \end{cases} \tag{3.1}$$

The characteristic variety of this system is the set

$$\{(\omega,k)\in\mathbb{R}^2|\omega^2=1+k^2\}.$$

Hypotheses 3 and 4 are therefore satisfied. For Hypothesis 5, suppose that $\omega^2=1+k^2$ and that for $m\in\mathbb{Z}$ one has $m^2\omega^2=1+m^2k^2$. It then follows that $m=\pm1$ and Hypothesis 5 is satisfied.

We now derive the asymptotic models corresponding to (3.1) in the geometrical and diffractive regimes. We refer the reader, for example, to [13] for the case of diffractive optics.

3.1.1. The geometrical optics regime. One searches an approximate solution in the form

$$\begin{pmatrix}u_0(t,x)\\v_0(t,x)\end{pmatrix}e^{i\frac{kx-\omega t}{\varepsilon}}+c.c.$$

Then one obtains

$$u_0=\frac{ik-1}{i\omega}v_0 \tag{3.2}$$

and

$$\partial_t u_0+\frac{k}{\omega}\partial_x u_0=\frac{4i}{\omega}|u_0|^2u_0\quad\text{for } t\in[0,T_0]. \tag{3.3}$$

3.1.2. Diffractive optics. One search an approximate solution in the form

$$\begin{pmatrix}u_1(t,x)\\v_1(t,x)\end{pmatrix}e^{i\frac{kx-\omega t}{\varepsilon}}+c.c.,$$

but on long time-scale with $u_1(0,x)=O(\sqrt{\varepsilon})$ and $v_1(0,x)=O(\sqrt{\varepsilon})$. One gets

$$u_1=\frac{ik-1}{i\omega}v_1 \tag{3.4}$$

and

$$\partial_t u_1+\frac{k}{\omega}\partial_x u_1-\frac{i\varepsilon}{\omega^3}\partial_x^2u_1=\frac{4i}{\omega}|u_1|^2u_1\quad\text{for } t\in\left[0,\frac{T_1}{\varepsilon}\right]. \tag{3.5}$$

3.1.3. The new model. One searches for a solution in the form

$$\begin{pmatrix}u_2(t,x)\\v_2(t,x)\end{pmatrix}e^{i\frac{kx-\omega t}{\varepsilon}}+c.c.$$

and one gets

$$u_2=\frac{i(k+\varepsilon D_x)-1}{i\sqrt{1+(k+\varepsilon D_x)^2}}v_2 \tag{3.6}$$

and

$$\begin{aligned}&\partial_t\begin{pmatrix}u_2\\v_2\end{pmatrix}+\frac{i}{\varepsilon}\left(\sqrt{1+(k+\varepsilon D_x)^2}-\sqrt{1+k^2}\right)\begin{pmatrix}u_2\\v_2\end{pmatrix}\\&=\begin{pmatrix}\frac{1}{2}&\frac{i(k+\varepsilon D_x)-1}{2i\sqrt{1+(k+\varepsilon D_x)^2}}\\\frac{i(k+\varepsilon D_x)+1}{2i\sqrt{1+(k+\varepsilon D_x)^2}}&\frac{1}{2}\end{pmatrix}\begin{pmatrix}-2(|u_2|^2+|v_2|^2)v_2-(u_2^2+v_2^2)\bar{v}_2\\-2(|u_2|^2+|v_2|^2)u_2-(u_2^2+v_2^2)\bar{u}_2\end{pmatrix}.\end{aligned} \tag{3.7}$$

Of course, thanks to (3.6), we can restrict ourselves to the first equation of (3.7) and setting

$$\mu^\varepsilon(D_x) = \frac{i(k+\varepsilon D_x)-1}{i\sqrt{1+(k+\varepsilon D_x)^2}},$$

one obtains

$$\partial_t u_2 + \frac{k\partial_x - i\varepsilon\partial_x^2}{\sqrt{1+(k+\varepsilon D_x)^2}+\sqrt{1+k^2}}u_2 = -(|u_2|^2+|v_2|^2)v_2 - 2(u_2^2+v_2^2)\bar{v}_2 + \mu^\varepsilon(D_x)\left[(|u_2|^2+|v_2|^2)u_2 + 2(u_2^2+v_2^2)\bar{u}_2\right] \tag{3.8}$$

with

$$v_2 = \frac{1}{\mu^\varepsilon(D_x)}u_2, \tag{3.9}$$

which is the complete system.

Finally, the same system with the Kerr nonlinearity is used in practical applications in [17] and reads

$$\partial_t u_3 + \frac{k\partial_x - i\varepsilon\partial_x^2}{\sqrt{1+(k+\varepsilon D_x)^2}+\sqrt{1+k^2}}u_3 = \frac{4i}{\omega}|u_3|^2u_3. \tag{3.10}$$

We will also compare our system with that one.

3.2. The numerical method. We restrict ourselves to the case $x \in [0,1]$ with periodic boundary conditions and we use a spectral method in the space variable x. For time discretization, we adopt a splitting technique.

• For system (3.1), suppose we have built an approximate solution $(u(n\delta t), v(n\delta t))$ at time $n\delta t$; one first integrates the linear part explicitly in Fourier variables with initial data $(u(n\delta t), v(n\delta t))$ over one time step. This gives an indermediate value (u_i, v_i). Then one integrates the nonlinear part

$$\partial_t \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} -\big(|u|^2+|v|^2\big)v \\ \big(|u|^2+|v|^2\big)u \end{pmatrix}$$

with initial value (u_i, v_i) explicitly over one time step. This gives $(u((n+1)\delta t), v((n+1)\delta t)$.

• For the geometrical optics (3.3) one has the exact solution

$$u_0(t,x) = A\left(x - \frac{k}{\omega}t\right)e^{\frac{4i}{\omega}|A(x-\frac{k}{\omega}t)|^2 t}, \tag{3.11}$$

where $A(x) = u_0(0,x)$.

• For the diffractive regime (3.5) we use the same strategy as that for (3.1). We omit the details since it is a standard procedure for the nonlinear Schrödinger equation (see [4] and the references therein for a more detailed study).

For the new model (3.8), suppose that one has the Fourier transform of $u_2(n\delta t)$: $\hat{u}_2(n\delta t)$. One solves the linear part of (3.8):

$$\partial_t \hat{u}_2 + \frac{ik\xi + \varepsilon\xi^2}{\sqrt{1+(k+\varepsilon\xi)^2}+\sqrt{1+k^2}}\hat{u}_2 = 0$$

with initial value $\hat{u}_2(n\delta t, \xi)$ on one time step. One gets an intermediate value $\hat{u}_{2i}$. One then obtains an intermediate value of $\hat{v}_2$ called $\hat{v}_{2i}$ using (3.9). We then perform an inverse Fourier transform of $\hat{u}_{2i}$ and $\hat{v}_{2i}$ in order to obtain u_{2i} and v_{2i} and then one constructs the nonlinear terms

$$NL1 := -(|u_{2i}|^2 + |v_{2i}|^2)v_{2i} - 2(u_{2i}^2 + v_{2i}^2)\bar{v}_{2i}$$

and

$$NL2 := (|u_{2i}|^2 + |v_{2i}|^2)u_{2i} + 2(u_{2i}^2 + v_{2i}^2)\bar{u}_{2i}.$$

Next we perform a Fourier transform of $NL1$ and $NL2$ and compute $\hat{NL}1+\mu^\varepsilon(\xi)\hat{NL}2$. The value of $\hat{u}_2((n+1)\delta t)$ is obtained by the explicit Euler scheme

$$\hat{u}_2((n+1)\delta t) = \hat{u}_{2i} + \delta t[\hat{NL}1 + \mu^\varepsilon(\xi)\hat{NL}2].$$

• For the modified system (3.10) the nonlinear step is explicit just like for (3.5) or (3.1).

All these schemes are of order 1 in time.

3.3. Numerical results. We have performed simulations with $\varepsilon = 10^{-2}$ or $\varepsilon = 10^{-3}$. All the results are given in the case where the numerical solution has converged, that is, a division of the time step by 2 and a multiplication by 2 of the number of points for the spatial discretization do not change the result. We use L^2-norms in order to compare the solutions. We take an initial value for u in the form

$$u(t=0,x) = \varepsilon^\alpha \left(e^{i\frac{kx}{\varepsilon}} \varphi(x) + c.c. \right)$$

for $\alpha \geq 0$. All the simulations are done with $k = 2\pi$ and $\omega = \sqrt{1 + (2\pi)^2}$. The initial value for v is obtained by using (3.9). That means that one takes $\psi(x) = \frac{1}{\mu^\varepsilon(D_x)}\varphi(x)$ and

$$v(t=0,x) = \varepsilon^\alpha \left(e^{i\frac{kx}{\varepsilon}} \psi(x) + c.c. \right).$$

The initial data for u_0, u_1, u_2, and u_3 is of course $\varphi(x)$. We call

$$e_{geo} = \max_{t\in[0,T]} \frac{|u(t,\cdot) - \varepsilon^\alpha (u_0(t,\cdot)e^{i\frac{kx-\omega t}{\varepsilon}} + c.c.)|_2}{|u(t,\cdot)|_2},$$

that is the maximum of the error between the exact solution of (3.1) and the approximate solution given by the geometrical optics approximation (3.3) on the time interval $[0,T]$. Here $|f|_2$ denotes the L^2-norm on $[0,1]$ of the function f. We also introduce

$$e_{diff} = \max_{t\in[0,T]} \frac{|u(t,\cdot) - \varepsilon^\alpha (u_1(t,\cdot)e^{i\frac{kx-\omega t}{\varepsilon}} + c.c.)|_2}{|u(t,\cdot)|_2},$$

that is the maximum of the error between the exact solution of (3.1) and the approximate solution given by the diffractive optics approximation (3.5) on the time interval $[0,T]$,

$$e_{new} = \max_{t\in[0,T]} \frac{|u(t,\cdot) - \varepsilon^\alpha (u_2(t,\cdot)e^{i\frac{kx-\omega t}{\varepsilon}} + c.c.)|_2}{|u(t,\cdot)|_2},$$

that is the maximum of the error between the exact solution of (3.1) and the approximate solution given by the new model on the time interval $[0,T]$, and

$$e_{newkerr} = \max_{t\in[0,T]} \frac{|u(t,\cdot) - \varepsilon^\alpha (u_3(t,\cdot)e^{i\frac{kx-\omega t}{\varepsilon}} + c.c.)|_2}{|u(t,\cdot)|_2},$$

that is the maximum of th error between the exact solution of (3.1) and the approximate solution given by the new model with Kerr nonlinearity given by (3.10) on the time interval $[0,T]$. We denote by N the number of Fourier modes in space and N_t the number of time steps.

3.3.1. Time of order 1.

Case 1. We begin with $\varphi(x) = e^{-75(x-\frac{1}{2})^2} e^{i10\cos(x)}$ with $\alpha = 0$ and we compute for $x \in [0,1]$ and $t \in [0,1]$. The errors at $T = 1$ are as follows.

	$\varepsilon = 10^{-2}$	$\varepsilon = 10^{-3}$
e_{geo}	2×10^{-2}	2.3×10^{-3}
e_{diff}	2×10^{-2}	2.3×10^{-3}
e_{new}	1.9×10^{-2}	2×10^{-3}
$e_{newkerr}$	2×10^{-2}	2.3×10^{-3}

For $\varepsilon = 10^{-2}$, the convergence on the errors is reached with $N = 1024$ and $N_t = 1600$. For $\varepsilon = 10^{-3}$ the convergence is reached with $N = 16384$ and $N_t = 12800$. For all cases, the error is of order ε as predicted by the theory. The simplest model (geometrical optics) is precise enough.

Case 2. We made a test for smaller solutions, namelly $\alpha = \frac{1}{2}$. The error at $T = 1$ are as follows.

	$\varepsilon = 10^{-2}$	$\varepsilon = 10^{-3}$
e_{geo}	3.3×10^{-3}	3.2×10^{-4}
e_{diff}	1.7×10^{-4}	1.8×10^{-6}
e_{new}	1.6×10^{-4}	1.9×10^{-6}
e_{newker}	1.9×10^{-4}	2×10^{-6}

For $\varepsilon = 10^{-2}$, the convergence on the errors is reached with $N = 1024$ and $N_t = 1600$. For $\varepsilon = 10^{-3}$, the convergence is reached with $N = 16384$ and $N_t = 12800$. Basically, the error for geometrical optics is the worst (of order ε), however, it remains very satisfactory. The others are of order ε^2 as predicted by the theory.

Case 3. For chirped initial data,

$$u(t=0,x) = \left(e^{-75(x-1/2)^2} e^{i15\cos(15x)} e^{i\frac{kx}{\varepsilon}} + c.c.\right),$$

$x \in [0,1]$. Such kind of solution can occur after diffraction webs for example or for laser with large spectrum. The errors at $T = 1$ are as follows.

	$\varepsilon = 10^{-2}$	$\varepsilon = 10^{-3}$
e_{geo}	0.8	5.7×10^{-2}
e_{diff}	0.17	1.2×10^{-2}
e_{new}	0.023	1.9×10^{-3}
$e_{newkerr}$	0.21	1.4×10^{-2}

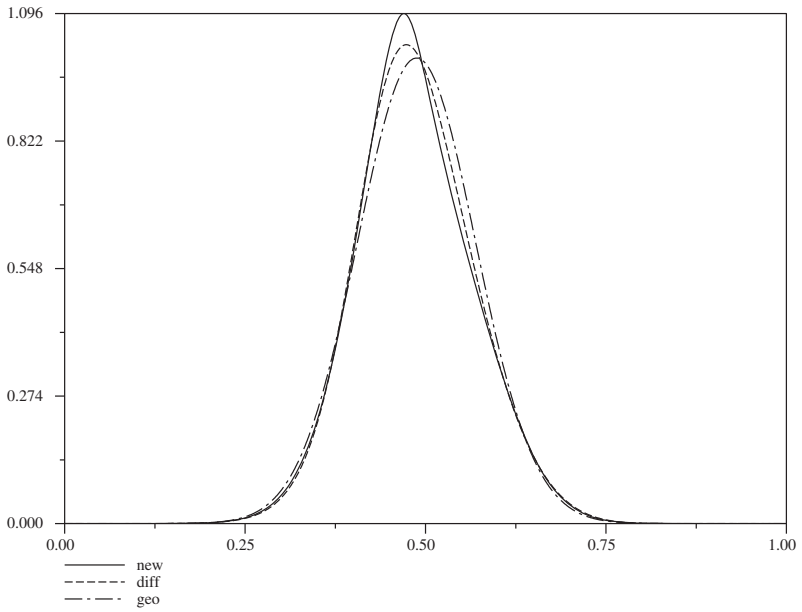

FIG. 3. *Amplitude of the first component of the approximate solution of* (3.1) *at the final time with chirped initial data,* $\alpha = 0$*, given by the geometrical optics, diffractive optics, and new approximations, Case* 3.

For $\varepsilon = 10^{-2}$, the convergence on the errors is reached with $N = 1024$ and $N_t = 1600$. For $\varepsilon = 10^{-3}$, the convergence is reached with $N = 16384$ and $N_t = 12800$. For $\varepsilon = 10^{-2}$, the error for the complete new model is 2.3%, the other errors are above 15%. Such errors are not acceptable in practical applications. As an illustration, one can find on Figure 3 the modulus of the amplitude (that is, without the phase factor $e^{i\frac{(kx-\omega t)}{\varepsilon}}$) of the first component for the three models: the new model, the geometrical optics, and the diffractive optics at the final time. As seen on the figure, the amplitude as well as the positions are false for the diffractive and geometrical optics regimes.

For $\varepsilon = 10^{-3}$, the result given by Shrödinger equation and the new model with the Kerr nonlinearity are correct. The geometrical optics give the worst error and the complete new model the smallest one.

Case 4. For smaller solutions, we made the same test but with $\alpha = \frac{1}{2}$. The errors are as follows.

	$\varepsilon = 10^{-2}$	$\varepsilon = 10^{-3}$
e_{geo}	0.91	6.9×10^{-2}
e_{diff}	0.32	2.3×10^{-3}
e_{new}	1.7×10^{-4}	1.7×10^{-6}
$e_{newkerr}$	2.1×10^{-3}	1.5×10^{-5}

For $\varepsilon = 10^{-2}$, the convergence on the errors is reached with $N = 1024$ and $N_t = 1600$. For $\varepsilon = 10^{-3}$, the convergence is reached with $N = 16384$ and $N_t = 12800$. As in the previous case, geometrical optics and the Schrödinger models give high errors for

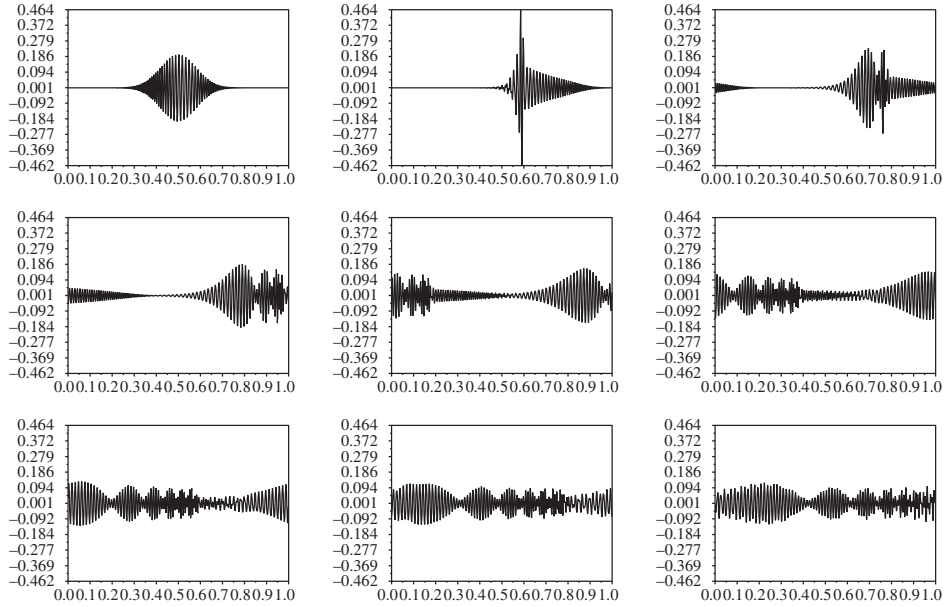

FIG. 4. *Real part of the first component of the solution of system* (3.1) *with* $\varepsilon = 0.01$ *at time* $t = n\frac{50}{8}$ *for* $n = 0, \ldots, 8$ *with chirped initial data and* $\alpha = 1/2$. *First line, from left to right,* $n = 0, 1, 2$, *second line, from left to right,* $n = 3, 4, 5$, *third line, from left to right,* $n = 6, 7, 8$, *Case* 6.

$\varepsilon = 0.01$. Both new models are correct however. For $\varepsilon = 10^{-3}$, the conclusions are the same as in the previous cases.

3.3.2. Diffractive time. We now consider long-time behavior: $T = 50$.

Case 5. We begin by a regular initial data and we take $\varphi(x) = e^{-75(x-\frac{1}{2})^2} e^{i\cos(x)}$ and $\alpha = \frac{1}{2}$. One gets the following errors.

	$\varepsilon = 10^{-2}$	$\varepsilon = 10^{-3}$
e_{geo}	0.13	1.3×10^{-2}
e_{diff}	2.4×10^{-3}	2×10^{-5}
e_{new}	1.7×10^{-4}	3×10^{-6}
$e_{newkerr}$	5.6×10^{-3}	5×10^{-5}

For $\varepsilon = 10^{-2}$, the convergence on the errors is reached with $N = 2048$ and $N_t = 80000$. For $\varepsilon = 10^{-3}$, the convergence is reached with $N = 8192$ and $N_t = 320000$. The geometrical optics gives of course a false result since diffractive effects are important. The result given by the new models are better than that of diffractive optics that is, however, perfectly correct. Any of the three models can be used in practical applications.

Case 6. For chirped initial data, we take

$$\varphi(x) = \left(e^{-75(x-1/2)^2} e^{i15\cos(15x)} + c.c. \right),$$

$x \in [0, 1]$ and $\alpha = \frac{1}{2}$ for $T = 50$. One gets the following errors.

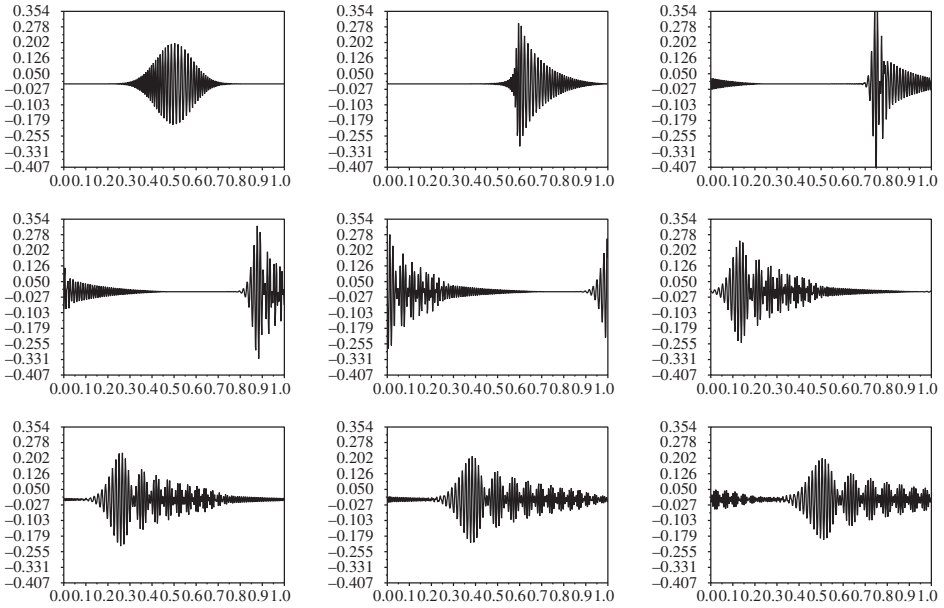

FIG. 5. *Real part of the approximate solution of system* (3.1) *with* $\varepsilon = 0.01$ *at time* $t = n\frac{50}{8}$ *for* $n = 0, \ldots, 8$ *with chirped initial data and* $\alpha = 1/2$ *given by diffractive optics approximation. First line, from left to right,* $n = 0, 1, 2$*, second line, from left to right,* $n = 3, 4, 5$*, third line, from left to right,* $n = 6, 7, 8$*, Case* 6.

	$\varepsilon = 10^{-2}$	$\varepsilon = 10^{-3}$
e_{geo}	1.5	1.6
e_{diff}	1.4	0.11
e_{new}	5×10^{-4}	3×10^{-6}
$e_{newkerr}$	0.08	8×10^{-4}

For $\varepsilon = 10^{-2}$, the convergence on the errors is reached with $N = 2048$ and $N_t = 80000$. For $\varepsilon = 10^{-3}$, the convergence is reached with $N = 8192$ and $N_t = 320000$. Only the complete new model gives an acceptable error. All the others give bad result. The new model with Kerr nonlinearity gives a satisfactory result for small ε but not for $\varepsilon = 0.01$. One can see the evolution of the solution at time $n\frac{50}{8}$ on Figure 4, and on Figure 5 the same but with the solution given by the Schrödinger equation which is far away from the reality.

3.3.3. Conclusion. For small times, chirped initial data or not, the diffractive model is satisfactory. For diffractive times and not chirped initial data, the diffractive model is satisfactory. The geometrical optics regime (that is the explicit solution) is valid on short times.

For diffractive times with chirped initial data, the new model is very useful. The new model with Kerr nonlinearity is intermediate in terms of quality. In any case the solution given by the new system cannot be distinguished from the exact one and will be therefore very useful in practical applications. We postpone the application of this theory to physical cases with more numerical tests to further work.

The main problem of our theory is the boundary conditions. Clearly, because of the pseudodifferential nature of the new model, it is not easy to take into account

nonperiodic boundary conditions. One of the possibility in this direction is to take one space variable as a variable of evolution. This process is under investigation.

REFERENCES

[1] D. Alterman and J. Rauch, *Nonlinear geometric optics for short pulses*, J. Differential Equations, 178 (2002), pp. 437–465.

[2] K. Barrailh and D. Lannes, *A general framework for diffractive optics and its applications to lasers with large spectrums and short pulses*, SIAM J. Math. Anal., 34 (2002), pp. 636–674.

[3] J.-D. Benamou, O. Lafitte, R. Sentis, and I. Solliec, *A Geometric Optics Based Numerical Method for High Frequency Electromagnetic Fields Computations Near Fold Caustics—Part* I, preprint, INRIA Tech. rep. RR-4422, INRIA, Le Chesnay, France 2002.

[4] C. Besse and B. Bidegaray, *Numerical study of self-focusing solutions to the Schrödinger-Debye system*, M2AN Math. Model. Numer. Anal., 1 (2001), pp. 35–55.

[5] B. Bidégaray, A. Bourgeade, D. Reignier, and R. Ziolkowski, *Multi-level Maxwell-Bloch simulations*, in Mathematical and Numerical Aspects of Wave Propagation, A. Bermúdez, D. Gómez, C. Hazard, P. Joly, and J. E. Roberts, eds., SIAM, Philadelphia, INRIA, Rocquercourt, FR 2000, pp. 221–225.

[6] J. Bona, M. Chen, and J.-C. Saut, *Boussinesq equations and other systems for small-amplitude long waves in nonlinear dispersive media.* I. *Derivation and linear theory*, J. Nonlinear Sci., 12 (2002), pp. 283–318.

[7] J. Bona, T. Colin, and D. Lannes, *Long wave models for water-waves*, preprint U-03-22, Université Bordeaux 1, Bordeaux, France, 2003, Arch. Ration. Mech. Anal., to appear.

[8] R. W. Boyd, *Nonlinear Optics*, Academic Press, Boston, 1992.

[9] T. Colin, *Rigorous derivation of the nonlinear Schrödinger equation and Davey-Stewartson systems from quadratic hyperbolic systems*, Asymptot. Anal., 31 (2002), pp. 69–91.

[10] T. Colin and B. Nkonga, *A numerical model for light interaction with a two-level atoms medium*, preprint LRC-03-17, Université Bordeaux 1, Bordeaux, France, 2002, Discrete Contin. Dyn. Syst. Ser. B, to appear.

[11] P. Donnat, *Quelques contributions mathématiques en optique non linéaire*, thèse École Polytechnique, Palaiseau, France, 1994.

[12] P. Donnat, J.-L. Joly, G. Métivier, and J. Rauch, *Diffractive nonlinear geometric optics*, Séminaire Sur Les Équations aux Dérivées Partielles 1995–1996, Exp. No. XVII, École Polytechnique, Palaiseau, France, 1996.

[13] P. Donnat and J. Rauch, *Dispersive nonlinear geometric optics*, J. Math. Phys., 38 (1997), pp. 1484–1523.

[14] J.-L. Joly, G. Métivier, and J. Rauch, *Generic rigorous asymptotic expansions for weakly nonlinear geometric optics*, Duke Math. J., 70 (1993), pp. 373–404.

[15] J.-L. Joly, G. Métivier, and J. Rauch, *Transparent nonlinear geometric optics and Maxwell-Bloch equations*, J. Differential Equations, 166 (2000), pp. 175–250.

[16] D. Lannes, *Dispersive effects for nonlinear geometrical optics with rectification*, Asymptot. Anal., 18 (1998), pp. 111–146.

[17] O. Morice, *Reference Manual of Code Miró*, CEA, France, 2002.

[18] A. C. Newell and J. V. Moloney, *Nonlinear Optics*, Addison–Wesley Reading, MA, 1991.

[19] T. Schäfer and E. Wayne, *Propagation of ultra-short optical pulses in nonlinear media*, preprint, Boston University, Boston, MA.

SIAM J. MATH. ANAL.
Vol. 36, No. 5, pp. 1689–1703

INVARIANT MEASURES FOR THE STOCHASTIC VON KARMAN PLATE EQUATION*

JONG UHN KIM†

Abstract. We prove the existence of an invariant measure for the von Karman plate equation with random noise. The nonlinear term which symbolizes the von Karman equation inhibits the standard procedure for the existence of an invariant measure. We propose a technically different approach to handle such intricate nonlinear equations.

Key words. von Karman plate equation, Brownian motion, stopping time, existence of a solution, invariant measure, probability distribution, tightness

AMS subject classifications. 35L65, 35R60, 60H15

DOI. 10.1137/S0036141003438854

1. Introduction.

In this paper, we will establish the existence of an invariant measure for a certain class of stochastic evolution equations with application to the stochastic von Karman plate equation. An invariant measure is an important object in stochastic dynamics. If the initial condition has the probability distribution equal to an invariant measure, then the probability distribution of the evolving solution is invariant in time. Some general results on the existence of invariant measures for stochastic evolution equations are presented in [6] and [7]. The basic method for the existence of invariant measures is due to Krylov and Bogolyubov [12]. However, there are some important equations which are not covered by the known theorems. Here we still follow the Krylov–Bogolyubov method, but with technically different adaptation, which has been motivated by the von Karman equation. For our method, we assume that the stochastic process associated with solutions has the Markov property with mean energy bounded uniformly in time, and that the probability distribution of the process is locally continuous with respect to a weaker norm. Typically, the first assumption is satisfied by a wide class of stochastic evolution equations with suitable dissipation. However, we need an additional condition for tightness of a family of probability measures which will yield an invariant measure. For parabolic equations, the regularizing property is crucially used to obtain tightness of a sequence of probability measures whose weak limit is an invariant measure; see [2]. Hyperbolic equations do not possess the regularizing property. But if the noise term has additional regularity in space variables and if more regular initial data can generate more regular solutions with a higher-order norm bounded uniformly in time, tightness of probability measures can be obtained in the same manner. There are equations which belong to neither case. The von Karman plate equation is a typical example. The advantage of this proposed approach lies in the second assumption, which is fairly mild and can be satisfied by equations such as the von Karman equation. We will highlight the utility of this procedure through the specific example of the von Karman equation.

The initial-boundary value problem for the von Karman plate is formulated as

*Received by the editors December 22, 2003; accepted for publication (in revised form) August 24, 2004; published electronically May 13, 2005.
http://www.siam.org/journals/sima/36-5/43885.html
†Department of Mathematics, Virginia Tech, Blacksburg, VA 24061-0123 (kim@math.vt.edu).

follows:

$$u_{tt} + \alpha u_t + \Delta^2 u - [u, v] = \sum_{j=1}^{\infty} g_j \frac{dB_j}{dt} \quad \text{in } (0,T) \times G, \tag{1.1}$$

$$\Delta^2 v + [u, u] = 0 \quad \text{in } (0,T) \times G, \tag{1.2}$$

$$u = \frac{\partial u}{\partial \nu} = 0, \quad v = \frac{\partial v}{\partial \nu} = 0 \quad \text{on } [0,T] \times \partial G, \tag{1.3}$$

$$u = u_0(x), \qquad u_t = u_1(x) \quad \text{at } t = 0. \tag{1.4}$$

Here G is a bounded domain in R^2 with smooth boundary ∂G, Δ is the Laplacian in R^2, $\frac{\partial}{\partial \nu}$ is the normal derivative on ∂G, and the bracket $[\cdot, \cdot]$ is defined by

$$[u, v] = \frac{\partial^2 u}{\partial x^2} \frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 v}{\partial x^2} \frac{\partial^2 u}{\partial y^2} - 2 \frac{\partial^2 u}{\partial x \partial y} \frac{\partial^2 v}{\partial x \partial y}. \tag{1.5}$$

Viscous damping is represented by a positive constant α, and B_j's are mutually independent standard Brownian motions over a given stochastic basis. When the right-hand side of (1.1) is replaced by a deterministic term, the existence of a weak solution to (1.1)–(1.4) was proved in [15], and more regular solutions were obtained in [4] and [8]. In fact, the weak solution belongs to the natural function class. Nevertheless, the uniqueness of the weak solution had been an open problem until the work of [1] and [8]. The existence and uniqueness of a solution to the stochastic problem (1.1)–(1.4) can be proved through a standard procedure based upon the known results from the deterministic case. The existence of statistical solutions was established in [3] and [10]. At present, the significant issue is the existence of an invariant measure.

Plate equations are neither hyperbolic nor parabolic while there is no regularizing property. In [4], it was shown that for large $\alpha > 0$ depending on the magnitudes of the given data, the bound of the global solution in a stronger norm is uniform in time. However, for small $\alpha > 0$, it is not known whether such an estimate is valid. Probably, it may not be true. This feature puts the above problem in a new category, which necessitates a technically different approach. Here we proceed in the opposite direction. Instead of trying to find uniform estimates in a stronger norm, we imbed the natural energy space into a larger function class, and obtain a probability measure on this larger space as a limit of a tight family of probability measures. We then prove that this is in fact an invariant measure on the original smaller space. For this, we need to show that the probability distribution of the solution depends continuously on initial data in a weaker norm for fixed time on each closed ball in the natural energy space. The main advantage of this procedure is that we do not need any additional estimates uniform in time other than uniform estimates in the natural energy space. Hence, we do not need either the assumption that $\alpha > 0$ is large or additional regularity of the noise term. We expect this procedure to be applied to other equations which behave like (1.1). Finally, the anonymous referee has informed the author that the idea of using a weaker topology was already used for interacting diffusions in [14] and for stochastic parabolic equations in [16] and [17].

2. Existence of invariant measures. Let $\{\Omega, \mathcal{F}_t, \mathcal{F}, P\}$ be a given stochastic basis and let $E(\cdot)$ denote the expectation with respect to P. Suppose that $X(t, s; z)$, $0 \le s \le t < \infty$ is a pathwise unique solution of a certain stochastic evolution equation such that $X(s, s; z) = z$. We assume

(I) $X(\cdot, s; z)$ is a Ξ-valued continuous process adapted to $\{\mathcal{F}_t\}_{t \ge s}$ for each $z \in \Xi$ and $s \ge 0$, where Ξ is a separable Banach space.

We define a function

$$\mathcal{P}(s,z;t,\Gamma) = P(X(t,s;z)\in\Gamma) \quad \text{for each } \Gamma\in\mathcal{B}(\Xi),\ 0\le s\le t<\infty,\ z\in\Xi, \tag{2.1}$$

where $\mathcal{B}(\Xi)$ is the Borel σ-algebra of Ξ. We assume

(II) $\mathcal{P}(\cdot,\cdot;\cdot,\cdot)$ is a time-homogeneous transition probability function. In other words, it satisfies the following conditions:

(i) $\mathcal{P}(s,z;t,\cdot)$ is a probability measure over $\{\Xi,\mathcal{B}(\Xi)\}$ for all $z\in\Xi$ and $0\le s<t<\infty$;

(ii) $\mathcal{P}(s,\cdot;t,\Gamma)$ is $\mathcal{B}(\Xi)$-measurable for all $0\le s<t<\infty$ and $\Gamma\in\mathcal{B}(\Xi)$;

(iii) for all $0\le s<t<\xi<\infty$ and $\Gamma\in\mathcal{B}(\Xi)$,

$$\mathcal{P}(s,z;\xi,\Gamma) = \int_\Xi \mathcal{P}(s,z;t,dy)\mathcal{P}(t,y;\xi,\Gamma);$$

(iv) $\mathcal{P}(s,\cdot;t,\cdot) = \mathcal{P}(s+h,\cdot;t+h,\cdot)$ for all $0\le s<t<\infty$ and $h>0$.

(III) There is some $z\in\Xi$ such that

$$E(\|X(t,0;z)\|_\Xi)\le M \quad \text{for all } t\ge 0 \tag{2.2}$$

for some positive constant M.

(IV) There is a Banach space Υ such that $\Xi\subset\Upsilon$, the imbedding $\Xi\to\Upsilon$ is continuous, and each closed ball of finite radius in Ξ is a compact subset of Υ. Furthermore, for each bounded continuous function ψ on Ξ, there is a sequence of continuous functions $\{\psi_k\}_{k=1}^\infty$ on Υ such that ψ_k is bounded uniformly in k and

$$\lim_{k\to\infty}\psi_k(y) = \psi(y) \quad \text{for each } y\in\Xi. \tag{2.3}$$

(V) For each fixed $0\le t<\infty$, and each fixed closed ball S of finite radius in Ξ, if $\{z_n\}_{n=1}^\infty$ is a sequence in S such that

$$z_n\to z \quad \text{in } \Upsilon, \tag{2.4}$$

then

$$E(\phi(X(t,0;z_n)))\to E(\phi(X(t,0;z))) \tag{2.5}$$

for every bounded continuous function ϕ on Υ.

Remark. If Ξ has a Schauder basis, the second part of assumption (IV) is automatically satisfied by using the continuous projection onto finite-dimensional subspaces. In fact, this is the case when we consider application to the von Karman plate equation.

THEOREM 2.1. *Under the assumptions* (I)–(V), *there is an invariant measure for the above process* $X(\cdot)$. *In other words, there is a probability measure* μ *on* Ξ *such that*

$$\int_\Xi E(\psi(X(t,0;y)))\mu(dy) = \int_\Xi \psi(y)\mu(dy) \tag{2.6}$$

for all $t\ge 0$ *and every bounded continuous function* ψ *on* Ξ.

Proof. Choose $z\in\Xi$ in the above assumption (III), and define a probability measure μ_T for each $T>0$ by

$$\mu_T(\Gamma) = \frac{1}{T}\int_0^T P(X(t,0;z)\in\Gamma)dt \tag{2.7}$$

for each $\Gamma \in \mathcal{B}(\Xi)$. This is well defined because $P(X(\cdot,0;z) \in \Gamma)$ is $\mathcal{B}([0,\infty))$-measurable. For this measurability, we argue as follows. For each bounded continuous function ϕ on Ξ, $E(\phi(X(t,0;z)))$ is continuous in t by assumption (I). Let Γ be a closed subset of Ξ and $\chi_\Gamma(\cdot)$ be the characteristic function of Γ. Then, there is a sequence of nonnegative bounded continuous functions $\{\phi_k\}_{k=1}^\infty$ on Ξ such that $\phi_k(y) \downarrow \chi_\Gamma(y)$ as $k \to \infty$ for each $y \in \Xi$. Hence, $E(\phi_k(X(t,0;z)))$ converges to $E(\chi_\Gamma(X(t,0;z)))$ as $k \to \infty$ for each t. Hence, $P(X(\cdot,0;z) \in \Gamma)$ is $\mathcal{B}([0,\infty))$-measurable. Let $\mathcal{S}$ be the collection of all subsets Γ such that $P(X(\cdot,0;z) \in \Gamma)$ is $\mathcal{B}([0,\infty))$-measurable. Then, $\mathcal{S}$ is a Dynkin system which includes all closed subsets of Ξ. Thus, $\mathcal{S}$ contains $\mathcal{B}(\Xi)$.

We now proceed to define

$$\tilde{\mu}_T(\Gamma) = \mu_T(\Gamma \cap \Xi) \tag{2.8}$$

for each $\Gamma \in \mathcal{B}(\Upsilon)$. Since the imbedding $\Xi \to \Upsilon$ is continuous, $\Gamma \cap \Xi$ is a Borel subset of Ξ for each $\Gamma \in \mathcal{B}(\Upsilon)$. Hence, $\tilde{\mu}_T$ is well defined and is a probability measure over $\{\Upsilon, \mathcal{B}(\Upsilon)\}$. For any $\epsilon > 0$, there is a positive number r_ϵ such that

$$P(\|X(t,0;z)\|_\Xi \le r_\epsilon) > 1 - \epsilon \quad \text{for all } t \ge 0 \tag{2.9}$$

which follows from assumption (III). Since the ball

$$S_{r_\epsilon} = \{y \in \Xi \mid \|y\|_\Xi \le r_\epsilon\} \tag{2.10}$$

is a compact subset of Υ by assumption (IV), the family of probability measures $\{\tilde{\mu}_T\}_{T>0}$ is tight. Hence, there is a sequence $\{\tilde{\mu}_{T_k}\}_{k=1}^\infty$ and a probability measure $\tilde{\mu}$ over $\{\Upsilon, \mathcal{B}(\Upsilon)\}$ such that $T_k \uparrow \infty$ as $k \to \infty$, and

$$\int_\Upsilon \phi(y)\tilde{\mu}_{T_k}(dy) \to \int_\Upsilon \phi(y)\tilde{\mu}(dy) \quad \text{as } k \to \infty \tag{2.11}$$

for every bounded continuous function ϕ on Υ. Since S_{r_ϵ} is a closed subset of Υ, it follows from (2.9) that

$$1 - \epsilon \le \limsup_{k\to\infty} \tilde{\mu}_{T_k}(S_{r_\epsilon}) \le \tilde{\mu}(S_{r_\epsilon}). \tag{2.12}$$

Since $\epsilon > 0$ is arbitrary and each Borel subset of Ξ is also a Borel subset of Υ, $\tilde{\mu}(\Xi) = 1$ and the restriction of $\tilde{\mu}$ to $\mathcal{B}(\Xi)$, written as μ, is a probability measure over $\{\Xi, \mathcal{B}(\Xi)\}$. Choose any bounded continuous function ϕ on Υ, and fix any $\epsilon > 0$. Then, there is $r > 0$ such that

$$\tilde{\mu}_{T_k}(S_r) = \mu_{T_k}(S_r) > 1 - \epsilon \quad \text{for all } k \ge 1. \tag{2.13}$$

Fix $t > 0$, and let

$$f(y) = E(\phi(X(t,0;y))) = \int_\Xi \mathcal{P}(0,y;t,dw)\phi(w). \tag{2.14}$$

Then, by assumption (V), $f(y)$ is continuous on S_r with respect to the norm of Υ. Since S_r is a closed subset of Υ, we can extend f to $\tilde{f}$ on Υ with the same bound such that $f(y) = \tilde{f}(y)$ for every $y \in S_r$. This follows from the Tietze extension theorem.

It is easy to see that

$$\begin{aligned}\int_\Upsilon \tilde{f}(y)\tilde{\mu}_{T_k}(dy) = &\int_{\Upsilon\setminus S_r} \tilde{f}(y)\tilde{\mu}_{T_k}(dy) \\ &+ \int_{S_r} \tilde{\mu}_{T_k}(dy) \int_\Xi \mathcal{P}(0,y;t,dw)\phi(w)\end{aligned} \tag{2.15}$$

and, by (2.13),

$$\left|\int_{S_r}\tilde{\mu}_{T_k}(dy)\int_{\Xi}\mathcal{P}(0,y;t,dw)\phi(w)-\int_{\Xi}\tilde{\mu}_{T_k}(dy)\int_{\Xi}\mathcal{P}(0,y;t,dw)\phi(w)\right|<M\epsilon, \tag{2.16}$$

where M is a positive constant such that $|\phi(y)|\le M$, for all $y\in\Upsilon$. Here we note that ϕ is also a continuous function on Ξ with respect to the norm of Ξ. It follows from assumption (II) that

$$\begin{aligned}\int_{\Xi}\mu_{T_k}(dy)&\int_{\Xi}\mathcal{P}(0,y;t,dw)\phi(w)\\ &=\frac{1}{T_k}\int_0^{T_k}\left(\int_{\Xi}\mathcal{P}(0,z;s,dy)\int_{\Xi}\mathcal{P}(0,y;t,dw)\phi(w)\right)ds\\ &=\frac{1}{T_k}\int_0^{T_k}\left(\int_{\Xi}\mathcal{P}(0,z;s+t,dy)\phi(y)\right)ds\\ &=\frac{1}{T_k}\int_t^{T_k+t}\left(\int_{\Xi}\mathcal{P}(0,z;\eta,dy)\phi(y)\right)d\eta.\end{aligned} \tag{2.17}$$

But we have

$$\lim_{k\to\infty}\left|\frac{1}{T_k}\int_t^{T_k+t}\left(\int_{\Xi}\mathcal{P}(0,z;\eta,dy)\phi(y)\right)d\eta-\int_{\Xi}\mu_{T_k}(dy)\phi(y)\right|=0, \tag{2.18}$$

$$\int_{\Xi}\mu_{T_k}(dy)\phi(y)=\int_{\Upsilon}\tilde{\mu}_{T_k}(dy)\phi(y), \tag{2.19}$$

and

$$\lim_{k\to\infty}\int_{\Upsilon}\tilde{\mu}_{T_k}(dy)\phi(y)=\int_{\Upsilon}\tilde{\mu}(dy)\phi(y)=\int_{\Xi}\mu(dy)\phi(y). \tag{2.20}$$

In the meantime, it holds that

$$\lim_{k\to\infty}\int_{\Upsilon}\tilde{f}(y)\tilde{\mu}_{T_k}(dy)=\int_{\Upsilon}\tilde{f}(y)\tilde{\mu}(dy), \tag{2.21}$$

$$\left|\int_{\Upsilon}\tilde{f}(y)\tilde{\mu}(dy)-\int_{S_r}f(y)\mu(dy)\right|<M\epsilon, \tag{2.22}$$

$$\int_{S_r}f(y)\mu(dy)=\int_{S_r}\mu(dy)E(\phi(X(t,0;y))), \tag{2.23}$$

and

$$\left|\int_{S_r}\mu(dy)E(\phi(X(t,0;y)))-\int_{\Xi}\mu(dy)E(\phi(X(t,0;y)))\right|<M\epsilon. \tag{2.24}$$

Thus, it follows from (2.15)–(2.24) that

$$\left|\int_{\Xi}\mu(dy)E(\phi(X(t,0;y)))-\int_{\Xi}\mu(dy)\phi(y)\right|<4M\epsilon. \tag{2.25}$$

Since $\epsilon>0$ is arbitrary, we have

$$\int_{\Xi}\mu(dy)E(\phi(X(t,0;y)))=\int_{\Xi}\mu(dy)\phi(y) \tag{2.26}$$

for all bounded continuous function ϕ on Υ for each $t > 0$. Next choose any bounded continuous function ψ on Ξ, and let $\{\psi_k\}_{k=1}^{\infty}$ be the sequence in assumption (IV). Then, for each $k \geq 1$, we have

$$\int_{\Xi} \mu(dy) E(\psi_k(X(t,0;y))) = \int_{\Xi} \mu(dy)\psi_k(y). \tag{2.27}$$

By passing $k \to \infty$, we have

$$\int_{\Xi} \mu(dy) E(\psi(X(t,0;y))) = \int_{\Xi} \mu(dy)\psi(y). \tag{2.28}$$

This completes the proof. □

3. Application to the stochastic von Karman equation. In this section, we present technical preliminaries to apply Theorem 2.1 to (1.1)–(1.4) and formulate the results.

Let $\{\phi_k\}_{k=1}^{\infty}$ be a complete orthonormal basis for $L^2(G)$ where each ϕ_k is an eigenfunction of

$$\begin{cases} \Delta^2 \phi_k = \lambda_k \phi_k & \text{in } G, \\ \phi_k = \dfrac{\partial \phi_k}{\partial \nu} = 0 & \text{on } \partial G. \end{cases} \tag{3.1}$$

Throughout this paper, $\langle \cdot, \cdot \rangle$ stands for the inner product of $L^2(G)$. It is easy to see that

$$\langle \Delta^2 \phi_j, \phi_k \rangle = \langle \Delta \phi_j, \Delta \phi_k \rangle = \lambda_j \delta_{jk} \quad \text{for all } j, k \geq 1. \tag{3.2}$$

$W^{m,p}(G)$, $H^m(G)$, and $H_0^m(G)$ denote the usual Sobolev spaces. Some of them can be characterized in terms of $\{\phi_k\}_{k=1}^{\infty}$:

$$H_0^2(G) \cap H^4(G) = \left\{ f = \sum_{k=1}^{\infty} a_k \phi_k \mid \sum_{k=1}^{\infty} \lambda_k^2 |a_k|^2 < \infty \right\}, \tag{3.3}$$

$$H_0^s(G) = \left\{ f = \sum_{k=1}^{\infty} a_k \phi_k \mid \sum_{k=1}^{\infty} \lambda_k^{s/2} |a_k|^2 < \infty \right\}, \quad 0 \leq s \leq 2, \ s \neq \frac{1}{2}, \frac{3}{2}, \tag{3.4}$$

$$H^{-s}(G) = \left\{ f = \sum_{k=1}^{\infty} a_k \phi_k \mid \sum_{k=1}^{\infty} \frac{1}{\lambda_k^{s/2}} |a_k|^2 < \infty \right\}, \quad 0 \leq s \leq 2, \ s \neq \frac{1}{2}, \frac{3}{2}. \tag{3.5}$$

We define the operator $\mathcal{G}$ on $H^{-2}(G)$ by

$$\mathcal{G}h = \sum_{k=1}^{\infty} \frac{1}{\lambda_k} a_k \phi_k \tag{3.6}$$

for $h = \sum_{k=1}^{\infty} a_k \phi_k \in H^{-2}(G)$. Obviously, $\mathcal{G}$ is the inverse of Δ^2 with the clamped boundary conditions. It is easy to see that for all $f, g \in L^2(G)$,

$$|\langle f, \mathcal{G} g \rangle| \leq \|f\|_{H^{-2}(G)} \|g\|_{H^{-2}(G)} \tag{3.7}$$

and

$$\langle f, \mathcal{G} f \rangle = \|f\|_{H^{-2}(G)}^2. \tag{3.8}$$

The following estimate was proved in [5] and [9]:

$$\|\mathcal{G}[f,g]\|_{W^{2,\infty}(G)} \le C\|f\|_{H^2(G)}\|g\|_{H^2(G)} \tag{3.9}$$

for all $f, g \in H^2(G)$.

Throughout this paper, $\{B_j\}_{j=1}^{\infty}$ is a sequence of mutually independent standard Brownian motions over the stochastic basis $\{\Omega, \mathcal{F}, \mathcal{F}_t, P\}$, where P is a probability measure over the σ-algebra $\mathcal{F}$, $\{\mathcal{F}_t\}$ is a right-continuous filtration over $\mathcal{F}$, and $\mathcal{F}_0$ contains all P-negligible sets. $E(\cdot)$ denotes the expectation with respect to P. When $\mathcal{X}$ is a Banach space, $\mathcal{B}(\mathcal{X})$ denotes the set of all Borel subsets of $\mathcal{X}$. An $\mathcal{X}$-valued function h is said to be $\mathcal{F}$-measurable if $h^{-1}(\mathcal{O}) \in \mathcal{F}$ for all $\mathcal{O} \in \mathcal{B}(\mathcal{X})$. This coincides with strong measurability for Bochner integrals when the range of h is separable. For $1 \le p < \infty$, $L^p(\Omega; \mathcal{X})$ denotes the set of all functions h which are $\mathcal{X}$-valued and strongly measurable with respect to $\mathcal{F}$ such that

$$\int_{\Omega} \|h\|_{\mathcal{X}}^p \, dP < \infty.$$

For general information on stochastic processes, see [11].

We assume the following condition on the noise term in (1.1). Each g_j depends only on the space variables, and

$$\sum_{j=1}^{\infty} \|g_j\|_{L^2(G)}^2 < \infty. \tag{3.10}$$

Under this assumption, we have the following existence result.

THEOREM 3.1. *For each $T > 0$ and $(u_0, u_1) \in H_0^2(G) \times L^2(G)$, there is a unique solution u of (1.1)–(1.4) such that (u, u_t) is adapted to $\{\mathcal{F}_t\}$, and*

$$(u, u_t) \in L^2(\Omega; C([0,T]; H_0^2(G) \times L^2(G))). \tag{3.11}$$

Here u satisfies (1.1) in the following sense. For almost all $\omega \in \Omega$, it holds that

$$\begin{aligned} &\langle u_t(t_2), \psi\rangle - \langle u_t(t_1), \psi\rangle + \int_{t_1}^{t_2} \langle \Delta u, \Delta\psi\rangle \, dt \\ &\quad + \alpha \int_{t_1}^{t_2} \langle u_t, \psi\rangle \, dt + \int_{t_1}^{t_2} \langle [u, \mathcal{G}[u,u]], \psi\rangle \, dt = \sum_{j=1}^{\infty} \int_{t_1}^{t_2} \langle g_j, \psi\rangle \, dB_j \end{aligned} \tag{3.12}$$

for all $\psi \in H_0^2(G)$ and all $0 \le t_1 < t_2 \le T$.

THEOREM 3.2. *There is an invariant measure on $H_0^2(G) \times L^2(G)$ for (1.1)–(1.4).*

4. Proof of Theorems 3.1 and 3.2. Let us define $\chi_N \in C_0^{\infty}(R)$ for each $N \ge 1$ by

$$\chi_N(y) = \begin{cases} 1 & \text{for } |y| \le 2N, \\ 0 & \text{for } |y| \ge 3N. \end{cases} \tag{4.1}$$

Then, it follows from (3.9) that

$$\begin{aligned} &\|\chi_N(\|u\|_{H^2(G)})[u, \mathcal{G}[u,u]] - \chi_N(\|w\|_{H^2(G)})[w, \mathcal{G}[w,w]]\|_{L^2(G)} \\ &\quad \le C_N \|u - w\|_{H^2(G)} \end{aligned} \tag{4.2}$$

for all $u, w \in H^2(G)$ and for some positive constant C_N. We now fix $N \geq 1$ and consider the modified problem

$$u_{tt} + \alpha u_t + \Delta^2 u + \chi_N(\|u\|_{H_0^2(G)})[u, \mathcal{G}[u,u]] = \sum_{j=1}^{\infty} g_j \frac{dB_j}{dt} \quad \text{in } (0,T)\times G, \tag{4.3}$$

$$u = \frac{\partial u}{\partial \nu} = 0 \quad \text{on } [0,T]\times \partial G, \tag{4.4}$$

$$u = u_0(x), \qquad u_t = u_1(x) \quad \text{at } t = 0. \tag{4.5}$$

By the general existence theorem in [6], for each $T > 0$ and $(u_0, u_1) \in H_0^2(G)\times L^2(G)$, there is a pathwise unique solution u of (4.3)–(4.5) such that (u, u_t) is adapted to $\{\mathcal{F}_t\}$, and

$$(u, u_t) \in L^2(\Omega; C([0,T]; H_0^2(G)\times L^2(G))). \tag{4.6}$$

This is still true when (u_0, u_1) is $\mathcal{F}_0$-measurable and $(u_0, u_1) \in L^2(\Omega; H_0^2(G)\times L^2(G))$, which follows from Kotelenez [13].

We introduce the projection P_m of $L^2(G)$ onto the subspace that is spanned by $\{\phi_1, \ldots, \phi_m\}$, and set

$$u_m = P_m u. \tag{4.7}$$

By taking the nonlinear term as a given function, we use the argument in [6, pp. 121–123] to obtain the following representation formula. For almost all $\omega \in \Omega$,

$$\begin{aligned} &\langle u_t(t_2), \psi\rangle - \langle u_t(t_1), \psi\rangle + \int_{t_1}^{t_2} \langle \Delta u, \Delta\psi\rangle\, dt + \alpha \int_{t_1}^{t_2} \langle u_t, \psi\rangle\, dt \\ &\quad + \int_{t_1}^{t_2} \langle \chi_N(\|u\|_{H_0^2(G)})[u, \mathcal{G}[u,u]], \psi\rangle\, dt = \sum_{j=1}^{\infty} \int_{t_1}^{t_2} \langle g_j, \psi\rangle\, dB_j \end{aligned} \tag{4.8}$$

for all $\psi \in H_0^2(G)$ and all $0 \leq t_1 < t_2 \leq T$. Thus, it follows that

$$\begin{aligned} d(u_{mt}) = {} & (-\Delta^2 u_m - \alpha u_{mt} - \chi_N(\|u\|_{H_0^2(G)}) P_m[u, \mathcal{G}[u,u]])dt \\ & + \sum_{j=1}^{\infty} P_m g_j\, dB_j \quad \text{for each } m \geq 1. \end{aligned} \tag{4.9}$$

By Ito's rule, we have, for all $0 \leq t_1 < t_2 \leq T$ and $m \geq 1$,

$$\begin{aligned} &\|u_{mt}(t_2)\|_{L^2(G)}^2 + \|\Delta u_m(t_2)\|_{L^2(G)}^2 \\ &\quad = \|u_{mt}(t_1)\|_{L^2(G)}^2 + \|\Delta u_m(t_1)\|_{L^2(G)}^2 - 2\alpha \int_{t_1}^{t_2} \|u_{mt}\|_{L^2(G)}^2\, dt \\ &\qquad - 2\int_{t_1}^{t_2} \langle \chi_N(\|u\|_{H_0^2(G)}) P_m[u, \mathcal{G}[u,u]], u_{mt}\rangle\, dt \\ &\qquad + 2\sum_{j=1}^{\infty} \int_{t_1}^{t_2} \langle P_m g_j, u_{mt}\rangle\, dB_j + \sum_{j=1}^{\infty} \int_{t_1}^{t_2} \|P_m g_j\|_{L^2(G)}^2 dt. \end{aligned} \tag{4.10}$$

It follows from (3.9) that

$$\begin{aligned} &\|\chi_N(\|u\|_{H_0^2(G)}) P_m[u, \mathcal{G}[u,u]] - \chi_N(\|u\|_{H_0^2(G)})[u_m, \mathcal{G}[u_m, u_m]]\|_{L^2(G)} \\ &\quad \leq \|\chi_N(\|u\|_{H_0^2(G)})(P_m[u, \mathcal{G}[u,u]] - [u, \mathcal{G}[u,u]])\|_{L^2(G)} \\ &\qquad + C_N \|u - u_m\|_{H_0^2(G)} \end{aligned} \tag{4.11}$$

and

$$\int_{t_1}^{t_2} 2\langle [u_m, \mathcal{G}[u_m, u_m]], u_{mt}\rangle dt = \|\Delta v_m(t_2)\|^2_{L^2(G)} - \|\Delta v_m(t_1)\|^2_{L^2(G)}, \tag{4.12}$$

where $v_m = \mathcal{G}[u_m, u_m]$. We now define a stopping time

$$\tau_N = \inf\{t > 0 \mid \|u(t)\|_{H_0^2(G)} \geq N\}. \tag{4.13}$$

By combining these and passing $m \to \infty$, we arrive at

$$\begin{aligned} &\|u_t(t_2)\|^2_{L^2(G)} + \|\Delta u(t_2)\|^2_{L^2(G)} + \|\Delta v(t_2)\|^2_{L^2(G)} \\ &\quad = \|u_t(t_1)\|^2_{L^2(G)} + \|\Delta u(t_1)\|^2_{L^2(G)} + \|\Delta v(t_1)\|^2_{L^2(G)} \\ &\qquad - 2\alpha \int_{t_1}^{t_2} \|u_t\|^2_{L^2(G)} dt + 2\sum_{j=1}^{\infty} \int_{t_1}^{t_2} \langle g_j, u_t\rangle \, dB_j + \sum_{j=1}^{\infty} \int_{t_1}^{t_2} \|g_j\|^2_{L^2(G)} dt \end{aligned} \tag{4.14}$$

for all $0 \leq t_1 \leq t_2 \leq \tau_N$ and for almost all $\omega \in \Omega$, where $v = \mathcal{G}[u, u]$. In the same way, we can also derive

$$\begin{aligned} &\langle u_t(t_2), u(t_2)\rangle + \frac{\alpha}{2}\|u(t_2)\|^2_{L^2(G)} \\ &\quad = \langle u_t(t_1), u(t_1)\rangle + \frac{\alpha}{2}\|u(t_1)\|^2_{L^2(G)} \\ &\qquad - \int_{t_1}^{t_2} (\|\Delta u\|^2_{L^2(G)} + \|\Delta v\|^2_{L^2(G)} - \|u_t\|^2_{L^2(G)}) dt + \sum_{j=1}^{\infty} \int_{t_1}^{t_2} \langle g_j, u\rangle dB_j \end{aligned} \tag{4.15}$$

for all $0 \leq t_1 \leq t_2 \leq \tau_N$ and for almost all $\omega \in \Omega$. We now write $u_N = u$ to signify the dependence of u on $\chi_N(\cdot)$. It follows from the Burkholder–Davis–Gundy inequality that

$$\begin{aligned} &E\left(\sup_{0\leq t\leq \tau_N \wedge T} \left| \sum_{j=1}^{\infty} \int_0^t \langle g_j, u_{Nt}\rangle \, dB_j \right| \right) \\ &\quad \leq ME\left(\sum_{j=1}^{\infty} \int_0^{\tau_N \wedge T} \|g_j\|^2_{L^2(G)} \|u_{Nt}\|^2_{L^2(G)} \, dt \right)^{1/2} \\ &\quad \leq \delta E\left(\sup_{0\leq t\leq \tau_N \wedge T} \|u_{Nt}\|^2_{L^2(G)} \right) + \frac{M^2 T}{4\delta} \sum_{j=1}^{\infty} \|g_j\|^2_{L^2(G)} \end{aligned} \tag{4.16}$$

for all $\delta > 0$ and for some positive constant M independent of N and T. Thus, we can derive from (4.14)

$$\begin{aligned} &E\left(\sup_{0\leq t\leq \tau_N \wedge T} (\|u_{Nt}(t)\|^2_{L^2(G)} + \|\Delta u_N(t)\|^2_{L^2(G)} + \|\Delta v_N(t)\|^2_{L^2(G)}) \right) \\ &\quad \leq C(\|u_0\|^2_{H_0^2(G)} + \|u_1\|^2_{L^2(G)} + \|\mathcal{G}[u_0, u_0]\|^2_{H_0^2(G)}) + CT\sum_{j=1}^{\infty} \|g_j\|^2_{L^2(G)} \end{aligned} \tag{4.17}$$

for some constant C independent of N and $T > 0$. Thus, we find that

$$P(\tau_N \leq T) \leq C_T/N^2 \to 0 \quad \text{as } N \to \infty. \tag{4.18}$$

Since $\tau_{N_1} \le \tau_{N_2}$ for $N_1 < N_2$, it follows that

$$\lim_{N\to\infty} \tau_N = \tau_\infty \ge T \quad \text{for almost all } \omega \in \Omega. \tag{4.19}$$

Since T is arbitrary,

$$\tau_N \uparrow \infty \quad \text{as } N \to \infty \tag{4.20}$$

for almost all $\omega \in \Omega$. By the pathwise uniqueness of u_N, we have $u_{N_1} = u_{N_2}$ on $[0, \tau_{N_1} \wedge \tau_{N_2}]$ for almost all $\omega \in \Omega$, and we can define

$$u(t) = u_N(t) \quad \text{for } t \in [0, \tau_N]. \tag{4.21}$$

Then, this u is the desired solution. Now (4.8) implies (3.12). Since each (u_N, u_{Nt}) is adapted to $\{\mathcal{F}_t\}$, (u, u_t) is adapted to $\{\mathcal{F}_t\}$. By Fatou's lemma, we derive from (4.17) and (4.20) that

$$\begin{aligned} &E\bigg(\sup_{0\le t\le T}(\|u_t(t)\|^2_{L^2(G)} + \|\Delta u(t)\|^2_{L^2(G)} + \|\Delta v(t)\|^2_{L^2(G)})\bigg) \\ &\qquad \le C(\|u_0\|^2_{H_0^2(G)} + \|u_1\|^2_{L^2(G)} + \|\mathcal{G}[u_0,u_0]\|^2_{H_0^2(G)}) + CT\sum_{j=1}^{\infty}\|g_j\|^2_{L^2(G)} \end{aligned} \tag{4.22}$$

for all $T > 0$, where $v = \mathcal{G}[u, u]$.

For the proof of pathwise uniqueness, we suppose that $(\tilde{u}, \tilde{u}_t)$ is another solution of (1.1)–(1.4) in $L^2(\Omega; C([0,T]; H_0^2(G) \times L^2(G)))$. Then, $u - \tilde{u}$ satisfies

$$u_{tt} - \tilde{u}_{tt} + \alpha(u_t - \tilde{u}_t) + \Delta^2(u - \tilde{u}) + [u, \mathcal{G}[u,u]] - [\tilde{u}, \mathcal{G}[\tilde{u}, \tilde{u}]] = 0 \tag{4.23}$$

for almost all $\omega \in \Omega$. Since (u, u_t) and $(\tilde{u}, \tilde{u}_t)$ belong to $C([0,T]; H_0^2(G) \times L^2(G))$ for almost all ω, we can apply the same argument as for the deterministic case to conclude that $u \equiv \tilde{u}$ for almost all $\omega \in \Omega$. This completes the proof of Theorem 3.1.

Next we will obtain uniform estimates. Fix any λ such that

$$0 < \lambda < \min(1, \alpha, \lambda_1), \tag{4.24}$$

where λ_1 is the first eigenvalue of (3.1), and define

$$\begin{aligned} Q(t) = {} & \|u_t(t)\|^2_{L^2(G)} + \|\Delta u(t)\|^2_{L^2(G)} + \|\Delta v(t)\|^2_{L^2(G)} \\ & + \lambda\langle u_t(t), u(t)\rangle + \frac{\alpha}{2}\lambda\|u(t)\|^2_{L^2(G)}. \end{aligned} \tag{4.25}$$

By virtue of (4.20) and (4.21), u satisfies (4.14) and (4.15) for all $0 \le t_1 < t_2 < \infty$ and for almost all ω. Since integrability is guaranteed by (4.22), it follows from (4.14) and (4.15) that

$$\begin{aligned} E(Q(t_2)) - E(Q(t_1)) = {} & -\lambda\int_{t_1}^{t_2} E(\|\Delta u\|^2_{L^2(G)} + \|\Delta v\|^2_{L^2(G)})dt \\ & - (2\alpha - \lambda)\int_{t_1}^{t_2} E(\|u_t\|^2_{L^2(G)})\,dt + \int_{t_1}^{t_2}\sum_{j=1}^{\infty}\|g_j\|^2_{L^2(G)}dt \end{aligned} \tag{4.26}$$

for all $0 \le t_1 < t_2 < \infty$. We can derive

$$\frac{d}{dt}E(Q(t)) \le -cE(Q(t)) + \sum_{j=1}^{\infty} \|g_j\|^2_{L^2(G)} \tag{4.27}$$

for all $t > 0$, where c is a positive constant depending on α, λ_1, and λ. This yields

$$E(Q(t)) \le C_M \quad \text{for all } t \ge 0, \tag{4.28}$$

where M is a constant such that $Q(0) \le M$, and C_M is a constant depending on M and the last term of (4.27). By virtue of (4.24), this yields (2.2).

According to the above argument for the existence of solutions, we could take any $s \ge 0$ as the initial time and $\zeta = (\zeta_0, \zeta_1)$ as the initial value for the Cauchy problem (1.1)–(1.3) if ζ is $H_0^2(G) \times L^2(G)$-valued $\mathcal{F}_s$-measurable such that $\zeta \in L^2(\Omega; H_0^2(G) \times L^2(G))$, and $\mathcal{G}[\zeta_0, \zeta_0] \in L^2(\Omega; H_0^2(G))$. We now write $X(t, s; \zeta) = (u, u_t)$, where u is the solution of (1.1)–(1.3) for $t \ge s$ satisfying $(u(s), u_t(s)) = \zeta$. Then, $X(\cdot, s; \zeta) \in L^2(\Omega; C([s, T]; H_0^2(G) \times L^2(G)))$ for all $T > s$, and (4.28) holds for all $t \ge s$. For each $0 \le s < t$, $z \in H_0^2(G) \times L^2(G)$, and $\Gamma \in \mathcal{B}(H_0^2(G) \times L^2(G))$, we set as in (2.1)

$$\mathcal{P}(s, z; t, \Gamma) = P(X(t, s; z) \in \Gamma).$$

LEMMA 4.1. *Choose any bounded continuous function ψ on $H_0^2(G) \times L^2(G)$, $t > s \ge 0$. Then,*

$$E(\psi(X(t, s; z))) = \int_{H_0^2(G) \times L^2(G)} \mathcal{P}(s, z; t, dy)\psi(y) \tag{4.29}$$

is continuous in $z \in H_0^2(G) \times L^2(G)$.

Proof. Let $\{z_n\}_{n=1}^{\infty}$ be a sequence in $H_0^2(G) \times L^2(G)$ such that $z_n \to z$ in $H_0^2(G) \times L^2(G)$. Let us fix any $t > s \ge 0$. By (4.22), we have

$$\begin{cases} E\left(\sup\limits_{s \le \eta \le t} \|X(\eta, s; z)\|_{H_0^2(G) \times L^2(G)} \right) \le M, \\ E\left(\sup\limits_{s \le \eta \le t} \|X(\eta, s; z_n)\|_{H_0^2(G) \times L^2(G)} \right) \le M \quad \text{for all } n \ge 1 \end{cases} \tag{4.30}$$

for some positive constant M. Let us fix any $\epsilon > 0$ and any bounded continuous function ψ on $H_0^2(G) \times L^2(G)$. Since $H_0^2(G) \times L^2(G)$ is a Polish space and $P(X(t, s; z) \in \cdot)$ is a probability measure over $\{H_0^2(G) \times L^2(G), \mathcal{B}(H_0^2(G) \times L^2(G))\}$, there is a compact subset $\mathcal{K}$ of $H_0^2(G) \times L^2(G)$ such that

$$P(X(t, s; z) \in \mathcal{K}) > 1 - \epsilon. \tag{4.31}$$

By virtue of (4.30), there is some $R > 0$ such that

$$\begin{cases} P\left(\sup\limits_{s \le \eta \le t} \|X(\eta, s; z_n)\|_{H_0^2(G) \times L^2(G)} \le R \right) > 1 - \epsilon \quad \text{for all } n \ge 1, \\ P\left(\sup\limits_{s \le \eta \le t} \|X(\eta, s; z)\|_{H_0^2(G) \times L^2(G)} \le R \right) > 1 - \epsilon. \end{cases} \tag{4.32}$$

By taking R larger, we also have

$$\mathcal{K} \subset \{y \in H_0^2(G) \times L^2(G) \mid \|y\|_{H_0^2(G) \times L^2(G)} \le R\}. \tag{4.33}$$

Let us fix such R and write for each n

$$A_n = \left\{ \sup_{s\le\eta\le t} \|X(\eta,s;z_n)\|_{H_0^2(G)\times L^2(G)} \le R \right\} \bigcap \left\{ \sup_{s\le\eta\le t} \|X(\eta,s;z)\|_{H_0^2(G)\times L^2(G)} \le R \right\} \bigcap \{X(t,s;z)\in\mathcal{K}\}. \tag{4.34}$$

We will estimate the integral on the right-hand side of

$$\begin{aligned} &|E(\psi(X(t,s;z))) - E(\psi(X(t,s;z_n)))| \\ &\quad \le \int_{A_n} |\psi(X(t,s;z)) - \psi(X(t,s;z_n))|\, dP + 6M\epsilon, \end{aligned} \tag{4.35}$$

where M is a positive constant such that $|\psi(y)| \le M$ for all y. By means of (3.9), we can derive from (4.23) that

$$\|X(t,s;z) - X(t,s;z_n)\|^2_{H_0^2(G)\times L^2(G)} \le C_R \|z_n - z\|^2_{H_0^2(G)\times L^2(G)} \tag{4.36}$$

for all $\omega \in \tilde{A}_n$, where $\tilde{A}_n \subset A_n$ and $P(A_n \setminus \tilde{A}_n) = 0$, and C_R is a constant independent of n. Since ψ is continuous on $H_0^2(G)\times L^2(G)$ and $\mathcal{K}$ is compact, there is $\delta > 0$ such that

$$|\psi(x) - \psi(y)| < \epsilon \tag{4.37}$$

for every $x \in \mathcal{K}$, $y \in H_0^2(G)\times L^2(G)$ satisfying $\|x-y\|_{H_0^2(G)\times L^2(G)} < \delta$. Hence, it follows from (4.36) that there is $N \ge 1$ such that for all $n \ge N$,

$$\int_{A_n} |\psi(X(t,s;z)) - \psi(X(t,s;z_n))|\, dP < \epsilon, \tag{4.38}$$

which yields

$$|E(\psi(X(t,s;z))) - E(\psi(X(t,s;z_n)))| < \epsilon + 6M\epsilon \tag{4.39}$$

for all $n \ge N$. Thus, $E(\psi(X(t,s;z)))$ is continuous in z. □

This implies that $\mathcal{P}(s,\cdot;t,\Gamma)$ is $\mathcal{B}(\Xi)$-measurable for all $0 \le s < t < \infty$ and $\Gamma \in \mathcal{B}(\Xi)$. This can be seen by the same argument as in the previous proof of the measurability of $P(X(\cdot,0;z))$.

LEMMA 4.2. *$X(\cdot)$ has the Markov property, and its transition probability function is time-homogeneous.*

Proof. By the uniqueness of solution, it holds that for any $0 \le r < s < t$ and $z \in H_0^2(G)\times L^2(G)$,

$$X(t,r;z) = X(t,s;X(s,r;z)) \tag{4.40}$$

for almost all ω. We have to show that

$$E(\psi(X(t,s;X(s,r;z))) \mid \mathcal{F}_s) = \mathcal{P}_{s,t}(\psi)(X(s,r;z)) \tag{4.41}$$

for almost all ω, for each bounded continuous function ψ on $H_0^2(G)\times L^2(G)$, where

$$\mathcal{P}_{s,t}\psi(y) = E(\psi(X(t,s;y))).$$

According to the proof of Theorem 3.1, the solution was obtained by the truncation method. Let $X_N = X_N(t,s;\zeta)$ denote the solution $(u_N, \partial_t u_N)$ of (4.3)–(4.4) satisfying $(u_N(s), \partial_t u_N(s)) = \zeta$, where $\zeta = (\zeta_0, \zeta_1)$ is $H_0^2(G)\times L^2(G)$-valued $\mathcal{F}_s$-measurable such that $\zeta \in L^2(\Omega; H_0^2(G)\times L^2(G))$ and $\mathcal{G}[\zeta_0,\zeta_0] \in L^2(\Omega; H_0^2(G))$. Then, we know that for each $T > s$,

$$X(t,s;\zeta) = \lim_{N\to\infty} X_N(t,s;\zeta) \quad \text{in } C([s,T]; H_0^2(G)\times L^2(G)) \tag{4.42}$$

for almost all ω. For each $N \geq 1$ and each bounded continuous function ψ on $H_0^2(G)\times L^2(G)$, it holds that

$$E(\psi(X_N(t,s;\zeta)) \mid \mathcal{F}_s) = \mathcal{P}_{s,t}^N(\psi)(\zeta) \tag{4.43}$$

for almost all ω, which follows directly from the argument in [6, p. 250]. Here $\mathcal{P}_{s,t}^N$ is defined by

$$\mathcal{P}_{s,t}^N\psi(y) = E(\psi(X_N(t,s;y))).$$

Since ψ is a bounded continuous function, we pass $N\to\infty$ to arrive at

$$E(\psi(X(t,s;\zeta)) \mid \mathcal{F}_s) = P_{s,t}(\psi)(\zeta) \tag{4.44}$$

for almost all ω. Hence $X(\cdot)$ has the Markov property.

Since g_j's are independent of time, we can apply the result in [6, p. 251] to see that the transition probability function is time-homogeneous. ∎

LEMMA 4.3. *Let $S_L = \{y \in H_0^2(G)\times L^2(G) \mid \|y\|_{H_0^2(G)\times L^2(G)} \leq L\}$, and let $\{z_n\}_{n=1}^\infty$ be a sequence in S_L such that $z_n \to z$ in $H_0^1(G)\times H^{-1}(G)$. If ϕ is a bounded continuous function on $H_0^1(G)\times H^{-1}(G)$, then for each $t > 0$,*

$$E(\phi(X(t,0;z_n))) \to E(\phi(X(t,0;z))) \tag{4.45}$$

as $z_n \to z$ in $H_0^1(G)\times H^{-1}(G)$.

Proof. Let us fix any $t^* > 0$, and write

$$Y_n(t) = X(t,0;z_n) - X(t,0;z). \tag{4.46}$$

Suppose that

$$\|X(t,0;z_n)\|_{H_0^2(G)\times L^2(G)} \leq R, \qquad \|X(t,0;z)\|_{H_0^2(G)\times L^2(G)} \leq R \tag{4.47}$$

for all $0 \leq t \leq t^*$ for some constant R. It follows from (4.23) and the basic inequality established in [1] that

$$\begin{aligned} &\|Y_n(t_2)\|^2_{H_0^1(G)\times H^{-1}(G)} - \|Y_n(t_1)\|^2_{H_0^1(G)\times H^{-1}(G)} \\ &\qquad \leq C_1 \log(1+\lambda_N) \int_{t_1}^{t_2} \|Y_n(s)\|^2_{H_0^1(G)\times H^{-1}(G)}\, ds + C_2 t^* \lambda_{N+1}^{-\beta} \end{aligned} \tag{4.48}$$

for all $0 \leq t_1 < t_2 \leq t^*$, all $N \geq N_0$, for some constant $0 < \beta < 1$, and for positive integer N_0. Here λ_N is the Nth eigenvalue of (3.1), and C_1 and C_2 are positive constants depending only on β and R. We partition $[0,t^*]$ such that

$$0 = t_0 < t_1 < \cdots < t_K = t^*, \quad t_k - t_{k-1} = t^*/K < \beta/C_1, \quad 1 \leq k \leq K. \tag{4.49}$$

By the Gronwall inequality, we can derive from (4.48) that

$$\|Y_n(t)\|^2_{H_0^1(G)\times H^{-1}(G)} \le (\|Y_n(t_k)\|^2_{H_0^1(G)\times H^{-1}(G)} + C_2 t^* \lambda_{N+1}^{-\beta})(1+\lambda_N)^{C_1(t-t_k)} \tag{4.50}$$

for all $t \in [t_k, t_{k+1}]$, all $N \ge N_0$, and for each $k = 0, \dots, K-1$. Since $\lambda_N \uparrow \infty$ as $N \to \infty$, we use (4.49) to infer from (4.50) that for given $\epsilon > 0$, there is $\epsilon_K > 0$ such that if $\|Y_n(t_{K-1})\|_{H_0^1(G)\times H^{-1}(G)} < \epsilon_K$,

$$\|Y_n(t_K)\|_{H_0^1(G)\times H^{-1}(G)} < \epsilon. \tag{4.51}$$

By iteration, we find that there is $\epsilon_1 > 0$ such that if $\|z_n - z\|_{H_0^1(G)\times H^{-1}(G)} < \epsilon_1$, (4.51) holds. By the same argument as in the proof of Lemma 4.1, we arrive at (4.45). □

LEMMA 4.4. *Let ψ be a bounded continuous function on $H_0^2(G) \times L^2(G)$. Then, there is a sequence $\{\psi_k\}_{k=1}^{\infty}$ such that each ψ_k is a continuous function on $H_0^1(G) \times H^{-1}(G)$ bounded uniformly in k, and*

$$\psi_k(y) \to \psi(y) \quad \text{as } k \to \infty \tag{4.52}$$

for each $y \in H_0^2(G) \times L^2(G)$.

Proof. It is enough to set

$$\psi_k(y) = \psi((P_k y_1, P_k y_2)) \quad \text{for } y = (y_1, y_2) \in H_0^1(G) \times H^{-1}(G),\ k = 1, 2, \dots, \tag{4.53}$$

where P_k is the projection onto the subspace spanned by $\{\phi_1, \dots, \phi_k\}$. □

Finally, we set

$$\Xi = H_0^2(G) \times L^2(G), \qquad \Upsilon = H_0^1(G) \times H^{-1}(G).$$

Then, assumptions (I)–(V) follow from the above lemmas, and the proof of Theorem 3.2 is complete.

REFERENCES

[1] A. BOUTET DE MONVEL AND I. CHUESHOV, *Uniqueness theorem for weak solutions of von Karman evolution equations*, J. Math. Anal. Appl., 221 (1998), pp. 419–429.

[2] P.L. CHOW AND R.Z. KHASMINSKII, *Stationary solutions of nonlinear stochastic evolution equations*, Stochastic Anal. Appl., 15 (1997), pp. 671–699.

[3] I. CHUESHOV, *Existence of statistical solutions of a stochastic system of von Karman equations in a bounded domain*, Sb. Math., 50 (1985), pp. 279–298.

[4] I. CHUESHOV, *Strong solutions and the attractor of the von Karman equations*, Sb. Math., 69 (1991), pp. 25–36.

[5] I. CHUESHOV AND I. LASIECKA, *Inertial manifolds for von Karman plate equations*, Appl. Math. Optim., 46 (2002), pp. 179–206.

[6] G. DA PRATO AND J. ZABCZYK, *Stochastic Equations in Infinite Dimensions*, Cambridge University Press, Cambridge, 1992.

[7] G. DA PRATO AND J. ZABCZYK, *Ergodicity for Infinite Dimensional Systems*, Cambridge University Press, Cambridge, 1996.

[8] A. FAVINI, M. HORN, I. LASIECKA, AND D. TATARU, *Global existence, uniqueness and regularity of solutions to a von Karman system with nonlinear boundary dissipation*, Differential Integral Equations, 9 (1996), pp. 267–294.

[9] A. FAVINI, M. HORN, I. LASIECKA, AND D. TATARU, *Addendum to the paper: Global existence, uniqueness and regularity of solutions to a von Karman system with nonlinear boundary dissipation*, Differential Integral Equations, 10 (1997), pp. 197–200.

[10] V.I. Gishlarkaev, *The existence of statistical solutions of the stochastic von Karman system in a bounded domain*, Math. Notes, 58 (1995), pp. 692–702.
[11] I. Karatzas and S. Shreve, *Brownian Motion and Stochastic Calculus*, 2nd ed., Springer, New York, Berlin, Heidelberg, 1997.
[12] N. Krylov and N. Bogolyubov, *La théorie générale de la mesure dans son application à l'étude des systémes de la Mécanique non linéaire*, Ann. of Math. (2), 38 (1937), pp. 65–113.
[13] P. Kotelenez, *A submartingale type inequality with applications to stochastic evolution equations*, Stochastics, 8 (1982), pp. 139–151.
[14] G. Leha and G. Ritter, *Lyapunov-type conditions for stationary distributions of diffusion processes on Hilbert spaces*, Stochastics Stochastics Rep., 48 (1994), pp. 195–225.
[15] J.L. Lions, *Quelques Méthodes de Résolution des Problémes aux Limites Non Linéaires*, Dunod, Paris, 1969.
[16] B. Maslowski and J. Seidler, *On sequentially weakly Feller solutions to SPDE's*, Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl., 10 (1999), pp. 69–78.
[17] B. Maslowski and J. Seidler, *Strong Feller solutions to SPDE's are strong Feller in the weak topology*, Studia Math., 148 (2001), pp. 111–129.

SIAM J. MATH. ANAL.
Vol. 36, No. 5, pp. 1704–1705

ERRATUM: ON THE HÖLDER CONTINUITY OF SOLUTIONS OF A CERTAIN SYSTEM RELATED TO MAXWELL'S EQUATIONS*

KYUNGKEUN KANG† AND SEICK KIM‡

Abstract. The main purpose of this erratum is to correct Lemmas 2.4 and 2.5 in [K. Kang and S. Kim, *SIAM J. Math. Anal.*, 34 (2002), pp. 87–100] and present their proofs. We also take this opportunity to rectify some flaws caused by those incorrectly stated lemmas.

AMS subject classifications. 35B45, 35J60, 35Q60

DOI. 10.1137/040612907

First we make a correction to the definition of $\mathcal{D}(\Omega)$ [2, p. 88, line 25] as follows:

$$\mathcal{D}(\Omega) = \mathcal{D}(\Omega;\mathbb{R}^3) = \left\{ f \in C^\infty(\Omega;\mathbb{R}^3) : f \text{ is compactly supported }, \nabla \cdot f = 0 \right\}.$$

In Theorems 2.1 and 2.2 and the other related parts of the article, $f \in \mathcal{H}^q_{\mathrm{loc}}(\Omega)$ should read $f \in \mathcal{H}^q(\Omega)$, and $\|f\|_{L^q(B)}$ should read $\|f\|_{L^q(\Omega)}$.

Then, Lemmas 2.4 and 2.5 should be corrected as follows.

LEMMA 2.4. *Let $\Omega \subset \mathbb{R}^3$ be an open set and assume $f \in \mathcal{D}(\Omega)$. Then there exists $g \in C^\infty(\Omega)$ such that $\nabla \times g = f$ and $\nabla \cdot g = 0$ in Ω. Moreover, we have $\|\nabla g\|_{L^p(\Omega)} \le C(p)\,\|f\|_{L^p(\Omega)}$ for $1 < p < \infty$.*

Proof. We define $g := -\nabla \times N(f)$, where $N(f)$ is the Newtonian potential of f over Ω (see, e.g., [1, p. 51] for the definition). Then from the vector identity

$$\nabla \times (\nabla \times F) = \nabla(\nabla \cdot F) - \Delta F, \tag{2.4}$$

we find

$$\nabla \times g = -\nabla \times (\nabla \times N(f)) = \Delta N(f) - \nabla(\nabla \cdot N(f)) = f - \nabla(N(\nabla \cdot f)) = f.$$

Also, by the Calderón–Zygmund theory (see, e.g., [1, Theorem 9.9]), we find $\|\nabla g\|_{L^p(\Omega)} \le C\,\|f\|_{L^p(\Omega)}$. Clearly, we have $\nabla \cdot g = 0$. ∎

LEMMA 2.5. *Suppose $F \in C^\infty(\overline{B};\mathbb{R}^3)$ satisfies $\nabla \times F = 0$ in B. Then there exists $\varphi \in C^\infty(B)$ such that $\nabla\varphi = F$ and $\int_B \varphi = 0$. Moreover, we have $\|\varphi\|_{L^2(B)} \le C\,\|F\|_{L^2(B)}$.*

Proof. Let φ be a solution to

$$\begin{cases} \Delta\varphi = \nabla \cdot F & \text{in} \quad B, \\ \frac{\partial\varphi}{\partial n} = F \cdot n & \text{on} \quad \partial B. \end{cases}$$

By subtracting a constant, we may assume $\int_B \varphi = 0$.

Denote $\omega := \nabla\varphi - F$. We have $\nabla \cdot \omega = 0$, $\nabla \times \omega = 0$ in B, and $\omega \cdot n = 0$ on ∂B. Therefore $\omega \equiv 0$ in B. We have thus shown that $\nabla\varphi = F$ in B. Since $\int_B \varphi = 0$, the Poincaré inequality implies $\|\varphi\|_{L^2(B)} \le C\,\|F\|_{L^2(B)}$. ∎

*Received by the editors August 5, 2004; accepted for publication (in revised form) August 24, 2004; published electronically May 13, 2005.

http://www.siam.org/journals/sima/36-5/61290.html

†Department of Mathematics, University of British Columbia, 121-1984 Mathematics Road, Vancouver V6T 1Z2, BC, Canada (kkang@math.ubc.ca).

‡Mathematics Department, University of Missouri, Columbia, MO 65211 (seick@math.missouri.edu).

Finally, we should make a slight change in the proof of Theorem 2.1.

On page 91, line 2, $f \in \mathcal{H}^q_{\mathrm{loc}}(\Omega) \cap C^\infty(\Omega)$ should read $f \in \mathcal{D}(\Omega)$.

On page 91, lines 7–10 should be replaced as follows:

Since $f \in \mathcal{D}(\Omega)$, we conclude from Lemma 2.4 that there exists a smooth vector g such that $f = \nabla \times g$ in Ω. By subtracting a constant vector, we may assume that $\int_{B_8} g = 0$. Then the Sobolev–Poincaré inequality implies

$$\|g\|_{L^{q^*}(B_8)} \le C \|\nabla g\|_{L^q(B_8)} \le C \|f\|_{L^q(\Omega)}, \quad q^* = nq/(n-q) > n. \tag{2.5}$$

Acknowledgment. We thank Professor Marius Mitrea for bringing these errors to our attention.

REFERENCES

[1] D. Gilbarg and N. Trudinger, *Elliptic Partial Differential Equations of Second Order,* 2nd ed., Springer-Verlag, Berlin, 1983.

[2] K. Kang and S. Kim, *On the Hölder continuity of solutions of a certain system related to Maxwell's equations*, SIAM J. Math. Anal., 34 (2002), pp. 87–100.

SIAM J. MATH. ANAL.
Vol. 36, No. 6, pp. 1707–1725

PERIODIC MOTIONS OF LINEAR IMPACT OSCILLATORS VIA THE SUCCESSOR MAP*

DINGBIAN QIAN† AND PEDRO J. TORRES‡

Abstract. We investigate the existence and multiplicity of nontrivial periodic bouncing solutions for linear and asymptotically linear impact oscillators by applying a generalized version of the Poincaré–Birkhoff theorem to an adequate Poincaré section called the successor map. The main theorem includes a generalization of a related result by Bonheure and Fabry and provides a sufficient condition for the existence of periodic bouncing solutions for Hill's equation with obstacle at $x \neq 0$.

Key words. impact oscillator, Hill's equation, periodic solution, successor map, Poincaré–Birkhoff twist theorem

AMS subject classifications. 34C15, 34C25, 34B30, 54H25

DOI. 10.1137/S003614100343771X

1. Introduction and main results.

In this paper, we study the existence of $2m\pi$-periodic bouncing solutions for the following linear impact oscillator:

$$\begin{cases} x'' + a(t)x = p(t) \quad \text{for } x(t) > 0; \\ x(t) \geq 0; \\ x(t_0) = 0 \ \Rightarrow x'(t_0+) = -x'(t_0-), \end{cases} \tag{1.1}$$

where $a(t),\ p(t)$ are 2π-periodic continuous functions and $p(t)$ satisfies

$$p(t) \leq 0 \quad \text{and} \quad \overline{p} = \frac{1}{2\pi}\int_0^{2\pi} p(t)dt < 0. \tag{1.2}$$

This system is included in a larger family of impact oscillators given by

$$\begin{cases} x'' + f(t,x,x') = 0 \quad \text{for } x(t) > q(t); \\ x(t) \geq q(t); \\ x(t_0) = q(t_0) \ \Rightarrow x'(t_0+) = -x'(t_0-) + 2q'(t_0), \end{cases} \tag{1.3}$$

where f is continuous and 2π-periodic with respect to t and $q \in C^2(\mathbb{R})$ is also 2π-periodic. From the viewpoint of mechanics this equation models the motion of a particle attached to a nonlinear spring and bouncing elastically against the barrier described by $q(t)$. Thus (1.3) is a model of dynamical system with discontinuity [23] that can be included in the wide family of *vibro-impact systems* [3]. Because of the range of applications in physics and engineering, vibro-impact systems have attracted the attention of a lot of researchers and in consequence the number of papers related to this topic is huge; see [4, 8, 10, 21, 22, 14] and their bibliographies only to mention

*Received by the editors November 14, 2003; accepted for publication (in revised form) August 17, 2004; published electronically May 20, 2005.
http://www.siam.org/journals/sima/36-6/43771.html

†Department of Mathematics, Suzhou University, Suzhou 215006, People's Republic of China (dbqian@suda.edu.cn). This author's research was supported by NNSF of China (10271085) and NSF of Jiangsu Province, China (BK2002037, 02KJB110003).

‡Universidad de Granada, Departamento de Matemática Aplicada, 18071 Granada, Spain (ptorres@ugr.es). This author's research was supported by D.G.I. BFM2002-01308, Ministerio de Ciencia y Tecnología, Spain.

some of them. There are also interesting relations with Fermi accelerator [15, 35], dual billiards [7], and celestial mechanics [9].

In spite of this, even for the simple case of a one-degree-of-freedom linear oscillator with impacts, the dynamics is far from being understood, although some results are known [6, 24, 25, 33]. Our purpose in this paper is to investigate the existence of nontrivial periodic bouncing solutions with prescribed number of impacts for linear and asymptotically linear impact oscillators. As it is known, the existence of subharmonics of arbitrary order is usually a hint of a complex dynamics. The following definition clarifies the concept of bouncing solution we mean here.

DEFINITION 1.1. *A continuous function $x : \mathbb{R} \to \mathbb{R}$ is a* bouncing solution *for problem* (1.3) *if the following conditions hold:*

1. $x(t) \geq q(t)$ *for all* $t \in \mathbb{R}$*;*
2. *the set* $W = \{t : x(t) = q(t)\}$ *is discrete and not empty;*
3. $x'(t_0+) = -x'(t_0-) + 2q'(t_0)$ *for any* $t_0 \in W$*;*
4. *given an interval* I*, if* $I \cap W = \emptyset$*, then* $x \in C^2(I, \mathbb{R}^+)$ *and it is a classical solution of* (1.3)*.*

Note that the change of variables $y(t) = x(t) - q(t)$ enables to assume without loss of generality that the barrier is fixed at zero. In this context, Lazer and McKenna [25] proved the existence of 2π-periodic bouncing solution for a linear oscillator with small amplitude forcing term and small viscous damping. Recently, Bonheure and Fabry [6] proved the existence of a 2π-periodic bouncing solution for the linear oscillator

$$x'' + \lambda x = p(t), \tag{1.4}$$

where $\lambda > 0$ is a constant and $p(t) < 0$. They also introduced the concept of *admissible* solution in [6] to treat the case where $p(t)$ changes its sign and showed some existence results for perturbations of a linear oscillator. The main feature of an admissible solution is that it can vanish on a whole interval. This is physically equivalent to an attachment of the particle to the barrier $x = 0$ during a whole interval of time. Due to the condition (1.2), we are able to work directly with the more specific concept of bouncing solution, which constitutes a particular class of admissible solutions.

Obviously, our model (1.1) includes (1.4) and also the bouncing problem for the Hill's equation

$$x'' + a(t)x = 0 \tag{1.5}$$

with obstacle $q(t) = d > 0$. Note that $x(t)$ is a bouncing solution of the problem

$$\begin{cases} x'' + a(t)x = 0 \quad \text{for } x(t) > d; \\ x(t) \geq d; \\ x(t_0) = d \ \Rightarrow x'(t_0+) = -x'(t_0-) \end{cases} \tag{1.6}$$

if and only if $y(t)$ is a solution of (1.1) with $p(t) = -a(t)d$ by means of the change $y(t) = x(t) - d$.

The approach of this paper is different from that in [25, 6]. We apply a new generalized version of Poincaré–Birkhoff twist theorem to the so-called successor map, defined as follows. For a given $\tau \in \mathbb{R}$ and $v \in \mathbb{R}^+$, let us denote by $x(t; \tau, v)$ the unique solution of the bouncing problem (1.1) with initial conditions $x(\tau; \tau, v) = 0$, $x'(\tau; \tau, v) = v > 0$. We assume conditions such that this solution is well defined and vanishes at some time $\hat{\tau} > \tau$. Thus $\hat{\tau}$ is the time of the next impact. As the bouncing is elastic, the velocity after this impact is

$$\hat{v} = -x'(\hat{\tau}; \tau, v).$$

If $\hat{v}$ is finite and positive, then the map

$$\mathcal{S} : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R} \times \mathbb{R}^+,$$
$$\mathcal{S}(\tau, v) = (\hat{\tau}, \hat{v})$$

is well defined, continuous, and one to one. Following [1, 31, 32, 33], this function is called *successor map*, although in this context "impact map" would be also adequate.

Let us state some notation to be used in the rest of the paper: given a 2π-periodic function $p(t)$, $\bar{p} = \frac{1}{2\pi}\int_0^{2\pi} p(t)\mathrm{d}t$ is the mean value of p and $\|p\|_\infty = \max_{0\le t\le 2\pi} |p(t)|$. The projection for the component i of a given vector is denoted by Π_i. All along the paper, the iteration of the successor map is denoted by $\mathcal{S}^n(\tau, v) = (\hat{\tau}^n(\tau, v),\ \hat{v}^n(\tau, v))$ and we will use $\hat{\tau}^n = \hat{\tau}^n(\tau, v),\ \hat{v}^n = \hat{v}^n(\tau, v)$ for short. Therefore, $\Pi_1(\mathcal{S}^n(\tau, v)) = \hat{\tau}^n,\ \Pi_2(\mathcal{S}^n(\tau, v)) = \hat{v}^n$. Both notations are used without distinction.

Our main result is the following.

THEOREM 1.2. *Assume that the successor map $\mathcal{S}$ is well defined for all $(\tau, v) \in \mathbb{R} \times \mathbb{R}^+$ and $p(t) \le 0$ for all t, $\bar{p} < 0$. Then for any $m, n \in \mathbb{N}$ such that $n > 2m(\sqrt{\|a\|_\infty})$, there exists at least one $2m\pi$-periodic bouncing solution of* (1.1) *with exactly n impacts in each period. Moreover, for any $m \in \mathbb{N}$ such that $2m(\sqrt{\|a\|_\infty}) < 1$, there exist at least two $2m\pi$-periodic solutions with one bouncing in each period.*

The following corollaries present two concrete situations where the successor map is well defined and the previous result applies.

COROLLARY 1.3. *Assume that $p(t) \le 0$ for all t, $\bar{p} < 0$, and $\bar{a} > 0$. Then, the conclusion of Theorem* 1.2 *holds.*

COROLLARY 1.4. *Assume that $p(t) \le 0$ for all t, $\bar{p} < 0$, and $a(t) \equiv 0$. Then for any $m, n \in \mathbb{N},\ n \ge 2$, there exists at least one $2m\pi$-periodic bouncing solution of* (1.1) *with exactly n impacts in each period. Moreover, for any $m \in \mathbb{N}$, there exist at least two $2m\pi$-periodic solutions with one bouncing in each period.*

Remark 1.5. In our opinion, the application of the Poincaré–Birkhoff twist theorem to the successor map instead of the Poincaré map (as it is done in [6]) is more natural and direct. For the linear impact oscillator (1.4) we can obtain at least two $2m\pi$-periodic bouncing solutions for (1.4) with exactly 1 impact in each period if $2m\sqrt{\lambda} < 1$, whereas in [6] only one solution is found. Moreover, we can deal with a nonconstant coefficient $a(t)$, in contrast with [6].

In order to understand some of the new phenomena arising in vibro-impact systems, it is interesting to consider in detail the Hill's equation with impacts (1.6) as a particular case. Note that if the obstacle is placed at $d = 0$, then a classical solution x of Hill's equation generates a bouncing solution $|x|$ of (1.6). Hence, in this case (1.6) inherits the dynamics of Hill's equation without impacts and in consequence its resonant or nonresonant character. However, if the obstacle is $d > 0$, the situation is different. Physically, this model corresponds to a kind of offset impact oscillator [18], consisting of a linear spring-mass system with a displaced wall with respect to the origin (see Figure 1(a)). The time-dependence of the stiffness coefficient $a(t)$ of the spring can be produced by periodic changes of the temperature or other physical variables. A periodic bouncing solution corresponds to a nontrivial periodic motion with prescribed impacts in one period. The following result holds.

COROLLARY 1.6. *Assume that $d > 0$, $a(t) \ge 0$ for all t, and $\bar{a} > 0$. Then for any $m, n \in \mathbb{N}$ such that $n > 2m(\sqrt{\|a\|_\infty})$, there exists at least one $2m\pi$-periodic bouncing solution of* (1.6) *with exactly n impacts in each period.*

The proof follows from Corollary 1.3 by means of the change of variables $y = x - d$. Thus, the Hill's equation could be unstable (equivalently, all nontrivial solutions are

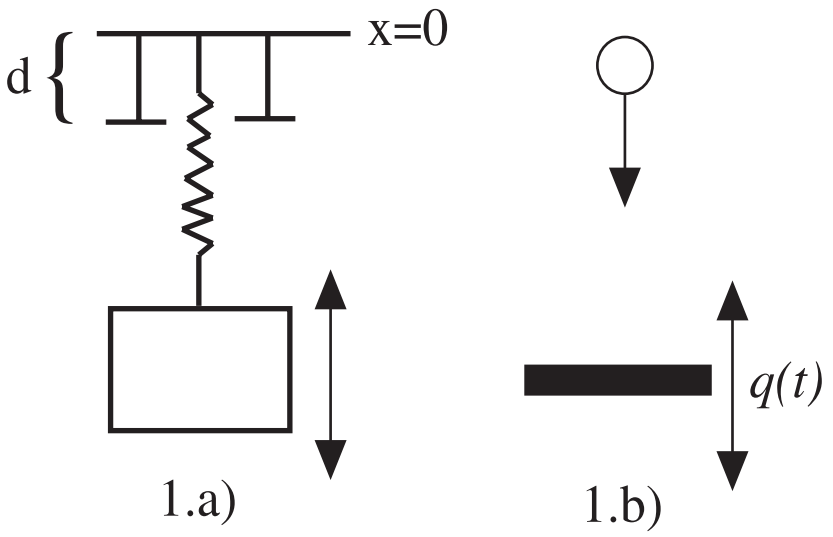

FIG. 1. (a) *The offset oscillator.* (b) *The "ping-pong" model.*

unbounded [26]) but nevertheless (1.6) has periodic bouncing solutions. In other words, possible parametric resonances are killed by the presence of an obstacle. This fact is a good example of the obstacle's influence in the dynamics of a given system.

Another simple but physically interesting model is the *"ping-pong" problem*, that is, a free ball moving in a vertical line subjected to gravity force and bouncing against a barrier or racket describing a periodic movement $q(t)$ (see Figure 1(b)). If G is the acceleration of gravity, the motion of the ball is described by

$$\begin{cases} x'' + G = 0 \quad \text{for } x(t) > q(t); \\ x(t) \geq q(t); \\ x(t_0) = q(t_0) \ \Rightarrow x'(t_0+) = -x'(t_0-) + 2q'(t_0). \end{cases}$$

This is a simple variation of Fermi's model that have deserved the attention of many researchers (see [19, 5, 13] and their references). After the change $y(t) = x(t) - q(t)$, the problem is transformed in (1.1) with $a(t) \equiv 0$ and $p(t) = -G - q''(t)$. Then, if $q''(t) > -G$ for any t, the ball experiences a diversity of periodic motions with a prescribed number of impacts as a consequence of Corollary 1.4.

Remark 1.7. The concept of bouncing solution could involve other new features and strong differences with the situation when working with differential equations without impacts. An interesting open problem is to prove or disprove the validity of Massera's theorem for impact oscillators. Massera's theorem asserts that in the framework of periodic differential equations the existence of a bounded solution implies the existence of a periodic solution [28]. This classical result is false in the context of equations with impacts in the sense that a bounded bouncing solution (using the definition in this paper) does not imply a periodic bouncing solution. To prove this, consider the Mathieu equation $a(t) = \gamma + \delta \cos t$ with obstacle $d = 0$ and parameters γ, δ placed in a stability region with irrational rotation number. Then any nontrivial solution of (1.5) is quasi-periodic (but not periodic) and in consequence every bouncing solution of (1.5) is bounded but there are no periodic bouncing solutions. Of course,

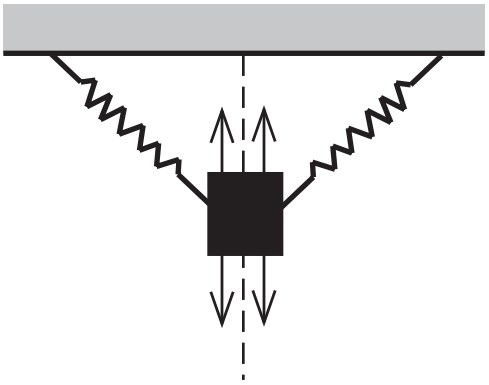

FIG. 2. *A spring-mass impact system.*

this is just an effect of the definition chosen here, since the trivial solution is excluded. Note that the trivial solution is not a bouncing solution but it is an admissible solution in the sense of [6]. So the exciting question of the validity of Massera's theorem for impact oscillators is still open: does the existence of a bounded bouncing solution imply the existence of a periodic admissible solution (including trivial solution)? We do not know the answer.

Our successor map approach is also suitable for use in nonlinear impact oscillators, as it is done in [34] for a singular equation. Here we include a result about the asymptotically linear impact oscillator.

THEOREM 1.8. *Let us assume that $g(t,x)$ is continuous, 2π-periodic with respect to t, and satisfies*

$$\limsup_{x\to 0^+}\left|\frac{g(t,x)}{x}\right| < +\infty, \qquad \lim_{x\to+\infty}\frac{g(t,x)}{x} = 0. \tag{1.7}$$

Besides, let us suppose that the successor map $\mathcal{S}$ of the bouncing problem

$$\begin{cases} x'' + a(t)x + g(t,x) = p(t) \quad \text{for } x(t) > 0; \\ x(t) \geq 0; \\ x(t_0) = 0 \ \Rightarrow x'(t_0+) = -x'(t_0-) \end{cases} \tag{1.8}$$

is well defined for all $(\tau, v) \in \mathbb{R}\times\mathbb{R}^+$ and $p(t) \leq 0$ for all t, $\bar{p} < 0$. Then, the conclusion of Theorem 1.2 *holds.*

A corollary of the previous result is the following.

COROLLARY 1.9. *Assume that $p(t) \leq 0$ for all t, $\bar{p} < 0$, $g(t,x)$ satisfies* (1.7) *and $a(t)x + g(t,x) \geq 0$ for any $x \geq 0$. Then, the conclusion of Theorem* 1.2 *holds.*

This result can be illustrated by a simple physical model presented in Figure 2. This mechanical system is a modification of the model presented in [2, 17] and consists of a single mass moving in a straight line, attached to the wall by two linear springs of constant k and natural length L and perturbed periodically by an external force $p(t)$. If it is assumed that the impacts between the mass and the wall are perfectly elastic, then the motion of the mass is governed by

$$\begin{cases} mx'' + 2k\left[x - \dfrac{Lx}{(c^2+x^2)^{1/2}}\right] = p(t) \quad \text{for } x(t) > 0; \\ x(t) \geq 0; \\ x(t_0) = 0 \ \Rightarrow x'(t_0+) = -x'(t_0-), \end{cases}$$

where $c > 0$ is the distance between the point of impact and the attachments of the springs (see [2, 17] for more details). If $p(t) \leq 0$ for all t, $\bar{p} < 0$, it is easy to verify that this problem is under the assumptions of Corollary 1.9 when $c > L$.

The rest of the paper is organized as follows. In section 2, the proof of Theorem 1.2 is given. It relies on a generalized version of Poincaré–Birkhoff theorem. Section 3 collects some auxiliary lemmas which are needed in the mentioned proof, more specifically the twist property of some iteration of the successor map is proved. Finally, section 4 is devoted to the study of the asymptotically linear impact oscillator.

2. Existence of periodic bouncing solutions. We will apply the Poincaré–Birkhoff twist theorem to the successor map $\mathcal{S}$ for proving the existence of 2π-periodic bouncing solutions for impact oscillators (1.1). The successor map was used recently by Ortega [33] for investigation of the boundedness of all the solutions for a linear impact oscillator by using Moser's twist theorem and the authors [34] for investigation of the periodic bouncing solutions for some singular impact oscillator. As a general idea, this successor map is just a different section of the flux and it goes back at least to Alekseev [1] and Moser [30].

The following generalized version of Poincaré–Birkhoff twist theorem is based on the theorems of Franks [16] and Ding [11] and is slightly different from the version used by others (see, for example, [20], [6], and [27]).

Let A and B be two annuli

$$A := \mathbf{S}^1 \times [a_1, a_2], \qquad B := \mathbf{S}^1 \times [b_1, b_2]$$

with $0 < b_1 < a_1 < a_2 < b_2 < +\infty$. A map $f : A \to B$ possesses a lift $\tilde{f} : \mathbb{R}\times[a_1, a_2] \to \mathbb{R} \times [b_1, b_2]$ with the form

$$\theta' = \theta + h(\theta, \rho), \quad \rho' = g(\theta, \rho),$$

where $h,\ g$ are continuous and 2π-periodic in θ. We say that $\tilde{f}$ satisfies the boundary twist condition if

$$h(\theta, a_1) \cdot h(\theta, a_2) < 0 \quad \text{for } \theta \in [0,\ 2\pi].$$

THEOREM 2.1. *Assume that $f : A \to B$ is an area-preserving homeomorphism homotopic to the inclusion such that $f(A) \cap \partial B = \emptyset$. Moreover, f possesses a lift $\tilde{f}$ satisfying the boundary twist condition and the area of the two connected components of the complement of $f(A)$ in B is the same as the area of the corresponding connected components of the complement of A in B. Then, f has at least two geometrically distinct fixed points (θ_i, ρ_i), $(i = 1, 2)$ satisfying $h(\theta_i, \rho_i) = 0$ for $i = 1, 2$.*

Proof. The proof basically combines the proofs from Franks [16] and Ding [11]. In [16], Franks showed that by using a result from Oxtoby and Ulam, one can extend f to an area-preserving homeomorphism $F : B \to B$ such that F is the identity on the boundary of B (see the proof and the remark of Theorem 4.2 in [16]). Then, we can assume further that F is an area-preserving homeomorphism of $D := \{(\theta, \rho) : \rho \leq b_2\}$ to its image such that $O \in F(D\backslash B)$. Now we meet all the assumptions of the argument in [11]. According to the argument of [11], we can prove that F, and then f, has at least two fixed points in A. Moreover, the fixed points (θ_i, ρ_i) satisfy $h(\theta_i, \rho_i) = 0$ for $i = 1, 2$ (see [11] and [12] for more details). Figure 3 illustrates the geometrical meaning of the hypotheses. □

Now, we apply the above Poincaré–Birkhoff theorem to the successor map $\mathcal{S}$. From the discussion in the next section we know that our successor map $\mathcal{S}$

$$\mathcal{S} : (\tau,\ v) \mapsto (\hat{\tau},\ \hat{v})$$

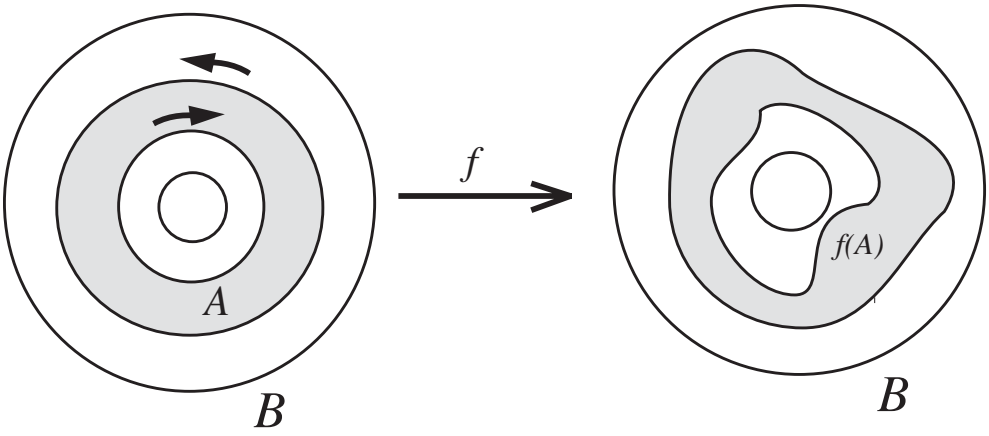

FIG. 3. *The Poincaré–Birkhoff theorem.*

is well defined, one to one, and continuous in its domain $\mathbb{R}\times\mathbb{R}^+$. Moreover, it satisfies

$$\mathcal{S}(\tau+2\pi,v)=\mathcal{S}(\tau,v)+(2\pi,0).$$

Thus, we can interpret τ and v as polar coordinates and $\mathcal{S}$ is an embedding homeomorphism on $\mathbf{S}^1\times\mathbb{R}^+$. It is easy to show that for any $n,m\in\mathbb{N}$, a fixed point of the map $\mathcal{S}^n(\tau,v)-(2m\pi,0)$ corresponds a $2m\pi$-periodic bouncing solution of the equation with n impacts in each period. We have the following lemma.

LEMMA 2.2. *$\mathcal{S}$ is an area-preserving map with the area element $vdvd\tau$. Moreover, $\mathcal{S}$ is area-preserving homotopic to the inclusion, and for any annuli $A\subset B\subset\mathbf{S}^1\times\mathbb{R}^+$ with $\mathcal{S}(A)\subset\overset{\circ}{B}$, the area of the two connected components of the complement of $\mathcal{S}(A)$ in B is the same as the area of the corresponding components of the complement of A in B.*

The proof of this lemma is similar to the proof of Lemma 1 in [20] and the proof of Proposition 2.3 in [31]. At first we can prove, under the assumption of the C^1-smoothness of a and p which implies the C^1-smoothness of $\mathcal{S}$, that $\mathcal{S}$ is an exact symplectic map in its domain; that is, for any C^1-closed path γ in its domain

$$\int_\gamma \frac{v^2}{2}d\tau=\int_{S\circ\gamma}\frac{v^2}{2}d\tau. \tag{2.1}$$

Moreover, note that $\mathcal{S}$ is an embedding homeomorphism on $\mathbf{S}^1\times\mathbb{R}^+$, then from Jordan separation theorem (see, for instance, [29]), we know that for any annuli $A\subset B\subset\mathbf{S}^1\times\mathbb{R}^+$ with $\mathcal{S}(A)\subset\overset{\circ}{B}$, there are two connected components of the complement of $\mathcal{S}(A)$ in B. Such components are the images of the two components of the complement of A in B. Hence, the geometric meaning of (2.1) is that the area of the components of the complement of $\mathcal{S}(A)$ in B are the same as the area of the corresponding components of the complement of A in B. The conclusion for the case of continuous functions a and p follows from an approximation argument.

Moreover, Lemmas 3.4 and 3.6 (see section 3) imply that, under the assumptions of Theorem 1.2, we can choose $v_-^{(n)}<v_+^{(n)}$ such that

$$\Pi_1(\mathcal{S}^n(\tau,v_-^{(n)}))-\tau<2m\pi,$$
$$\Pi_1(\mathcal{S}^n(\tau,v_+^{(n)}))-\tau>2m\pi\quad\text{for }\tau\in[0,2\pi].$$

Hence, let A be the annulus bounded by $\mathbf{S}^1 \times \{v_-^{(n)}\}$ and $\mathbf{S}^1 \times \{v_+^{(n)}\}$ and let B be the annulus bounded by $\mathbf{S}^1 \times \{v_*\}$ and $\mathbf{S}^1 \times \{v^*\}$. We can prove, as showed in section 3, that $f(A) \subset \overset{\circ}{B}$ for $v_* > 0$ sufficiently small and v^* sufficiently large, where $f : A \to B$ is defined by

$$f(\tau, v) = \mathcal{S}^n(\tau, v) - (2m\pi, 0).$$

It is easy to see that f is an area-preserving homeomorphism homotopic to the inclusion and $\tilde{f}$ satisfies the boundary twist condition. Thus the conclusion of Theorem 1.2 follows by a direct application of Theorem 2.1. Note that in any case we get two fixed points of $\mathcal{S}^n(\tau, v) - (2m\pi, 0)$. However, if the number of bouncings n is greater than or equal to 2, these two fixed points provided by Theorem 2.1 may correspond to the same bouncing solution, so we can only assure the existence of two different $2m\pi$-periodic bouncing solutions when there is only one impact in each period.

3. Twist property for the successor map. The aim of this section is to provide the necessary properties for the application of the Poincaré–Birkhoff theorem yet done in section 2. Basically, our goal is to prove that the rotation for some iteration of the successor map is *slow* for small velocities and *fast* for large velocities. This will be done through some auxiliary lemmas concerning the asymptotic dynamics of the solutions for (1.1). Lemma 3.1 gives a second-order differential inequality to be used later. Lemma 3.2 shows that $\mathcal{S}$ is well defined for $v \ll 1$, Lemma 3.3 shows that the impact velocity $\hat{v}$ is small if the initial velocity v is small enough and in consequence, Lemma 3.4 gives the slow rotation for some iteration of $\mathcal{S}$ for small initial velocities. Lemma 3.5 discusses, under the assumption that the successor map is well defined, the fast rotation of $\mathcal{S}$ for large velocities by using the Sturm comparison theorem. This fact implies (Lemma 3.6) the fast rotation for some iteration of $\mathcal{S}$ for large initial velocities. At the end of this section, we discuss, in Lemmas 3.7 and 3.8, when the successor map $\mathcal{S}$ is well defined by using some oscillatory properties of the solutions of the Hill's equation.

LEMMA 3.1. *Suppose that $x_1(t)$ is a solution of the equation $x'' = Mx$ for $t \in I$, where $M > 0$, and $x_2(t)$ satisfies the differential inequality $x'' \le Mx$ for $t \in I$, with the same initial conditions $x_1(\tau) = x_2(\tau)$, $x_1'(\tau) = x_2'(\tau)$. Then $x_1(t) \ge x_2(t)$ for $t \in I$.*

Proof. Let $z_n(t) = x_n(t) - x_2(t)$, where $x_n(t)$ is the solution of $x'' = Mx$ with the initial condition $x_n(\tau) = x_2(\tau)$, $x_n'(\tau) = x_2'(\tau) + \frac{1}{n}$. Then $z_n(\tau) = 0$, $z_n'(\tau) = \frac{1}{n} > 0$ which implies that $z_n(t) > 0$ for $t > \tau$ and t close to τ. Moreover, $z_n''(t) \ge M z_n(t)$ for $t > \tau$. Thus $z_n'(t) > z_n'(\tau) > 0$ and $z_n(t)$ increases strictly for $t > \tau$. Hence $x_n(t) > x_2(t)$ for $t > \tau$. Let $n \to \infty$. Then $x_n(t) \to x_1(t)$ in any compact interval by using the continuous dependence on initial values. Therefore, $x_1(t) \ge x_2(t)$ for $t \in I$. □

LEMMA 3.2. *If $p(t) \le 0$ and $\overline{p} = \frac{1}{2\pi}\int_0^{2\pi} p(t)dt < 0$, then every solution $x(t; \tau, v)$ of (1.1) starting from $x(\tau; \tau, v) = 0$, $x'(\tau; \tau, v) = v > 0$ does not satisfy $x(t; \tau, v) = x'(t; \tau, v) = 0$ for any t in its domain. Moreover, $\mathcal{S}$ is well defined and one to one for $v \ll 1$.*

Proof. Note that every solution of (1.1) starting from $x(\tau; \tau, v) = 0$, $x'(\tau; \tau, v) = v > 0$ satisfies

$$x' = y, \quad y' = -a(t)x + p(t)$$

in (x, y)-plane before it meets $x = 0$ again. Then $x'(t; \tau, v) > 0$ when it is in the half-plane $y > 0$ which implies that $x(t; \tau, v) > 0$ for $t > \tau$ and close to τ. Moreover,

if there are $\tau_1,\ \tau_2$ such that

$$x'(\tau_1;\tau,v)=x'(\tau_2;\tau,v)=0,\qquad x'(s;\tau,v)>0\quad \text{for } s\in(\tau_1,\tau_2),$$

then

$$x(\tau_2;\tau,v)>x(\tau_1;\tau,v). \tag{3.1}$$

If $x'(\tau_3;\tau,v)=0$ and $x'(s;\tau,v)<0$ for $s\in(\tau_3,t)$, then by using polar coordinates

$$x=r\cos\theta,\qquad y=r\sin\theta$$

in the half-plane $y\le 0$ we get

$$r'=(1-a(t))r\cos\theta\sin\theta+p(t)\sin\theta\ge -Kr,$$

where $K=\max_{0\le t\le 2\pi}|1-a(t)|$. Thus

$$r(t)\ge r(\tau_3)\exp(-K(t-\tau_3)). \tag{3.2}$$

Therefore, either $x(t;\tau,v)$ has no any impact in $t>\tau$ or $x(t;\tau,v)$ has its next impact at $t=\hat\tau$. In this case, (3.1), (3.2) imply that

$$x'(\hat\tau;\tau,v)\le -x(\tilde\tau;\tau,v)\exp(-K(\hat\tau-\tau))<0,$$

where $t=\tilde\tau$ is the first time $x(t;\tau,v)$ meets $y=0$ after τ. The conclusion of the first part of the lemma is thus proved.

Next, note that when $x(t;\tau,v)$ is remaining in half-plane $x>0$,

$$x''(t;\tau,v)=-a(t)x(t;\tau,v)+p(t)\le Mx(t;\tau,v),$$

where $M=\max_{0\le t\le 2\pi}|a(t)|$. Then Lemma 3.1 implies that

$$x(t;\tau,v)\le M_0=\frac{v}{2\sqrt{M}}(\exp(2\pi\sqrt{M})-\exp(-2\pi\sqrt{M})) \tag{3.3}$$

for $t\in(\tau,\tau+2\pi)$. Thus,

$$x'(t;\tau,v)=v-\int_\tau^t(a(s)x(s;\tau,v)-p(s))ds\le O(v)+\int_\tau^t p(s)ds.$$

Because $\bar p<0$, there must be $\tilde\tau\in(\tau,\tau+2\pi)$ such that

$$x(\tilde\tau;\tau,v)>0,\qquad x'(\tilde\tau;\tau,v)=0,\qquad x'(s;\tau,v)>0\quad \text{for } s\in(\tau,\tilde\tau),$$

provided that $v\ll 1$. Moreover, for $t\in(\tilde\tau,\tau+2\pi)$, we have

$$x(t;\tau,v)=x(\tilde\tau;\tau,v)-\int_{\tilde\tau}^t\int_{\tilde\tau}^w(a(s)x(s;\tau,v)-p(s))dsdw=O(v)+\int_{\tilde\tau}^t\int_{\tilde\tau}^w p(s)dsdw$$

from which it follows that there exists $\hat\tau\in(\tilde\tau,\tau+2\pi)$ such that

$$x(\hat\tau;\tau,v)=0,\qquad x'(\hat\tau;\tau,v)<0,\qquad x(t;\tau,v)>0\quad \text{for } t\in[\tilde\tau,\hat\tau),$$

provided that $v\ll 1$ and $\bar p<0$. The lemma is thus proved. □

The next lemma clarifies the behavior of the next impact velocity $\hat{v}$ for small v.

LEMMA 3.3. *If $p(t) \le 0$ and $\overline{p} < 0$, then the next velocity $\hat{v}$ of the successor map satisfies*

$$\lim_{v \to 0^+} \hat{v}(\tau, v) = 0$$

uniformly for $\tau \in [0, 2\pi)$.

Proof. As it is shown in Lemma 3.2, for $v > 0$ small enough, we have a well-defined $\hat{\tau} \in (\tau, \tau + 4\pi)$. Moreover,

$$\max_{\tau \le t \le \hat{\tau}} x(t; \tau, v) = O(v) \quad \text{as } v \to 0^+. \tag{3.4}$$

By contradiction, let us assume that there exist $\{\tau_n\}$ belonging to $[0, 2\pi)$ and $\{v_n\}$ with $v_n \to 0^+$ as $n \to \infty$, such that $\hat{v}(\tau_n, v_n) \le -\delta < 0$. Then there exist $t_n \in (\tau_n, \hat{\tau}(\tau_n, v_n))$ satisfying

$$x'(t_n; \tau_n, v_n) = -\frac{\delta}{2}, \qquad x'(s; \tau_n, v_n) \le -\frac{\delta}{2} \quad \text{for } s \in [t_n, \hat{\tau}(\tau_n, v)].$$

Denote by $P = \|p\|_\infty$, then

$$\begin{aligned} -\frac{\delta}{2} &\ge \hat{v}(\tau_n, v_n) - x'(t_n; \tau_n, v_n) = -\int_{t_n}^{\hat{\tau}(\tau_n, v_n)} (a(s)x(s; \tau_n, v_n) - p(s))ds \\ &\ge -(M + P)(\hat{\tau}(\tau_n, v_n) - t_n), \end{aligned}$$

provided that $\max_{t_n \le t \le \hat{\tau}(\tau_n, v_n)} x(t; \tau_n, v_n) \le 1$ (this is guaranteed for v_n small by (3.4)). Thus we can estimate

$$\begin{aligned} \max_{\tau_n \le t \le \hat{\tau}_n} x(t; \tau_n, v_n) &\ge x(t_n; \tau_n, v_n) - x(\hat{\tau}(\tau_n, v_n); \tau_n, v_n) \\ &= -\int_{t_n}^{\hat{\tau}(\tau_n, v_n)} x'(s; \tau_n, v_n) \mathrm{d}s \ge \frac{\delta}{2} \cdot (\hat{\tau}(\tau_n, v_n) - t_n) \ge \frac{\delta^2}{4(M + P)}, \end{aligned}$$

which contradicts (3.4). The result is thus proved. ∎

Let us recall that we write $\mathcal{S}^n(\tau, v) = (\hat{\tau}^n(\tau, v),\ \hat{v}^n(\tau, v))$ and we will use the abbreviation $\hat{\tau}^n = \hat{\tau}^n(\tau, v)$, $\hat{v}^n = \hat{v}^n(\tau, v)$. Then, it is deduced from Lemma 3.3 that for all $n \in \mathbb{N}$ and $v_n > 0$, there exists $v_0 > 0$, such that

$$|x'(t; \tau, v)| \le v_n \quad \text{for } v \in (0, v_0],\ t \in [\tau, \hat{\tau}^n].$$

Now, suppose that there are $c > 0$ and $\delta > 0$ such that $p(t) \le -c$ for $t \in [\tau_0 - 2\delta, \tau_0 + 2\delta]$. Then,

$$\hat{\tau}^n - \tau < \delta \quad \text{for } v \ll 1 \text{ and } \tau \in [\tau_0 - \delta, \tau_0 + \delta]. \tag{3.5}$$

Actually,

$$|\hat{v}^j + \hat{v}^{j-1}| = |x'(\hat{\tau}^j; \tau, v) - x'(\hat{\tau}^{j-1}; \tau, v)| = \int_{\hat{\tau}^{j-1}}^{\hat{\tau}^j} (a(t)x(t; \tau, v) - p(t))dt \ge \frac{c}{2}(\hat{\tau}^j - \hat{\tau}^{j-1}),$$

provided that (3.4) and $[\hat{\tau}^{j-1}, \hat{\tau}^j] \subset [\tau_0 - 2\delta, \tau_0 + 2\delta]$. Then

$$\hat{\tau}^n - \tau \le \frac{4}{c} \sum_{j=1}^{n} \hat{v}^j < \delta \tag{3.6}$$

if we choose $v \ll 1$ and $\tau \in [\tau_0 - \delta, \tau_0 + \delta]$.

Now, we can prove the twist property of the successor map for $v \ll 1$.

LEMMA 3.4. *Let us suppose that $p(t) \leq 0$ and $\bar{p} < 0$. Then, for all $n, m \in \mathbb{N}$, there exists $v_n > 0$ such that*

$$\Pi_1(\mathcal{S}^n(\tau, v)) - \tau < 2m\pi \quad \text{for } v \in (0, v_n] \text{ and } \tau \in [0, 2\pi].$$

Proof. Since $p(\cdot)$ is continuous and $\bar{p} < 0$, there are $c > 0$, $\delta > 0$, and $\tau_0 \in [0, 2\pi]$ such that $p(t) \leq -c$ for $t \in [\tau_0 - 2\delta, \tau_0 + 2\delta]$. Then, there exists $v \ll 1$ such that

$$\hat{\tau}^n(\tau, v) - \tau < \delta \quad \text{for } \tau \in [\tau_0 - \delta, \tau_0 + \delta]. \tag{3.7}$$

For $\tau \in (\tau_0 + \delta, 2\pi + \tau_0 - \delta)$ either $\hat{\tau}^n(\tau, v) \leq 2\pi + \tau_0 - \delta$ which implies that

$$\hat{\tau}^n - \tau < 2\pi - 2\delta, \tag{3.8}$$

or there exists $t \in (\hat{\tau}^{j-1}, \hat{\tau}^j) \cap [2\pi + \tau_0 - \delta, 2\pi + \tau_0]$ for some $j \in \{1, 2, \ldots, n\}$. Then, by estimating like in (3.6) it is proved that, if v small enough, $\hat{\tau}^j - t \leq \frac{\delta}{n}$. From here it is deduced that

$$\hat{\tau}^j \in \left(2\pi + \tau_0 - \delta, 2\pi + \tau_0 + \frac{\delta}{n}\right] \tag{3.9}$$

and then $\hat{\tau}^n - \hat{\tau}^j < \frac{n-1}{n}\delta$ which implies that

$$\hat{\tau}^n - \tau = \hat{\tau}^n - \hat{\tau}^j + \hat{\tau}^j - t + t - \tau < \frac{n-1}{n}\delta + \frac{\delta}{n} + 2\pi + \tau_0 - (\tau_0 + \delta) = 2\pi. \tag{3.10}$$

Since $\mathcal{S}$ is continuous on $\mathbb{R} \times \mathbb{R}^+$ (this is a consequence of the uniqueness of the solution for the initial value problem for linear equation), the above estimations are uniform for the compact interval $[0, 2\pi]$. Therefore, (3.7)–(3.10) complete the proof of the lemma. ∎

Under the assumption that the successor map $\mathcal{S}$ is well defined, our next result proves the twist property for large velocities.

LEMMA 3.5. *Assume that $\mathcal{S} : (\tau, v) \mapsto (\hat{\tau}, \hat{v})$ for $(\tau, v) \in \mathbb{R} \times \mathbb{R}^+$ is well defined and $p(t) \leq 0$, $\bar{p} < 0$. Then*

$$\liminf_{v \to +\infty} [\hat{\tau}(\tau, v) - \tau] \geq \frac{\pi}{\sqrt{\|a\|_\infty}} \tag{3.11}$$

uniformly for $\tau \in [0, 2\pi)$. If $a(t) \equiv 0$, then

$$\lim_{v \to +\infty} [\hat{\tau}(\tau, v) - \tau] = +\infty \tag{3.12}$$

uniformly for $\tau \in [0, 2\pi)$.

Proof. Suppose firstly that $\|a\|_\infty > 0$ and there are $\tau_n \in [0, 2\pi)$ and $v_n > 0$ with $v_n \to +\infty$ as $n \to \infty$ such that $\hat{\tau}(\tau_n, v_n) - \tau_n \leq \pi/\sqrt{\|a\|_\infty} - \gamma$ with $\gamma > 0$. Then there are $\tau_* \in [0, 2\pi]$ and $\hat{\tau}_* \in (\tau_*, \tau_* + \pi/\sqrt{\|a\|_\infty} - \gamma]$ such that $\tau_n \to \tau_*$ and $\hat{\tau}(\tau_n, v_n) \to \hat{\tau}_*$ as $n \to \infty$. Moreover, $y_n(t) = x(t; \tau_n, v_n)/v_n$ is the solution of the equation

$$x'' + a(t)x = \frac{1}{v_n}p(t)$$

with the initial conditions $y_n(\tau_n) = 0,\ y_n'(\tau_n) = 1$. By continuous dependence of the solutions with respect to initial value and parameters we have that

$$\lim_{n\to\infty} y_n(t) = y_0(t) \quad \text{and} \quad \lim_{n\to\infty} y_n'(t) = y_0'(t) \tag{3.13}$$

uniformly on compact intervals, where $y_0(t)$ is the solution of Hill's equation (1.5) with the initial condition $y_0(\tau_*) = 0,\ y_0'(\tau_*) = 1$. Thus $y_0(\hat{\tau}_*) = 0$ due to the continuous dependence of the solutions with respect to initial values and parameters. On the other hand, by using Sturm comparison theorem, it is proved that

$$\tau' - \tau \geq \frac{\pi}{\sqrt{\|a\|_\infty}},$$

where τ' and τ are two consecutive zeros of $y_0(t)$, so in consequence $\hat{\tau}_* - \tau_* \geq \pi/\sqrt{\|a\|_\infty}$. This is a contradiction. If $a(t) \equiv 0$, then any solution $x(t;\tau,v)$ of the equation $x'' = p(t)$ has the derivative $x'(t;\tau,v) = v + \int_\tau^t p(s)\mathrm{d}s$. Hence, for any fixed $v > 0$ there exists a $\hat{\tau} > \tau$ such that $x(\hat{\tau};\tau,v) = \int_\tau^{\hat{\tau}}(v + \int_\tau^t p(s)\mathrm{d}s)\mathrm{d}t = 0$ and $\lim_{v\to+\infty}(\hat{\tau} - \tau) = +\infty$. Therefore, the lemma is proved. □

From the above estimation we can prove the twist property of successor map for $v \gg 1$. Recall that $r(t) = (x^2(t;\tau,v) + (x'(t;\tau,v))^2)^{1/2}$ satisfies

$$-Kr(t) - P \leq r'(t) \leq Kr(t) + P \quad \text{for } t \in (\tau, \hat{\tau}),$$

where $K = \max_{0\leq t\leq 2\pi} |1 - a(t)|$ and $P = \max_{0\leq t\leq 2\pi} |p(t)|$. Then, by using Gronwall inequality,

$$\left(v + \frac{P}{K}\right)\exp(-KT) \leq \hat{v} + \frac{P}{K} \leq \left(v + \frac{P}{K}\right)\exp(KT), \tag{3.14}$$

provided that $\hat{\tau} - \tau \leq T$. Suppose that $\Pi_1(\mathcal{S}^n(\tau,v)) - \tau \leq 2m\pi$, then

$$\Pi_1(\mathcal{S}^{i+1}(\tau,v)) - \Pi_1(\mathcal{S}^i(\tau,v)) \leq 2m\pi \quad \text{for } i = 0, 1, \ldots, n-1.$$

This implies that

$$\begin{aligned}\left(\Pi_2(\mathcal{S}^i(\tau,v)) + \frac{P}{K}\right)\exp(-2m\pi K) &\leq \Pi_2(\mathcal{S}^{i+1}(\tau,v)) + \frac{P}{K} \\ &\leq \left(\Pi_2(\mathcal{S}^i(\tau,v)) + \frac{P}{K}\right)\exp(2m\pi K)\end{aligned}$$

for $i = 0, 1, \ldots, n-1$. Thus for a given $v_0^+ > 0$ we have $v_{n,m}^+ > 0$ such that

$$\text{if } v > v_{n,m}^+ \text{ and } \Pi_1(\mathcal{S}^n(\tau,v)) - \tau \leq 2m\pi, \text{ then } \Pi_2(\mathcal{S}^i(\tau,v)) > v_0^+ \tag{3.15}$$

for $i = 0, 1, \ldots, n-1$. Hence, the following result is obtained.

LEMMA 3.6. *Let us suppose that $p(t) \leq 0$ and $\bar{p} < 0$. Let $n, m \in \mathbb{N}$ be such that $n > 2m(\sqrt{\|a\|_\infty})$. Then, there exists $v_{n,m}^+ > 0$ such that*

$$\Pi_1(\mathcal{S}^n(\tau,v)) - \tau > 2m\pi \quad \textit{for } v \geq v_{n,m}^+ \textit{ and } \tau \in [0, 2\pi].$$

Proof. From Lemma 3.5 we know that there exists $v_0^+ > 0$ such that

$$\Pi_1(\mathcal{S}(\tau,v)) - \tau \geq \frac{\pi}{\sqrt{\|a\|_\infty}} \quad \text{for } v \geq v_0^+ \text{ and } \tau \in [0, 2\pi). \tag{3.16}$$

By the periodicity of the equation, it is verified that

$$\mathcal{S}(\tau + 2\pi, v) = \mathcal{S}(\tau, v) + (2\pi, 0).$$

This means that the function $f(\tau) = \Pi_1(\mathcal{S}(\tau, v)) - \tau$ is 2π-periodic. Therefore, (3.16) holds for all $\tau \in \mathbb{R}$. Taking $v^+_{n,m}$ as in (3.15), if $v > v^+_{n,m}$ then either $\Pi_1(\mathcal{S}^n(\tau, v)) - \tau > 2m\pi$ or $\Pi_1(\mathcal{S}^n(\tau, v)) - \tau \leq 2m\pi$. In the second case, it follows from (3.15) that $\Pi_2(\mathcal{S}^i(\tau, v)) > v_0^+$ for $i = 0, 1, \ldots, n-1$, and in consequence for every $i = 1, \ldots, n-1$ we have

$$\Pi_1(\mathcal{S}^{i+1}(\tau, v)) - \Pi_1(\mathcal{S}^i(\tau, v)) \geq \frac{\pi}{\sqrt{\|a\|_\infty}} \quad \text{for } v \geq v^+_{n,m} \text{ and } \tau \in [0, 2\pi).$$

Adding the previous inequalities for $i = 1, \ldots, n-1$ with (3.16),

$$\Pi_1(\mathcal{S}^n(\tau, v)) - \tau \geq n\frac{\pi}{\sqrt{\|a\|_\infty}}.$$

Now, taking into account that $n > 2m(\sqrt{\|a\|_\infty})$, the result is done. □

The rest of this section is devoted to the discussion of the conditions implying that the successor map is well defined. At first we can prove as in the proof of Lemma 3.2 that successor map is well defined if $p(t) \leq 0$ for all t, $\bar{p} < 0$ and $a(t) \equiv 0$. We will prove in the following that $\bar{a} > 0$ is also enough to assure that the successor map is well defined. With this, the proofs of Corollaries 1.3 and 1.4 are completed.

Consider the solution $x(t; \tau, v)$ of the impact oscillator (1.1) starting from

$$x(\tau; \tau, v) = 0, x'(\tau; \tau, v) = v > 0.$$

Lemma 3.2 implies that either there exists $\hat{\tau} > \tau$ such that $x(\hat{\tau}; \tau, v) = 0$ and $x(t; \tau, v) > 0$ for $t \in (\tau, \hat{\tau})$, or

$$x(t; \tau, v) > 0 \quad \text{for all } t > \tau, \tag{3.17}$$

and in consequence $x(t; \tau, v)$ is a (classical) solution of the equation $x'' + a(t)x = p(t)$, with $t > \tau$. If (3.17) holds, we will show that there is a constant $\delta > 0$ such that $x(t; \tau, v) \geq \delta$ for sufficiently large $t > \tau$. Actually, we will show firstly that (3.17) implies that

$$|x(t; \tau, v)| + |x'(t; \tau, v)| \geq 2\delta. \tag{3.18}$$

By contradiction, let us suppose that there exists $\tau_1 > \tau$ such that $x(\tau_1; \tau, v) = \alpha \geq 0$, $x'(\tau_1; \tau, v) = \beta$ with $|\alpha| + |\beta| < 2\delta$. Then as in Lemma 3.2 it is shown that

$$x(t; \tau, v) \leq \frac{1}{2\sqrt{M}}((\sqrt{M}\alpha + \beta)\exp(2\pi\sqrt{M}) + (\sqrt{M}\alpha - \beta)\exp(-2\pi\sqrt{M}))$$

for $t \in (\tau_1, \tau_1 + 2\pi)$, being $M = \|a\|_\infty$. Thus,

$$\begin{aligned} x(t; \tau, v) &= \alpha + \int_{\tau_1}^t \left(\beta + \int_{\tau_1}^s (-a(\xi)x(\xi; \tau_1, v) + p(\xi)\mathrm{d}\xi)\mathrm{d}s\right) \\ &= O(|\alpha| + |\beta|) + \int_{\tau_1}^t \int_{\tau_1}^s (p(\xi))\mathrm{d}\xi\mathrm{d}s. \end{aligned}$$

This implies that, using $\bar{p} < 0$, if δ is small enough, then there must be some $\hat{\tau} \in (\tau_1, \tau_1 + 2\pi)$ such that $x(\hat{\tau}; \tau, v) = 0$. This contradicts (3.17).

Note that (3.18) implies that $v \geq 2\delta$. Moreover, there exists $t_1 > \tau$ such that $x(t_1; \tau, v) \geq \delta$. We claim that

$$x(t; \tau, v) \geq \delta \quad \text{for } t \geq t_1. \tag{3.19}$$

If (3.19) is not true, let $t_2 = \inf\{t : t \geq t_1,\ x(t; \tau, v) < \delta\}$. Then $x'(t_2; \tau, v) \leq 0$. If $x'(t; \tau, v) \leq 0$ for $t \geq t_2$, then $x(t; \tau, v) \leq x(t_2; \tau, v) \leq \delta$ for $t \geq t_2$, and (3.18) implies that $x'(t; \tau, v) < -\delta$ for $t \geq t_2$. Thus

$$x(t; \tau, v) = x(t_2; \tau, v) + \int_{t_2}^{t} x'(s; \tau, v)\mathrm{d}s \leq -\delta(t - t_2) + \delta < 0$$

for $t > t_2 + 1$ which contradicts (3.17). Hence, we can define $t_3 = \inf\{t : t \geq t_2,\ x'(t; \tau, v) > 0\}$. Clearly, $x'(t_3; \tau, v) = 0$ and

$$x(t_3; \tau, v) = x(t_2; \tau, v) + \int_{t_2}^{t_3} x'(s; \tau, v)\mathrm{d}s \leq x(t_2; \tau, v) \leq \delta$$

which contradicts (3.18). Therefore, we have proved the following result.

LEMMA 3.7. *There exists $\delta > 0$ independent of (τ, v) such that if $\mathcal{S}$ is not defined for some $(\tau, v) \in \mathbb{R} \times \mathbb{R}^+$, then $x(t; \tau, v) \geq \delta$ for $t \gg 1$.*

Now we assume that Hill's equation is oscillatory, that is, all nonzero solution of Hill's equation have infinitely many zeros. It is a known fact (see [26]) that these zeros correspond to a sequence tending to $+\infty$.

LEMMA 3.8. *Let us assume that Hill's equation* (1.5) *is oscillatory. Then there exist $\beta_0 > 0$ and $\varepsilon_0 > 0$ such that any solution $x(t; \tau, v)$ of* (1.1) *such that $x(\tau_1; \tau, v) = \alpha,\ x'(\tau_1; \tau, v) = \beta$ with $\beta \geq \beta_0$ and $0 \leq \alpha \leq \varepsilon_0\beta$ will have a next zero $\hat{\tau} > \tau_1$.*

Proof. Let $y_\beta(t) := \frac{1}{\beta}x(t; \tau, v)$. Then, $y_\beta(t)$ is a solution of the equation

$$x'' + a(t)x = \frac{1}{\beta}p(t)$$

for $t > \tau_1$ and $y_\beta(s) > 0$ for $s \in (\tau_1, t)$ with initial conditions

$$y_\beta(\tau_1) = \frac{\alpha}{\beta}, \qquad y'_\beta(\tau_1) = 1.$$

If $y_0(t)$ is the solution of the Hill's equation $x'' + a(t)x = 0$ with initial conditions $y_0(\tau_1) = 0,\ y'_0(\tau_1) = 1$, by continuous dependence of the solutions with respect to initial value and parameters we have that

$$\lim_{\beta \to +\infty} y_\beta(t) = y_0(t) \quad \text{and} \quad \lim_{\beta \to +\infty} y'_\beta(t) = y'_0(t) \tag{3.20}$$

uniformly on compact interval. Let $\hat{\tau}_0(\tau_1)$ be the next zero of $y_0(t)$ after τ_1 (that is, $y_0(\hat{\tau}_0) = 0$ and $y_0(t) > 0$ for all $\tau_1 < t < \hat{\tau}_0$). Then, $\hat{\tau}_0(\tau_1)$ is a simple zero with $y'_0(\hat{\tau}_0) < 0$ independent of β. Thus (3.20) implies that for β large enough and $\frac{\alpha}{\beta}$ small enough there exists $\hat{\tau}(\tau_1)$ such that $y_\beta(\hat{\tau}) = 0$. The lemma is thus proved. ∎

A direct consequence of the above lemma is that the successor map $\mathcal{S}$ for the impact oscillator (1.1) is well defined for $v \gg 1$. As shown in [26], the condition $\bar{a} > 0$ implies that Hill's equation (1.5) is oscillatory. This condition is also enough to

assure that our successor map is well defined. Actually, if $\mathcal{S}$ is not defined for some (τ, v) with $v > 0$, then by using Lemmas 3.7 and 3.8, the solution $x(t;\tau,v)$ will satisfy $x(t;\tau,v) \geq \delta$ and $|\frac{x'(t;\tau,v)}{x(t;\tau,v)}| < \max\{\frac{\beta_0}{\delta}, \frac{1}{\varepsilon_0}\}$ for t large enough. Now, by integrating (1.1) in $[2l\pi, 2k\pi]$ for $l, k \in \mathbb{N}$ we have

$$\int_{2l\pi}^{2k\pi} \left(\frac{x''}{x}\right) \mathrm{d}t + 2\pi(k-l)\bar{a} \leq 0,$$

but this implies that

$$2\pi(k-l)\bar{a} + \int_{2l\pi}^{2k\pi} \frac{(x')^2}{x^2} \mathrm{d}t - 2\max\left\{\frac{\beta_0}{\delta}, \frac{1}{\varepsilon_0}\right\} \leq 0.$$

It is clear that this is not possible if k is large enough. Therefore the successor map $\mathcal{S}$ is well defined for all (τ, v) with $v > 0$, provided that $p(t) \leq 0$ for all $t, \bar{p} < 0$ and $\bar{a} > 0$.

4. Asymptotically linear impact oscillators. Finally, we discuss the case of the asymptotically linear impact oscillator (1.8). Throughout this section, it is understood that the assumptions of Theorem 1.8 hold. Such assumptions imply that there exist $M, P > 0$ such that $|a(t)x + g(t,x) - p(t)| \leq Mx + P$ for $x \geq 0$ and for all t. Moreover, the successor map $\mathcal{S}$ of the problem (1.8) is well defined for all $(\tau, v) \in \mathbb{R} \times \mathbb{R}^+$ and $p(t) \leq 0$ for all t, $\bar{p} < 0$. Then, by using similar arguments as in Lemmas 3.2 and 3.3, it is easy to prove that the conclusions of Lemma 3.4 are still valid for the successor map of problem (1.8). Roughly speaking, Lemma 3.4 means that the rotation of some iteration of the successor map is slow for small velocities. On the other hand, it is necessary to control the behavior of the successor map for large velocities (that is, an analogous to Lemma 3.6). To this purpose, some lemmas are needed. For the moment, let us assume that $\|a\|_\infty > 0$.

LEMMA 4.1. *The solution $x(t;\tau,v)$ of problem (1.8) with initial conditions $x(\tau;\tau,v) = 0$, $x'(\tau;\tau,v) = v > 0$ satisfies*

$$\left(r(\tau_1) + \frac{P}{M+1}\right)\exp(-(M+1)(\tau_2 - \tau_1)) \leq r(\tau_2) + \frac{P}{M+1} \leq \left(r(\tau_1) + \frac{P}{M+1}\right)\exp((M+1)(\tau_2 - \tau_1)) \tag{4.1}$$

for $\tau_2 - \tau_1 \geq 0$, where $r(t) = ((x(t;\tau,v))^2 + (x'(t;\tau,v))^2)^{1/2}$.

Proof. This inequality is proved by using the Gronwall lemma as in (3.14). ∎

Let $n, m \in \mathbb{N}$ be such that $n > 2m(\sqrt{\|a\|_\infty})$. Then, there exists $\sigma > 0$, such that $n > 2m(\sqrt{\|a\|_\infty} + 2\sigma)$. Let us fix the positive numbers $T = \frac{n\pi}{\sqrt{\|a\|_\infty}}$ and

$$\delta = \frac{1}{2}\left|\frac{\pi}{\sqrt{\|a\|_\infty + 2\sigma}} - \frac{\pi}{\sqrt{\|a\|_\infty + \sigma}}\right|.$$

Note that σ (and in consequence δ) can be chosen arbitrarily small. By using the assumption (1.7), it is possible to take $d > 0$ (depending on σ) such that

$$\max_{0 \leq t \leq 2\pi} |a(t)x + g(t,x) - p(t)| \leq (\|a\|_\infty + \sigma)x \quad \text{for } x \geq d.$$

The following estimation is obtained by using the previous lemma.

LEMMA 4.2. *Let $x(t;\tau,v)$ be the solution of* (1.8) *with initial conditions*

$$x(\tau;\tau,v)=0,\quad x'(\tau;\tau,v)=v>0.$$

Then for d, δ, and $T>0$ as given before, there exists $v_\delta>0$ such that if $v\geq v_\delta$, then there exists $\tau_d^+>\tau$ such that $x(\tau_d^+;\tau,v)=d$ and $x(t;\tau,v)<d$ for $t\in(\tau,\tau_d^+)$, and moreover $|\tau_d^+-\tau|<\delta$. Besides, if there exists $\tau_d^->\tau_d^+$ such that $x(\tau_d^-;\tau,v)=d$, $x'(\tau_d^-;\tau,v)<0$, and $|\tau_d^--\tau|<T$, then there exists $\hat\tau>\tau_d^+$ such that $x(\hat\tau;\tau,v)=0$ and $|\hat\tau-\tau_d^-|<\delta$. Moreover, if δ is small enough, then

$$\frac{v_d^+}{2}\leq v\leq 2v_d^+. \tag{4.2}$$

Proof. Firstly, the global existence of $x(t;\tau,v)$ right to τ is assured from the assumptions. Suppose there is no time $t\in(\tau,\tau+1)$ such that $x(t;\tau,v)=d$, that is, $0<x(t;\tau,v)<d$ for $t\in(\tau,\tau+1)$. Then

$$x'(t;\tau,v)>\left(v+\frac{P}{M+1}\right)\exp(-(M+1))-d-\frac{P}{M+1},$$

so an integration gives

$$x(\tau+1;\tau,v)>\left(v+\frac{P}{M+1}\right)\exp(-(M+1))-d-\frac{P}{M+1}>d$$

if v is large enough. Thus we have proved the existence of τ_d^+. Moreover,

$$v_d^+=x'(\tau_d^+;\tau,v)>\left(v+\frac{P}{M+1}\right)\exp(-(M+1)(\tau_d^+-\tau))-d-\frac{P}{M+1}.$$

Hence

$$\begin{aligned}d&=\int_\tau^{\tau_d^+}x'(s;\tau,v)\mathrm{d}s\\&\geq\left[\left(v+\frac{P}{M+1}\right)\exp(-(M+1)(\tau_d^+-\tau))-d-\frac{P}{M+1}\right](\tau_d^+-\tau)\end{aligned}$$

and in consequence for a given δ we get $|\tau_d^+-\tau|<\delta$ by taking v large enough. The discussion for $\hat\tau$ is similar. Finally, if δ is small enough, then

$$\left(v_d^++\frac{P}{M+1}\right)\exp(-(M+1)\delta)-\frac{P}{M+1}\geq\frac{v_d^+}{2}$$

and

$$\left(v_d^++d+\frac{P}{M+1}\right)\exp(-(M+1)\delta)-\frac{P}{M+1}\leq 2v_d^+.$$

Now, the estimation (4.2) follows easily from (4.1). ∎

Define now

$$h(t,x)=\begin{cases}\dfrac{a(t)x+g(t,x)-p(t)}{x},&x\geq d;\\[2ex]\dfrac{a(t)d+g(t,d)-p(t)}{d},&x<d.\end{cases}$$

Then $h(t,x)$ is continuous and 2π-periodic with respect to t and verifies $|h(t,x)| \leq \|a\|_\infty + \sigma$ for $x \geq 0$ and for all t. Let $x(t;\tau,v)$ be the solution of the equation $x''+h(t,x)x = 0$ satisfying initial conditions $x(\tau;\tau,v) = 0$, $x'(\tau;\tau,v) = v > 0$. On the other hand, let $y_0(t;\tau,v)$ be the solution of the equation $x''+(\|a\|_\infty+\sigma)x = 0$ satisfying the same initial conditions as $x(t;\tau,v)$. By using Sturm comparison theorem,

$$\hat{\tau}(h) - \tau \geq \frac{\pi}{\sqrt{\|a\|_\infty + \sigma}},$$

where $\hat{\tau}(h)$ is the next zero of $x(t;\tau,v)$ right to τ. Moreover, we have the following lemma.

LEMMA 4.3. *Let $x(t;\tau,v)$ be the solution of the equation $x'' + h(t,x)x = 0$ satisfying the initial conditions*

$$x(\tau;\tau,v) = 0, \quad x'(\tau;\tau,v) = v > 0.$$

Then, there is $v_\delta > 0$ such that if $v \geq v_\delta$, then there exist τ_d^+, τ_d^- such that $x(\tau_d^+;\tau,v) = d$, $x'(\tau_d^+;\tau,v) = v_d^+ > 0$, $x(t;\tau,v) < d$ for $t \in (\tau,\tau_d^+)$ and $x(\tau_d^-;\tau,v) = d$, $x'(\tau_d^-;\tau,v) = v_d^- < 0$, $x(t;\tau,v) > 0$ for $t \in (\tau,\tau_d^-)$, respectively. Moreover,

$$\tau_d^- - \tau_d^+ > \frac{\pi}{\sqrt{\|a\|_\infty + 2\sigma}}.$$

Proof. Recall that $|h(t,x)x| \leq (\|a\|_\infty + \sigma)x$ for $x \geq 0$ and for all t, so the conclusion of Lemmas 4.1 and 4.2 are valid for $x(t;\tau,v)$ if $v > 0$ is sufficiently large, thus we have $\tau_d^+ - \tau < \delta$. Note that $h(t,x)x = a(t)x + g(t,x) - p(t)$ for $x \geq d$ and for all t. This implies, under the assumption of Theorem 1.8, that there exists $\tau_d^- > \tau_d^+$ such that $x(\tau_d^-;\tau,v) = d$, $x'(\tau_d^-;\tau,v) = v_d^- < 0$, and $x(t;\tau,v) > 0$ for $t \in (\tau,\tau_d^-)$. By contradiction, if

$$\tau_d^- - \tau_d^+ \leq \frac{\pi}{\sqrt{\|a\|_\infty + 2\sigma}},$$

then $\tau_d^- - \tau < \tau_d^- - \tau_d^+ + \delta < T$, and Lemma 4.2 implies that the zero $\hat{\tau}(h)$ right to τ exists and $\hat{\tau}(h) - \tau_d^- < \delta$. Hence,

$$\tau_d^- - \tau_d^+ > \hat{\tau}(h) - \tau - 2\delta \geq \frac{\pi}{\sqrt{\|a\|_\infty + \sigma}} - 2\delta = \frac{\pi}{\sqrt{\|a\|_\infty + 2\sigma}}.$$

This contradiction completes the proof of Lemma 4.3. ∎

Finally, let us consider $x(t;\tau,v)$ the solution of (1.8) with initial conditions

$$x(\tau;\tau,v) = 0, \quad x'(\tau;\tau,v) = v > 0.$$

Let $\hat{\tau}$ be the first zero right to τ. If v is large enough, then there exist τ_d^-, $\tau_d^+ \in (\tau,\hat{\tau})$ such that $x(\tau_d^\pm;\tau,v) = d$, $v_d^+ = x'(\tau_d^+;\tau,v) > 0$, $v_d^- = x'(\tau_d^-;\tau,v) < 0$, and $x(t;\tau,v) < d$ for $t \in (\tau,\tau_d^+) \cup (\tau_d^-,\hat{\tau})$. Moreover, $|\tau_d^+ - \tau| < \delta$ and v_d^+ is arbitrarily large by using the estimation (4.2). On the other hand, let $x_h(t)$ be the solution of the equation $x'' + h(t,x)x = 0$ satisfying $x_h(\tau_d^+) = d$, $x_h'(\tau_d^+) = v_d^+ > 0$. If τ_h is such that $x_h(\tau_h) = 0$, $x_h(t) > 0$ for $t \in (t_h,\tau_d^+)$, then the estimation (4.2) implies that the initial velocity $v_h = x_h'(\tau_h)$ is arbitrarily large. Taking into account that $h(t,x)x = a(t)x + g(t,x) - p(t)$ for $x \geq d$ and for all t, Lemma 4.2 implies that the time in which the solution $x(t;\tau,v)$ of the equation $x'' + a(t)x + g(t,x) = p(t)$ moves

from (d_σ, v_d^+) to (d_σ, v_d^-) is larger than $\frac{\pi}{\sqrt{\|a\|_\infty+2\sigma}}$. In consequence, if v large enough (more explicitly, $v \geq v_\delta$), then we have

$$\hat{\tau} - \tau \geq \frac{\pi}{\sqrt{\|a\|_\infty + 2\sigma}}. \tag{4.3}$$

Looking for the estimation of $\Pi_1(\mathcal{S}^n(\tau, v)) - \tau$, note that by using the argument leading to (3.15), it results that for a given $v_\delta > 0$ there is $v_{n,m}^+(\delta) > 0$ such that

$$\text{if } v > v_{n,m}^+(\delta) \quad \text{and} \quad \Pi_1(\mathcal{S}^n(\tau, v)) - \tau \leq 2m\pi, \text{ then } \Pi_2(\mathcal{S}^i(\tau, v)) > v_\delta \tag{4.4}$$

for $i = 0, 1, \ldots, n-1$. Hence, following the arguments of section 3, we can prove that the conclusions of Lemma 3.6 are true for the successor map of the problem (1.8) under the assumptions of Theorem 1.8. Note that if $a(t) \equiv 0$, then T is not well defined, but it is easy to prove, by using similar arguments as before, that $\Pi_1(\mathcal{S}(\tau, v)) - \tau \geq 2m\pi$ for v sufficiently large. Now, Theorem 1.8 can be proved by mimicking the arguments of sections 2 and 3 with minor modifications.

The property that the successor map $\mathcal{S}$ is well defined is not easy to check. For example, consider the equation $x'' - x = -1$. It has a singular point $(1, 0)$ in $x - x'$ phase plane and the solution $x(t; \tau, 1)$ starting from $x(\tau; \tau, 1) = 0$, $x'(\tau; \tau, 1) = 1$ will tend to $(1, 0)$ in $x - x'$ phase plane as $t \to +\infty$. Thus we can construct an equation by modifying the above equation such that the new equation is asymptotically linear and the successor map $\mathcal{S}$ of this equation is well defined for v sufficiently small and sufficiently large but $\mathcal{S}$ is not well defined for $v = 1$. In spite of that, in the following we show that $a(t)x + g(t, x) \geq 0$ is a sufficient condition to have $\mathcal{S}$ well defined. Actually, let us note that

$$\begin{aligned} x'(t; \tau, v) &= v - \int_\tau^t (a(s)x + g(s, x))\mathrm{d}s + \int_\tau^t p(s)\mathrm{d}s \\ &\leq v + \int_\tau^t p(s)\mathrm{d}s \to -\infty \quad \text{as } t \to +\infty. \end{aligned}$$

Thus, for any fixed $v > 0$, there exists a $\hat{\tau} > \tau$ such that $x(\hat{\tau}; \tau, v) = 0$ which implies that $\mathcal{S}$ is well defined for (τ, v). Hence Corollary 1.9 is proved.

Acknowledgments. The authors thank Prof. R. Ortega for his help in understanding the bouncing problem. Finally, they thank the referees for their help and valuable suggestions.

REFERENCES

[1] V.M. Alekseev, *Quasirandom dynamical systems.* II. *One-dimensional nonlinear oscillations in a field with periodic perturbation*, Sb. Math., 6 (1968), pp. 506–560.
[2] T.W. Arnold and W. Case, *Nonlinear effects in a simple mechanical system*, Amer. J. Phys., 50 (1982), pp. 220–224.
[3] V.I. Babitsky, *Theory of Vibro-Impact Systems*, Springer-Verlag, Berlin, 1998.
[4] C.N. Bapat, *Periodic motions of an impact oscillator*, J. Sound Vibration, 209 (1998), pp. 43–60.
[5] C.N. Bapat, S. Sankar, and N. Popplewell, *Repeated impacts on a sinusoidally vibrating table reappraised*, J. Sound Vibration, 108 (1986), pp. 99–115.
[6] D. Bonheure and C. Fabry, *Periodic motions in impact oscillators with perfectly elastic bouncing*, Nonlinearity, 15 (2002), pp. 1281–1298.
[7] P. Boyland, *Dual billiards, twist maps and impact oscillators*, Nonlinearity, 9 (1996), pp. 1411–1438.

[8] L. Brindeu, *Stability of the periodic motions of the vibro-impact systems*, Chaos Solitons Fractals, 11 (2000), pp. 2493–2503.
[9] M. Corbera and J. Llibre, *Periodic orbits of a collinear restricted three body problem*, Celestial Mech. Dynam. Astronom., 86 (2003), pp. 163–183.
[10] K. Czolczynski and T. Kapitaniak, *Influence of the mass and stiffness ratio on a periodic motion of two impacting oscillators*, Chaos Solitons Fractals, 17 (2003), pp. 1–10.
[11] W. Ding, *A generalization of the Poincaré-Birkhoff theorem*, Proc. Amer. Math. Soc., 88 (1983), pp. 341–346.
[12] T. Ding and F. Zanolin, *Periodic solutions of Duffing's equations with superquadratic potential*, J. Differential Equations, 79 (1992), pp. 328–378.
[13] R.M. Everson, *Chaotic dynamics of a bouncing ball*, Phys. D, 19 (1986), pp. 355–383.
[14] W. Fang and J.A. Wickert, *Response of a periodically driven impact oscillator*, J. Sound Vibration, 170 (1994), pp. 397–409.
[15] E. Fermi, *On the origin of cosmic radiation*, Phys. Rev., 75 (1949), pp. 1169–1174.
[16] J. Franks, *Generalizations of the Poincaré-Birkhoff theorem*, Ann. of Math. (2), 128 (1988), pp. 139–151.
[17] L.K. Forbes, *A series analysis of forced transverse oscillations in a spring-mass system*, SIAM J. Appl. Math., 49 (1989), pp. 704–719.
[18] M.B. Hindmarsh and D.J. Jefferiest, *On the motions of the offset impact oscillator*, J. Phys. A, 17 (1984), pp. 1791–1803.
[19] P.J. Holmes, *The dynamics of repeated impacts on a sinusoidally vibrating table*, J. Sound Vibration, 84 (1982), pp. 173–189.
[20] H. Jacobowitz, *Periodic solutions of $x'' + f(x,t) = 0$ via the Poincaré-Birkhoff theorem*, J. Differential Equations, 20 (1976), pp. 37–52.
[21] T. Kapitaniak and M. Wiercigroch, *Dynamics of impact systems*, Chaos Solitons Fractals, 11 (2000), pp. 2411–2412.
[22] S.A. Kember and V.I. Babitsky, *Excitation of vibro-impact systems by periodic impulses*, J. Sound Vibration, 227 (1999), pp. 427–447.
[23] M. Kunze, *Non-Smooth Dynamical Systems*, Lecture Notes in Math. 1744, Springer-Verlag, Berlin, 2000.
[24] H. Lamba, *Chaotic, regular and unbounded behaviour in the elastic impact oscillator*, Phys. D, 82 (1995), pp. 117–135.
[25] A. Lazer and P. McKenna, *Periodic bouncing for a forced linear spring with obstacle*, Differential Integral Equations, 5 (1992), pp. 165–172.
[26] W. Magnus and S. Winkler, *Hill's Equation*, Dover, New York, 1966.
[27] A. Margheri, C. Rebelo, and F. Zanolin, *Maslov index, Poincaré-Birkhoff theorem and periodic solutions of asymptotically linear planar Hamiltonian systems*, J. Differential Equations, 183 (2002), pp. 342–367.
[28] J. Massera, *The existence of periodic solutions of systems of differential equations*, Duke Math. J., 17 (1950), pp. 457–475.
[29] W. Massey, *Singular Homology Theory*, Springer-Verlag, New York, 1980.
[30] J. Moser, *Stable and Random Motions in Dynamical Systems*, Ann. of Math. Stud., Princeton University Press, Princeton, NJ, 1973.
[31] R. Ortega, *Asymmetric oscillators and twist mappings*, J. London Math. Soc. (2), 53 (1996), pp. 325–342.
[32] R. Ortega, *Boundedness in a piecewise linear oscillator and a variant of the small twist theorem*, Proc. London Math. Soc. (3), 79 (1999), pp. 381–413.
[33] R. Ortega, *Dynamics of a forced oscillator with obstacle*, in Variational and Topological Methods in the Study of Nonlinear Phenomena, V. Benci et al., eds., Birkhäuser, Boston, 2001, pp. 77–89.
[34] D. Qian and P.J. Torres, *Bouncing solutions of an equation with attractive singularity*, Proc. Roy. Soc. Edinburgh Sect. A, 134 (2004), pp. 210–213.
[35] V. Zharnitsky, *Invariant curve theorem for quasiperiodic twist mappings and stability of motion in the Fermi-Ulam problem*, Nonlinearity, 13 (2000), pp. 1123–1136.

SIAM J. MATH. ANAL.
Vol. 36, No. 6, pp. 1726–1741

SPECTRAL STABILITY OF LOCAL DEFORMATIONS OF AN ELASTIC ROD: HAMILTONIAN FORMALISM*

S. LAFORTUNE† AND J. LEGA‡

Abstract. Hamiltonian methods are used to obtain a necessary and sufficient condition for the spectral stability of pulse solutions to two coupled nonlinear Klein–Gordon equations. These equations describe the near-threshold dynamics of an elastic rod with circular cross section. The present work completes and extends a recent analysis of the authors' [*Phys. D*, 182 (2003), pp. 103–124], in which a sufficient condition for the instability of "nonrotating" pulses was found by means of Evans function techniques.

Key words. spectral stability, Hamiltonian methods, elastic filament

AMS subject classifications. 37K45, 37K50, 37K05, 35P05, 35Q72

DOI. 10.1137/S0036141004439350

1. Introduction.

Coherent structures, such as fronts, pulses, defects, or solitons, play an important role in the dynamics of many physical, optical, or biological systems (see for instance [1]). When such systems are modeled in terms of partial differential equations, a coherent structure is often described as one or a family of solutions, which asymptotically connects simple plane waves or stationary states of the system. Numerous numerical and analytical techniques have been developed to study the stability of coherent structures. In particular, the Evans function [2, 3, 4, 5, 6, 7, 8, 9] can be used to obtain information on the point spectrum of one-dimensional linear operators. This approach is very general [9], but only a sufficient condition for the instability of a solution is typically found analytically. A numerical investigation of the number of zeros of the Evans function in the right half complex plane is therefore often necessary to obtain complete spectral stability results (see for instance [10] and the references therein). On the other hand, in the case of Hamiltonian systems with symmetries, global techniques are available, which give sufficient conditions for orbital stability (or instability) [11, 12].

In this paper, we apply and also extend such techniques to obtain a necessary and sufficient criterion for the spectral stability of a family of pulse solutions of two coupled nonlinear Klein–Gordon equations. These equations describe the near-threshold dynamics of an elastic filament with circular cross section [13]. The family of coherent structures we are interested in has two parameters, the speed c at which each pulse travels, and the frequency ω at which the filament rotates about its axis. In a recent paper [14], we used Evans function techniques to obtain a criterion that guarantees the instability of "nonrotating" ($\omega = 0$) pulses. Because the nonlinear Klein–Gordon equations are Hamiltonian and invariant under space translations as well as gauge invariant [15], we show here that one can take advantage of these properties to obtain spectral stability results for both "rotating" ($\omega \neq 0$) as well as nonrotating pulses.

*Received by the editors January 5, 2004; accepted for publication (in revised form) August 2, 2004; published electronically May 20, 2005. This material is based upon work supported by the National Science Foundation under grant DMS-0075827 to J. L.

http://www.siam.org/journals/sima/36-6/43935.html

†Department of Mathematics, College of Charleston, 66 George Street, Charleston, SC 29424-0001 (lafortunes@cofc.edu).

‡Department of Mathematics, University of Arizona, 617 N. Santa Rita, P.O. Box 210089, Tucson, AZ 85721-0089 (lega@math.arizona.edu).

The Hamiltonian formalism we use below is described in [12], and is a generalization to systems whose symmetry group has a dimension larger than 1, of the technique discussed in [11]. For the purpose of this study, the method can be simplified and summarized as follows. Consider a Hamiltonian system of the form

$$\frac{\partial \mathbf{u}}{\partial t} = JE'(\mathbf{u}), \tag{1.1}$$

where $\mathbf{u}(x,t) \in X$ is an n-dimensional vector which depends on the space coordinate x and on time t, X is a real Hilbert space with inner product denoted by $(\cdot\,,\cdot)$, J is an invertible $n \times n$ skew-symmetric matrix, E is a functional of $\mathbf{u}$, and E' is its Fréchet derivative. Assume that this Hamiltonian system is invariant under a one-parameter group T of unitary transformations on X which commute with J, and that there exists a one-parameter family of solutions $\mathbf{u}(x,t)$ to (1.1) which can be written as

$$\mathbf{u}(x,t) = T(st)\ \mathbf{u}_s(x), \tag{1.2}$$

where $\mathbf{u}_s(x)$ only depends on space and is parametrized by the one-dimensional parameter s. In order to study the linear stability of $\mathbf{u}(x,t)$, one first linearizes (1.1) about this one-parameter family of solutions. As shown in Appendix A, the linearized system reads

$$\frac{\partial \mathbf{w}}{\partial t} = J\, H_s\, \mathbf{w}, \tag{1.3}$$

where the perturbation to $\mathbf{u}(x,t)$ is $T(st)\mathbf{w}(x,t)$, H_s is the self-adjoint operator given by

$$H_s = E''(\mathbf{u}_s) - s\, Q''(\mathbf{u}_s),$$

and Q is the conserved quantity associated with the invariance T. It can be found from T by means of an infinite-dimensional version of Noether's theorem [11, 12, 16], and is given by $Q(\mathbf{u}) = \frac{1}{2}\,\langle \mathcal{B}\mathbf{u}\,, \mathbf{u}\rangle$, where $\langle \mathbf{u}^*, \mathbf{v}\rangle \equiv (\mathbf{u}, \mathbf{v})$ for all $\mathbf{u}$, $\mathbf{v} \in X$ (X^* is the dual of X and X is identified with X^{**}), and the linear operator $\mathcal{B}$ is such that $J\,\mathcal{B}$ is an extension of $T'(0)$.

The method described in [11, 12] takes advantage of the fact that H_s is self-adjoint and therefore relatively simple to analyze, to obtain information on the spectrum of JH_s, which is the linearization of (1.1) about the coherent structure (1.2). More precisely, it is shown in [11] that if H_s has exactly one negative eigenvalue, the convexity requirement

$$\frac{d^2}{ds^2}\, d(s) > 0,$$

where the scalar function $d(s)$ is given by

$$d(s) = E(\mathbf{u}_s) - s\, Q(\mathbf{u}_s),$$

is a necessary and sufficient condition for the stability of (1.2), provided the continuous spectrum of H_s is positive and bounded away from the origin. By stability, we mean (nonlinear) orbital stability: by starting close enough to a solution (1.2), $\mathbf{u}_0(x,t)$, with say $s = s_0$, one can guarantee that at each time the system will remain close to the orbit of $\mathbf{u}_0$ under changes in the parameter s; i.e., at each time t the solution will

be arbitrarily close to some $T(st)\mathbf{u}_s(x)$ where s may depend on t and is of course not necessarily equal to s_0.

This approach can be extended [12] to the case where the symmetry group of the Hamiltonian system is m-dimensional ($1 < m < \infty$). Then, H_s and $d(s)$ are replaced by

$$H_{\mathbf{s}} = E''(\mathbf{u}_{\mathbf{s}}) - \sum_{i=1}^{m} s_i\, Q_i''(\mathbf{u}_{\mathbf{s}}), \qquad d(\mathbf{s}) = E(\mathbf{u}_{\mathbf{s}}) - \sum_{i=1}^{m} s_i\, Q_i(\mathbf{u}_{\mathbf{s}}),$$

where $\{Q_i\}_{i=1,\dots,m}$ is the set of conserved quantities associated with the unitary representation T of the m-dimensional symmetry group, and $\mathbf{s}$ is the vector with components s_i. In this case, it is shown in [12] that if $H_{\mathbf{s}}$ has an appropriate spectral decomposition, if the scalar function $d(\mathbf{s})$ is nondegenerate at $\mathbf{s}$, and if the number of negative eigenvalues of $H_{\mathbf{s}}$ is equal to the number of positive eigenvalues of the Hessian of $d(\mathbf{s})$ at $\mathbf{s}$, then the coherent structure is stable.

This paper is organized as follows. In section 2, we introduce the coupled nonlinear Klein–Gordon equations as well as the family of pulses, the stability of which we want to analyze. In section 3, we rewrite the Klein–Gordon equations in their Hamiltonian form, define the symmetries of this system and the associated conserved quantities, as well as $H_{\mathbf{s}}$. We then show that the self-adjoint operator $H_{\mathbf{s}}$ has exactly one negative eigenvalue, and that its continuous spectrum is positive but touches the origin. As a consequence, one of the hypotheses of [12], which says that $H_{\mathbf{s}}$ has a closed positive subspace, is not satisfied. This implies that it is difficult to prove orbital stability in X. Indeed, the continuous spectrum of JH_s is not bounded away from the origin and in such a case one typically resorts to the introduction of weighted spaces in order to obtain nonlinear stability results. The goal of this paper is to consider spectral stability instead. In section 4, we point out that only nonnegativity of the bilinear form $\langle H_{\mathbf{s}}\,\mathbf{u}, \mathbf{u}\rangle$ on the closure of the positive subspace of $H_{\mathbf{s}}$ is in fact needed to obtain spectral stability. We then show that, under this less restrictive assumption, the results of [12] can be extended to give a necessary and sufficient condition for the nonexistence of positive point spectrum of the linearization $J\,H_s$ about the two-parameter family of pulses. Since the continuous spectrum of $J\,H_s$ is on the imaginary axis, this is equivalent to the spectral stability of these solutions. Finally, we derive an explicit expression for this stability condition, and we show that it is consistent with the criterion found in [14] in the case of nonrotating pulses. In section 5, we summarize our results and mention possible extensions of this work to more complex models for the dynamics of elastic filaments.

2. The coupled nonlinear Klein–Gordon equations. We consider the following dimensionless coupled nonlinear Klein–Gordon equations, which describe the near-threshold dynamics of an elastic rod with circular cross section, subject to constant twist:

$$\begin{aligned} &\frac{\partial^2 A}{\partial t^2} - c_0^2 \frac{\partial^2 A}{\partial x^2} = \mu A - A|A|^2 + A\frac{\partial B}{\partial x}, \\ &\frac{\partial^2 B}{\partial t^2} - \frac{\partial^2 B}{\partial x^2} = -\frac{\partial |A|^2}{\partial x}, \end{aligned} \tag{2.1}$$

where $A(x,t)$ denotes the (scaled) slowly varying complex amplitude of the helical mode which grows above the bifurcation threshold, $B(x,t)$ is the (scaled) real axial

twist, c_0 is the group velocity of amplitude deformations, relative to the group velocity of twist deformations, and μ measures the distance from the threshold above which filaments subject to increasing constant twist tend to assume a helical shape [13]. Both c_0 and μ are real parameters. We are interested in pulse solutions [15, 17] of (2.1) of the form

$$A = a_0(\xi)e^{i\omega t}, \qquad B = b_0(\xi), \qquad \xi = x - ct, \tag{2.2}$$

where

$$a_0(\xi) = \alpha\ \mathrm{sech}(\beta\xi)\exp\left(i\frac{\omega c}{c^2 - c_0^2}\,\xi\right), \qquad b_0(\xi) = \frac{\alpha^2}{\beta(1-c^2)}\tanh(\beta\xi), \tag{2.3}$$

and

$$\alpha^2 = \frac{2\beta^2}{c^2}(c^2-1)(c^2-c_0^2), \qquad \beta^2 = \frac{\mu(c^2-c_0^2) - \omega^2 c_0^2}{(c^2-c_0^2)^2}. \tag{2.4}$$

As in [14], we only consider the case $\mu < 0$, since the continuous spectrum of the linearization of (2.1) about any pulse solution (2.2–2.4) would otherwise intersect the right half complex plane. As a consequence, α^2 and β^2 in (2.4) are positive if and only if

$$c^2 < c_0^2\left(1 + \frac{\omega^2}{\mu}\right) \ \text{ and } \ c^2 - 1 < 0. \tag{2.5}$$

We also note for later reference that a_0 and b_0 in (2.3) satisfy the following ordinary differential equations in ξ:

$$\begin{aligned}(c^2 - c_0^2)\,\frac{d^2a_0}{d\xi^2} - 2\,i\,c\,\omega\frac{da_0}{d\xi} &= a_0\left(\omega^2 + \mu - |a_0|^2 + \frac{db_0}{d\xi}\right),\\ (c^2-1)\frac{d^2b_0}{d\xi^2} &= -\frac{d}{d\xi}(|a_0|^2).\end{aligned} \tag{2.6}$$

As indicated in [15], system (2.1) may be written in Hamiltonian form. To this end, we first introduce the real variables P, Q, R, S, and U such that

$$A = P + iQ, \qquad A_t = R + iS, \qquad U = B_t. \tag{2.7}$$

Then (2.1) reads

$$\frac{\partial\,\mathbf{v}}{\partial t} = J\ E'(\mathbf{v}), \tag{2.8}$$

where $\mathbf{v} = [R, S, U, P, Q, B]^T$, the energy E is given by

$$E(\mathbf{v}) = \int_{-\infty}^{\infty} h\,\mathrm{d}x, \tag{2.9}$$

where

$$\begin{aligned}h = c_0^2\left(P_x^2 + Q_x^2\right) - \mu\left(P^2+Q^2\right) + \frac{1}{2}\left(P^2+Q^2\right)^2 - \left(P^2+Q^2\right)B_x\\ +\ R^2 + S^2 + \frac{1}{2}B_x^2 + \frac{1}{2}U^2,\end{aligned} \tag{2.10}$$

and the invertible, skew-symmetric matrix J reads

$$J = \begin{pmatrix} 0 & 0 & 0 & -1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1/2 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \\ 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}. \tag{2.11}$$

3. Hamiltonian formalism. As indicated in the introduction, we first look for the group of unitary transformations which leave the Hamiltonian system invariant [18] and define the corresponding conserved quantities. Let $\mathbf{v}_0 = [r_0, s_0, u_0, p_0, q_0, b_0]^T$ be the six-dimensional vector corresponding to the solution of (2.6), with a_0 and b_0 given in (2.3) and

$$a_0 = p_0 + i\,q_0, \quad r_0 = -c\,p_0' - \omega\,q_0, \quad s_0 = -c\,q_0' + \omega\,p_0, \quad u_0 = -c\,b_0'. \tag{3.1}$$

We require perturbations of $\mathbf{v}_0$ to be in $X = L^2 \times L^2 \times L^2 \times H^1 \times H^1 \times H^1$, even though the last component b_0 of $\mathbf{v}_0$ is nondecaying at infinity.

3.1. Symmetries. There are two unitary transformations that preserve the Hamiltonian structure (2.8) and which are elements of the symmetry group of (2.1) [14]. They are

1. The translational invariance, denoted by T_1 and such that

$$T_1(x_0)\,A(x,t) = A(x - x_0, t), \quad T_1(x_0)\,B(x,t) = B(x - x_0, t), \tag{3.2}$$

   where x_0 is a real arbitrary constant.
2. The gauge invariance, T_2, such that

$$T_2(\theta)\,A(x,t) = A(x,t)\,e^{i\theta}, \quad T_2(\theta)\,B(x,t) = B(x,t), \tag{3.3}$$

   where θ is a real arbitrary constant.

Using these definitions, the pulse solution given in (2.2) and (2.3) can be written as a function of x and t whose temporal dependence is generated by the action of these transformations. In other words, any pulse solution of (2.1) reads

$$[A(x,t), B(x,t)] = T_1(ct) \circ T_2(\omega t)\,[a_0(x), b_0(x)], \tag{3.4}$$

where a_0 and b_0 are defined in (2.3) and depend on c and ω.

The conserved quantities corresponding to the symmetries T_1 in (3.2) and T_2 in (3.3) are, respectively,

$$Q_1(\mathbf{v}) = -\int_{-\infty}^{\infty} \left(2\,(RP_x + SQ_x) + UB_x\right) dx, \tag{3.5}$$

$$Q_2(\mathbf{v}) = 2\int_{-\infty}^{\infty} (SP - QR)\,dx, \tag{3.6}$$

and any pulse solution defined in (2.2) and (2.3) is a critical point of the functional

$$I(\mathbf{v}) = E(\mathbf{v}) - c\,Q_1(\mathbf{v}) - \omega\,Q_2(\mathbf{v}). \tag{3.7}$$

Indeed, the system of equations given by

$$I'(\mathbf{v}) = \begin{pmatrix} 2(R + cP_x + \omega Q) \\ 2(S + cQ_x - \omega P) \\ U + cB_x \\ 2(-c_0^2 P_{xx} + P(-\mu + P^2 + Q^2 - B_x) - cR_x - \omega S) \\ 2(-c_0^2 Q_{xx} + Q(-\mu + P^2 + Q^2 - B_x) - cS_x + \omega R) \\ -B_{xx} + (P^2 + Q^2)_x - cU_x \end{pmatrix} = 0 \tag{3.8}$$

is equivalent to (2.6) with the first equation split into its real and imaginary parts. Equation (3.8) can be rewritten as

$$E'(\mathbf{v}) = Q'(\mathbf{v})\ \kappa, \tag{3.9}$$

where $\kappa = [c, \omega]^T$ and $Q'(\mathbf{v})$ is the 6×2 matrix $[Q_1'(\mathbf{v})\ Q_2'(\mathbf{v})]$. By taking the directional derivative of (3.9) in the direction $\sigma = [c_1, \omega_1]^T$ and evaluating it at $\mathbf{v} = \mathbf{v_0}$, we obtain

$$I''(\mathbf{v_0})\ \partial_\sigma \mathbf{v_0} = Q'(\mathbf{v_0})\ \sigma, \tag{3.10}$$

which is equation (3.3) of [12]. Here $\partial_\sigma \mathbf{v_0}$ denotes the directional derivative of $\mathbf{v_0}$, which depends on κ, in the direction σ. In what follows, we denote $I''(\mathbf{v_0})$ by $H_{c,\omega}$, and the linearization of (2.1) about the pulse solution $\mathbf{v_0}$ reads

$$\frac{\partial \mathbf{w}}{\partial t} = J\, H_{c,\omega}\ \mathbf{w},$$

where the perturbation to $[A, B]$ is $[(p(\xi, t) + iq(\xi, t)) \exp(i\omega t), b(\xi, t)]$ and where $\mathbf{w} = [r, s, u, p, q, b]^T$ is such that

$$r = p_t - c\, p_\xi - \omega\, q, \qquad s = q_t - c\, q_\xi + \omega\, p, \qquad u = b_t - c\, b_\xi.$$

Below, we show that the self-adjoint operator $H_{c,\omega}$ has a two-dimensional kernel, has only one negative eigenvalue, and is such that its continuous spectrum is nonnegative.

3.2. Kernel of $H_{c,\omega}$. In [14], it is shown that the kernel of $J\, H_{c,\omega}$ is two-dimensional and generated by

$$-T_1'(0)\mathbf{v}_0 = \mathbf{v}_0' \qquad \text{and} \qquad T_2'(0)\mathbf{v}_0 = \begin{pmatrix} -s_0 \\ r_0 \\ 0 \\ -q_0 \\ p_0 \\ 0 \end{pmatrix}, \tag{3.11}$$

where T_1 and T_2 are the Hamiltonian symmetries defined in (3.2) and (3.3) and $\mathbf{v}_0$ is the six-dimensional pulse solution defined at the beginning of section 3. Since J is invertible, $\mathrm{Ker}(J\, H_{c,\omega}) = \mathrm{Ker}(H_{c,\omega})$.

3.3. Negative subspace of $H_{c,\omega}$. In what follows, we show that the dimension of the negative subspace of $H_{c,\omega}$ is equal to 1. To do so, we first rewrite the quadratic form $\langle H_{c,\omega}\mathbf{w},\mathbf{w}\rangle$, where $\mathbf{w}\in X$, as a sum of squares plus terms of the form $\langle\tilde{\mathcal{L}}_i\, z_i, z_i\rangle$, $i=1,2$, where the $\tilde{\mathcal{L}}_i$ are second-order differential operators. We then find the number of negative eigenvalues of these two operators, using results from Sturm–Liouville theory. Finally, we use this information together with the min-max principle, to find the sign of the two lowest eigenvalues of $H_{c,\omega}$.

The linear operator $H_{c,\omega}$ reads

$$H_{c,\omega}\equiv I''(\mathbf{v}_0)=E''(\mathbf{v}_0)-c\,Q_1''(\mathbf{v}_0)-\omega\,Q_2''(\mathbf{v}_0)=\begin{pmatrix} D^2 & C\\ C^* & L\end{pmatrix}, \tag{3.12}$$

where

$$D^2=\begin{pmatrix}2&0&0\\0&2&0\\0&0&1\end{pmatrix},\quad C=\begin{pmatrix}2c\partial_x & 2\omega & 0\\ -2\omega & 2c\partial_x & 0\\ 0&0&c\partial_x\end{pmatrix},$$

$$C^*=\begin{pmatrix}-2c\partial_x & -2\omega & 0\\ 2\omega & -2c\partial_x & 0\\ 0&0&-c\partial_x\end{pmatrix}=-C, \tag{3.13}$$

and

$$L=\begin{pmatrix} 2\left(-c_0^2\,\partial_{xx}-\partial_x\,b_0-\mu+q_0^2+3p_0^2\right) & 4p_0\,q_0 & -2p_0\,\partial_x\\ 4p_0\,q_0 & 2\left(-c_0^2\partial_{xx}-\partial_x\,b_0-\mu+p_0^2+3q_0^2\right) & -2q_0\,\partial_x\\ 2\left(p_0\,\partial_x+\partial_x\,p_0\right) & 2\left(q_0\,\partial_x+\partial_x\,q_0\right) & -\partial_{xx}\end{pmatrix}. \tag{3.14}$$

Here ∂_x and ∂_{xx} refer to first and second partial derivatives with respect to x. We denote a six-dimensional vector in X by $\mathbf{w}=[w_1,w_2]^T$, where $i=1,2$ and the w_i are three-dimensional. Then,

$$\begin{aligned}\langle H_{c,\omega}\mathbf{w},\mathbf{w}\rangle &= \langle D^2\,w_1,w_1\rangle+\langle C\,w_2,w_1\rangle+\langle C^*\,w_1,w_2\rangle+\langle L\,w_2,w_2\rangle\\ &= (D\,w_1,D\,w_1)+2\langle C\,w_2,w_1\rangle+\langle L\,w_2,w_2\rangle\\ &= \left|\left|D\,w_1+D^{-1}C\,w_2\right|\right|^2+\langle(L+C^2D^{-2})\,w_2,w_2\rangle. \end{aligned}\tag{3.15}$$

The linear operator $L_1=L+C^2D^{-2}$ is such that

$$L_1=\begin{pmatrix} & L_2 & \begin{matrix}-2p_0\,\partial_x\\ -2q_0\,\partial_x\end{matrix}\\ 2\left(p_0\,\partial_x+\partial_x\,p_0\right) & 2\left(q_0\,\partial_x+\partial_x\,q_0\right) & -(1-c^2)\,\partial_{xx}\end{pmatrix}, \tag{3.16}$$

where

$$L_2=\begin{pmatrix} 2\left((c^2-c_0^2)\partial_{xx}-\partial_x\,b_0-\mu-\omega^2+q_0^2+3p_0^2\right) & 4\left(p_0\,q_0+c\,\omega\,\partial_x\right)\\ 4\left(p_0\,q_0-c\,\omega\,\partial_x\right) & 2\left((c^2-c_0^2)\partial_{xx}-\partial_x\,b_0-\mu-\omega^2+p_0^2+3q_0^2\right)\end{pmatrix}. \tag{3.17}$$

With $\mathbf{y} = [y_1, y_2, y_3]^T$, we have

$$\begin{aligned}\langle L_1\,\mathbf{y},\mathbf{y}\rangle &= \left\langle L_2 \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \right\rangle \\ &\quad + (1-c^2)\,(\partial_x y_3, \partial_x y_3) + 2\sqrt{1-c^2}\,\left(K^T\,\partial_x y_3, \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}\right) \\ &= \left\langle \left(L_2 - K^T K\right) \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \right\rangle + \left\| \sqrt{1-c^2}\,\partial_x y_3 + K \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \right\|^2,\end{aligned} \tag{3.18}$$

where

$$K = -\frac{2}{\sqrt{1-c^2}}\,(p_0, q_0)\,.$$

Finally, let

$$\mathcal{L} \equiv L_2 - K^T K \tag{3.19}$$

and introduce the change of variable

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} \cos\left(\frac{\omega c x}{c^2-c_0^2}\right) & \sin\left(\frac{\omega c x}{c^2-c_0^2}\right) \\ -\sin\left(\frac{\omega c x}{c^2-c_0^2}\right) & \cos\left(\frac{\omega c x}{c^2-c_0^2}\right) \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \equiv M \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}. \tag{3.20}$$

Then

$$\left\langle \mathcal{L} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \right\rangle = \left\langle \tilde{\mathcal{L}} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}, \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \right\rangle, \tag{3.21}$$

where

$$\tilde{\mathcal{L}} = \begin{pmatrix} \tilde{\mathcal{L}}_1 & 0 \\ 0 & \tilde{\mathcal{L}}_2 \end{pmatrix} \tag{3.22}$$

and

$$\begin{aligned}\tilde{\mathcal{L}}_1 &= \tilde{\mathcal{L}}_2 - \frac{4c^2}{1-c^2}(p_0^2 + q_0^2), \\ \tilde{\mathcal{L}}_2 &= 2\left((c^2 - c_0^2)\,\partial_{xx} + \frac{\omega^2\,c_0^2}{c^2-c_0^2} + p_0^2 + q_0^2 - \partial_x\,b_0 - \mu\right).\end{aligned} \tag{3.23}$$

The operator $\tilde{\mathcal{L}}_2$ has 0 as an eigenvalue since, from (2.6), one can show that

$$\tilde{\mathcal{L}}_2\,|a_0| = 0, \tag{3.24}$$

where a_0 is given in (2.3). Since the pulse has no zero, Sturm–Liouville theory (see, for instance, [19, p. 104]) indicates that $\tilde{\mathcal{L}}_2$ does not have any negative eigenvalue. By taking the derivative of (3.24) with respect to x we obtain

$$\tilde{\mathcal{L}}_1\,|a_0|' = 0. \tag{3.25}$$

Since $|a_0|'$ has exactly one zero, $\tilde{\mathcal{L}}_1$ has one negative eigenvalue.

We now use the above information together with the min-max principle (see for instance, Theorem XIII.1, page 76 [20]; see pages 87–90 [21], for a discussion in the finite-dimensional case), to prove that $H_{c,\omega}$ has exactly one negative eigenvalue. We have shown that

$$\begin{aligned}\langle H_{c,\omega}\,\mathbf{w},\mathbf{w}\rangle = \left|\left|D\,w_1 + D^{-1}C\,w_2\right|\right|^2 + \left|\left|\sqrt{1-c^2}\,\partial_x y_3 + K\begin{pmatrix} y_1 \\ y_2\end{pmatrix}\right|\right|^2 \\ +\langle\tilde{\mathcal{L}}_1\,z_1,z_1\rangle + \langle\tilde{\mathcal{L}}_2\,z_2,z_2\rangle,\end{aligned}$$

where $\mathbf{w} = [w_1,w_2]^T$, $w_2 = [y_1,y_2,y_3]$, $[z_1,z_2]^T = M\,[y_1,y_2]^T$, and M is defined in (3.20). Let $\tilde{\mathbf{w}} = [\tilde{w}_1,\tilde{w}_2]^T$, $||\tilde{\mathbf{w}}|| = 1$, be such that

$$\begin{aligned}\tilde{w}_1 &= -D^{-2}C\,\tilde{w}_2, \quad \tilde{w}_2 = [\tilde{y}_1,\tilde{y}_2,\tilde{y}_3], \\ \partial_x\,\tilde{y}_3 &= \frac{-1}{\sqrt{1-c^2}}\,K\begin{pmatrix}\tilde{y}_1\\ \tilde{y}_2\end{pmatrix}, \quad \begin{pmatrix}\tilde{y}_1\\ \tilde{y}_2\end{pmatrix} = M^{-1}\begin{pmatrix}\tilde{z}_1\\ 0\end{pmatrix},\end{aligned}$$

with $\tilde{z}_1$ chosen to be an eigenvector of $\tilde{\mathcal{L}}_1$ with eigenvalue $\kappa < 0$. Note that the $\tilde{z}_2$ variable associated with $\tilde{\mathbf{w}}$ is zero. By the min-max principle, the smallest eigenvalue $\mu_1(H_{c,\omega})$ of $H_{c,\omega}$ is such that

$$\mu_1(H_{c,\omega}) = \inf_{\mathbf{w}\in X,||\mathbf{w}||=1}\langle H_{c,\omega}\,\mathbf{w},\mathbf{w}\rangle.$$

Thus,

$$\mu_1(H_{c,\omega}) \le \langle H_{c,\omega}\,\tilde{\mathbf{w}},\tilde{\mathbf{w}}\rangle = \langle\tilde{\mathcal{L}}_1\tilde{z}_1,\tilde{z}_1\rangle = \kappa\langle\tilde{z}_1,\tilde{z}_1\rangle < 0;$$

i.e., $H_{c,\omega}$ has at least one negative eigenvalue. Similarly, the second smallest eigenvalue $\mu_2(H_{c,\omega})$ of $H_{c,\omega}$ is given by

$$\mu_2(H_{c,\omega}) = \sup_{\phi_1\in X}\ \inf_{\mathbf{w}\in X,||\mathbf{w}||=1,\mathbf{w}\perp\phi_1}\langle H_{c,\omega}\,\mathbf{w},\mathbf{w}\rangle.$$

Choose $\tilde{\phi}_1 \neq \mathbf{0}$ such that $\tilde{\phi}_1 = [0,0,0,\tilde{\varphi}_1,\tilde{\varphi}_2,0]^T$ with $[\tilde{\varphi}_1,\tilde{\varphi}_2]^T = M^{-1}\,[1,0]^T \perp M^{-1}\,[0,1]^T$. Then, $\mathbf{w}\perp\tilde{\phi}_1$ is equivalent to having the z_1 variable associated with $\mathbf{w}$ equal to zero, the other variables being arbitrary. Thus,

$$\begin{aligned}&\mu_2(H_{c,\omega}) \\ &\ge \inf_{\mathbf{w}\in X,||\mathbf{w}||=1,\mathbf{w}\perp\tilde{\phi}_1}\langle H_{c,\omega}\,\mathbf{w},\mathbf{w}\rangle \\ &= \inf_{\substack{\mathbf{w}\in X,\\ ||\mathbf{w}||=1,\\ z_1=0}}\left[\left|\left|D\,w_1 + D^{-1}C\,w_2\right|\right|^2 + \left|\left|\sqrt{1-c^2}\,\partial_x y_3 + K\begin{pmatrix}y_1\\ y_2\end{pmatrix}\right|\right|^2 + \langle\tilde{\mathcal{L}}_2\,z_2,z_2\rangle\right] \\ &\ge \inf_{\mathbf{w}\in X,||\mathbf{w}||=1,z_1=0}\langle\tilde{\mathcal{L}}_2\,z_2,z_2\rangle = 0,\end{aligned}$$

the last equality being a consequence of the fact that $\tilde{\mathcal{L}}_2$ has the bottom of its point spectrum at zero. Thus, $H_{c,\omega}$ has at most one negative eigenvalue. Since it also has at least one negative eigenvalue, $H_{c,\omega}$ has exactly one negative eigenvalue.

3.4. Continuous spectrum of $H_{c,\omega}$. To find the continuous spectrum of $H_{c,\omega}$, we define

$$H^{\infty}_{c,\omega} \equiv \lim_{x\to\infty} H_{c,\omega} = \begin{pmatrix} D^2 & C \\ C^* & L^{\infty} \end{pmatrix}, \tag{3.26}$$

where C and D are given in (3.13) and

$$L^{\infty} = \begin{pmatrix} 2\left(-c_0^2\,\partial_{xx} - \mu\right) & 0 & 0 \\ 0 & 2\left(-c_0^2\partial_{xx} - \mu\right) & 0 \\ 0 & 0 & -\partial_{xx} \end{pmatrix}. \tag{3.27}$$

The continuous spectrum of $H_{c,\omega}$ is then given [22, Theorem A.2, p. 140] by the set

$$S = \left\{\lambda \in \mathbb{C} \,:\, \det\left(H^{\infty}_{c,\omega}(k) - \lambda\, I_6\right) = 0 \ \text{ for some } \ k \in \mathbb{R}\right\}, \tag{3.28}$$

where $H^{\infty}_{c,\omega}(k)$ is obtained from $H^{\infty}_{c,\omega}$ by replacing ∂_x with $i\,k$. The six eigenvalues of $H^{\infty}_{c,\omega}(k)$ are

$$\begin{aligned} \lambda_1^{\pm}(k) &= \frac{1}{2}\left(1 + k^2 \pm \sqrt{(1-k^2) + 4k^2c^2}\right), \\ \lambda_2^{\pm}(k) &= -\mu + c_0^2k^2 + 1 \pm \sqrt{(\mu + 1 - c_0^2k^2)^2 + 4(\omega - kc)^2}, \\ \lambda_3^{\pm}(k) &= -\mu + c_0^2k^2 + 1 \pm \sqrt{(\mu + 1 - c_0^2k^2)^2 + 4(\omega + kc)^2}. \end{aligned}$$

We now show that $\lambda_1^{\pm}(k) \geq 0$ and that $\lambda_{2,3}^{\pm}(k) > 0$ for all k's. First, the condition $\lambda_1^{\pm}(k) \geq 0$ is equivalent to $4k^2(1-c^2) \geq 0$, which is always satisfied since $c^2 < 1$, by the second inequality of (2.5). Since $\lambda_1^{-}(0) = 0$, the set $\{\lambda_1^{\pm}(k),\ k \in \mathbb{R}\}$ is in fact equal to $[0, +\infty)$. Second, the condition $\lambda_2^{\pm}(k) > 0$ is equivalent to $P(k) \equiv \omega^2 - 2\,k\,c\,\omega + \mu + (c^2 - c_0^2)\,k^2 < 0$. To see that this condition is satisfied for all k's, we proceed as follows. Note that as $k \to \pm\infty$, $P(k)$ is negative since $c^2 < c_0^2$ by the first inequality of (2.5). The derivative of P vanishes at $k = k_0 \equiv c\,\omega/(c^2 - c_0^2)$, and P reaches its maximum at that point. Substitution of the expression of k_0 into the formula for $P(k)$ gives

$$P(k_0) = \frac{\mu}{c^2 - c_0^2}\left[c^2 - c_0^2\left(1 + \frac{\omega^2}{\mu}\right)\right],$$

which is strictly negative since $\mu < 0$ by hypothesis and since $c^2 - c_0^2(1 + \omega^2/\mu) < 0$, whence $c^2 - c_0^2 < 0$, by the first inequality of (2.5). Thus, $\lambda_2^{\pm}(k) > 0$ for all k's. Finally, since the above argument does not depend on the sign of ω and since $\lambda_3^{\pm}(k)$ can be obtained from $\lambda_2^{\pm}(k)$ by changing ω into $-\omega$, we also have that $\lambda_3^{\pm}(k) > 0$ for all k's. Therefore, S is the whole nonnegative real axis.

A similar calculation shows that the continuous spectrum of $J\,H_{c,\omega}$ is the whole imaginary axis, and therefore contains the origin. This was to be expected since the symmetry $B \to B + \text{constant}$ of (2.1) indicates that the origin is in the continuous spectrum of the linearization of (2.1) about any solution whose B-component converges to a constant as ξ goes to infinity. Given this, it is also not surprising that the continuous spectrum of $H_{c,\omega}$ touches the origin, as shown above.

An important consequence of this discussion is that the stability theorems of [12] are not directly applicable, since the hypothesis that the positive subspace of $H_{c,\omega}$

be closed is not satisfied. As already mentioned in the introduction, the presence of continuous spectrum of $H_{c,\omega}$ all the way to the origin—or, as exemplified by the above calculations, the fact that the continuous spectrum of $J\,H_{c,\omega}$ includes the origin—typically rules out the possibility of having nonlinear stability results for perturbations $\mathbf{w} \in X$. In the next section, we thus restrict ourselves to the question of the spectral stability of pulse solutions. We show that the theorems of [12] can be adapted to the present situation, and obtain a necessary and sufficient condition for the spectral stability of the family of pulses.

4. Spectral stability criterion. Consider the scalar function of c and ω given by

$$d(c,\omega) = E(\mathbf{v}_0) - c\,Q_1(\mathbf{v}_0) - \omega\,Q_2(\mathbf{v}_0), \tag{4.1}$$

where E, Q_1, and Q_2 are defined in (2.9), (3.5), and (3.6) and $\mathbf{v}_0 = [r_0, s_0, u_0, p_0, q_0, b_0]^T$ is, as before, the six-dimensional solution of (3.8) defined by (3.1) with a_0 and b_0 given in (2.3). Note that d defined in (4.1) depends on c and ω both explicitly and implicitly through the dependence of $\mathbf{v}_0$ on c and ω.

Even though Assumption 3 of [12] is not satisfied here since the positive subspace of $H_{c,\omega}$ is not closed (in other words, since the continuous spectrum of $H_{c,\omega}$ is not bounded away from zero), the following theorems carry through.

THEOREM 1 (from Theorem 3.1 of [12]). *Let* $X_1 = \{\mathbf{u} \in X | \langle Q_i'(\mathbf{v_0}), \mathbf{u}\rangle = 0,\ i = 1, 2\}$, *and* Π_1 *be the orthogonal projection of* X *onto* X_1. *If* d *is nondegenerate, the reduced Hamiltonian* $H_1 = \Pi_1^*\,H_{c,\omega}\,\Pi_1$ *has the negative index (i.e., a negative subspace of dimension)*

$$n(H_1) = n(H_{c,\omega}) - p\,(d''), \tag{4.2}$$

where $p\,(d'')$ *is the number of positive eigenvalues of the Hessian* d'' *of* d, *and* $n(H_{c,\omega})$ *is the dimension of the negative subspace of* $H_{c,\omega}$.

In our case, $n(H_{c,\omega}) = 1$, and (4.2) therefore implies that

$$p\,(d'') \leq 1. \tag{4.3}$$

THEOREM 2 (from Theorem 5.1 of [12]). *Let* d'' *be nonsingular,* $n(H_{c,\omega}) - p\,(d'')$ *be odd, and* X *be separable. Then* $J\,H_{c,\omega}$ *has at least one pair of real nonzero eigenvalues.*

Note that only nonnegativity of the bilinear form $\langle H_{c,\omega}\,\mathbf{u}, \mathbf{u}\rangle$ on the closure of the positive subspace of $H_{c,\omega}$ is needed in the proof of this theorem given in [12].

Therefore, if

$$\det\,(d'') = d_{cc}\,d_{\omega\omega} - d_{c\omega}^2 > 0, \tag{4.4}$$

the number of positive eigenvalues of d'' is even, and by (4.3), it is equal to zero. Equation (4.2) of Theorem 1 then implies that $n(H_1) = n(H_{c,\omega}) = 1$, and by Theorem 2, the linearization $J\,H_{c,\omega}$ has at least one pair $(\lambda, -\lambda)$, $\lambda \in \mathbb{R}$, of real nonzero eigenvalues. Thus, the linearization of (2.1) about a pulse solution has positive point spectrum, and the pulse solution is dynamically unstable (section VI of [12]). Condition (4.4) reads

$$\begin{aligned} &-12\,c^2{c_0}^4\left(c_0^2-1\right)\left(-3\,{c_0}^2+2\,c^2+c^4\right)\omega^4 \\ &-18\,\mu\,{c_0}^2\left(c^4-3\,c^2{c_0}^2+c^2+{c_0}^2\right)\left(c^2-c_0^2\right)^2\omega^2 \\ &+3\,\left(c^2-1\right)\mu^2\left(c^4{c_0}^2+2\,c^4-9\,c^2{c_0}^2+6\,{c_0}^4\right)\left(c^2-c_0^2\right)^2 > 0. \end{aligned} \tag{4.5}$$

In the case $\omega = 0$, we see that pulses are thus unstable if

$$T \equiv 3c_0^2\left(3c^2 - 2c_0^2\right) - c^4\left(c_0^2 + 2\right) > 0, \tag{4.6}$$

since $c^2 - 1 < 0$. This is exactly the condition found in [14] by means of Evans function techniques. As discussed in [14], the condition $T > 0$ defines a nonempty set of speeds c. There also exists values of c such that $T < 0$. For instance, the stable propagation of a pulse solution with $\omega = 0$, and for which $T = -27.078$, is illustrated in the numerical simulation of Figure 11 of [15] (or equivalently Figure 1 of [14]).

We now show that if $\det(d'') < 0$, $J\,H_{c,\omega}$ has no positive point spectrum. In this case, since the continuous spectrum of the linearization $J\,H_{c,\omega}$ is on the imaginary axis, this operator has no positive spectrum, and the corresponding pulse solution is therefore spectrally stable. First note that if $\det(d'') < 0$, then $p\,(d'') = 1$ and $n(H_1) = 0$ by Theorem 1 above. We now prove that $n(H_1) = 0$ implies that it is not possible for $J\,H_{c,\omega}$ to have any nonzero eigenvalue. Suppose that $\mathbf{y}$ is an eigenvector of $J\,H_{c,\omega}$ with nonzero eigenvalue a. Then $\mathbf{y} \in X_1$, where X_1 is defined in Theorem 1. Indeed, for $i = 1, 2$, we have

$$\begin{aligned}
\langle Q_i'(\mathbf{v_0}), \mathbf{y}\rangle &= \langle \mathcal{B}_i\,\mathbf{v_0}, \mathbf{y}\rangle \qquad \text{by (A.6)} \\
&= \langle J^{-1}\,T_i'(0)\,\mathbf{v_0}, \mathbf{y}\rangle \qquad \text{since } J\,\mathcal{B}_i \text{ is an extension of } T_i'(0) \\
&= \langle J^{-1}\,T_i'(0)\,\mathbf{v_0}, a^{-1}JH_{c,\omega}\,\mathbf{y}\rangle \\
&= -\langle a^{-1}H_{c,\omega}\,\mathbf{y}, T_i'(0)\,\mathbf{v_0}\rangle \qquad \text{since } J \text{ is skew} \\
&= -\langle H_{c,\omega}\,T_i'(0)\,\mathbf{v_0}, a^{-1}\,\mathbf{y}\rangle = 0,
\end{aligned}$$

since $T_i'(0)\,\mathbf{v_0}$ is in the kernel Z of $H_{c,\omega}$ by (A.11) with T replaced by T_i. Note that X_1 contains Z, since (see [12, p. 316]) if $\mathbf{z} \in Z$, then

$$\forall \sigma = [c_1, \omega_1]^T, \qquad 0 = \langle H_{c,\omega}\,\mathbf{z}, \partial_\sigma \mathbf{v_0}\rangle = \langle H_{c,\omega}\,\partial_\sigma \mathbf{v_0}, \mathbf{z}\rangle = \langle Q'(\mathbf{v_0})\;\sigma, \mathbf{z}\rangle,$$

where $Q'(\mathbf{v_0})$ is defined at the end of section 3.1, and we made use of (3.10). Then, let (see equation (5.1) of [12])

$$X_2 = \left\{\mathbf{u} \in X \middle| \langle Q_i'(\mathbf{v_0}), \mathbf{u}\rangle = 0 = (T_i'(0)\,\mathbf{v_0}, \mathbf{u}),\ i = 1, 2\right\}.$$

From the definition of X_2, one has the orthogonal decomposition $X_1 = X_2 + Z$, and $\mathbf{y} \in X_1$ can therefore be written as $\mathbf{y} = \mathbf{x_2} + \mathbf{z}$ with $\mathbf{x_2} \in X_2$ and $\mathbf{z} \in Z$. Then,

$$\langle H_{c,\omega}\,\mathbf{y}, \mathbf{y}\rangle = \langle a\,J^{-1}\,\mathbf{y}, \mathbf{y}\rangle = 0$$

since J^{-1} is skew. This implies that $\langle H_{c,\omega}\,\mathbf{x_2}, \mathbf{x_2}\rangle = \mathbf{0}$ since

$$\langle H_{c,\omega}\,\mathbf{y}, \mathbf{y}\rangle = \langle H_{c,\omega}\,\mathbf{x_2}, \mathbf{x_2}\rangle + \langle H_{c,\omega}\,\mathbf{x_2}, \mathbf{z}\rangle = \langle H_{c,\omega}\,\mathbf{x_2}, \mathbf{x_2}\rangle.$$

As a consequence, $\langle H_1\,\mathbf{x_2}, \mathbf{x_2}\rangle = 0$. Finally, since $n(H_1) = 0$, H_1 is a nonnegative self-adjoint operator on X_1. By the spectral decomposition theorem, $\langle H_1\,\mathbf{x_2}, \mathbf{x_2}\rangle = 0$ implies $\mathbf{x_2} = \mathbf{0}$, whence $\mathbf{y} \in Z = \operatorname{Ker} H_{c,\omega}$. This contradicts the hypothesis that $\mathbf{y}$ is an eigenvector of $J\,H_{c,\omega}$ with nonzero eigenvalue a. Therefore, we have the following theorem.

THEOREM 3. *Let d'' be nonsingular. Then, pulse solutions of* (2.1) *given by* (2.2) *and* (2.3) *are spectrally stable if and only if $n(H_1) = 0$, i.e., if and only if* $\det(d'') < 0$.

In other words, pulses are unstable if condition (4.5) is satisfied, and they are spectrally stable if the left-hand side of (4.5) is negative. More generally, if $n(H_1) = 0$, i.e., if $n(H_{c,\omega}) = p(d'')$, and if d'' is nonsingular, then pulse solutions are spectrally stable.

5. Conclusion. We have obtained a necessary and sufficient condition for the spectral stability of a family of pulse solutions to two coupled Klein–Gordon equations. This analysis, based on Hamiltonian methods [11, 12], confirms and extends the results of [14] obtained by means of Evans functions techniques for nonrotating pulses, and also applies to the case of rotating pulses. As discussed in the introduction and in sections 3 and 4, the theorems of [12] do not directly apply here because the continuous spectrum of the linearization of the coupled Klein–Gordon equations about the family of pulse solutions contains the origin. As a consequence, one cannot use these results to establish orbital stability of the solutions. However, we have shown that one can modify the results of [12] in order to obtain spectral stability. The system of partial differential equations considered here models the near-threshold dynamics of an elastic rod with circular cross section. More general envelope equations taking into account other physical properties of elastic filaments, such as extensibility, the existence of a tension mode, or the presence of a noncircular cross section, can be found in the literature (see for example [23]). All of these equations can be written in Hamiltonian form and it should be possible to extend the analysis presented in this paper to these more general models.

Appendix A. Linearization about a coherent structure. This discussion is entirely based on Section II of [11], and is presented here to make this paper self-contained. Consider the following Hamiltonian system:

$$\frac{\partial \mathbf{u}}{\partial t} = J E'(\mathbf{u}), \tag{A.1}$$

where $\mathbf{u} \in X$, X is a real Hilbert space with inner product denoted by $(\cdot\,,\cdot)$, $J : X^* \to X$ is invertible and skew-symmetric (X^* is the dual of X), E is a functional of $\mathbf{u}$, and $E' : X \to X^*$ is its Fréchet derivative. Assume that $\mathbf{U}_s(x,t) = T(st)\ \mathbf{u}_s(x)$ is a solution of this system, where $T \equiv \{T(s) : X \to X,\ s \in \mathbb{R}\}$, is a one-parameter group of unitary transformations, which has the following properties:

1. T commutes with J, i.e.,

$$T(s)\, J = J\, T^*(-s), \tag{A.2}$$

   where $T^*(s)$ is the adjoint of $T(s)$.
2. E is invariant under T, i.e.,

$$E\,(T(s)\mathbf{u}) = E(\mathbf{u}) \tag{A.3}$$

   for all $\mathbf{u} \in X$ and for all $s \in \mathbb{R}$.

By taking the Fréchet derivative of (A.3) with respect to $\mathbf{u}$, one obtains

$$E'(\mathbf{u}) = T^*(s)\, E'(T(s)\mathbf{u}). \tag{A.4}$$

The conserved quantity associated with the invariance T is given by

$$Q(\mathbf{u}) = \frac{1}{2}\langle \mathcal{B}\mathbf{u}\,, \mathbf{u}\rangle, \tag{A.5}$$

where $\langle \mathbf{u}^*, \mathbf{v}\rangle \equiv (\mathbf{u}, \mathbf{v})$ for all $\mathbf{u}$, $\mathbf{v} \in X$ (and X is identified with X^{**}), and the linear operator $\mathcal{B} : X \to X^*$ is such that $J\,\mathcal{B}$ is an extension of $T'(0) : X \to X$ [11, 12, 16]. By differentiating (A.5), one obtains

$$Q'(\mathbf{u}) = \mathcal{B}\,\mathbf{u} \qquad \text{and} \qquad \mathcal{B} = Q''(\mathbf{u}), \qquad \forall\ \mathbf{u} \in X. \tag{A.6}$$

Since T is a representation of a one-parameter group of operators, we have $T(s+\epsilon) = T(s)\,T(\epsilon),\ \forall\,(s,\epsilon) \in \mathbb{R}^2$, which implies

$$T'(s) = T'(0)\,T(s). \tag{A.7}$$

To see that Q is invariant under T, calculate

$$\begin{aligned}
\frac{d}{ds}\left[Q(T(s)\mathbf{u})\right] &\equiv \langle Q'(T(s)\,\mathbf{u}), T'(s)\,\mathbf{u}\rangle \\
&= \langle Q'(T(s)\,\mathbf{u}), T'(0)\,T(s)\,\mathbf{u}\rangle \qquad \text{by (A.7)} \\
&= \langle Q'(T(s)\,\mathbf{u}), J\,\mathcal{B}\,T(s)\,\mathbf{u}\rangle \qquad \text{since } J\,\mathcal{B} \text{ extends } T'(0) \\
&= \langle B\,T(s)\,\mathbf{u}, J\,\mathcal{B}\,T(s)\,\mathbf{u}\rangle \qquad \text{by (A.6)} \\
&= 0 \qquad \text{since } \langle \mathbf{u}^*, J\,\mathbf{u}^*\rangle = 0,\ \forall\,\mathbf{u}^* \in X^*.
\end{aligned}$$

Thus, $Q(T(s)\,\mathbf{u}) = Q(\mathbf{u})$ and by differentiation

$$T^*(s)\ Q'(T(s)\,\mathbf{u}) = Q'(\mathbf{u}), \tag{A.8}$$

which, together with the first equation of (A.6), implies

$$T^*(s)\,\mathcal{B}\,T(s) = \mathcal{B}. \tag{A.9}$$

To perform a linear stability analysis of the coherent structure, let

$$\mathbf{u}(x,t) = T(st)\left(\mathbf{u}_s(x) + \mathbf{w}(x,t)\right),$$

and substitute into (A.1). We have

$$\begin{aligned}
\frac{\partial \mathbf{u}}{\partial t} &= \frac{\partial}{\partial t}\left[T(st)\left(\mathbf{u}_s(x) + \mathbf{w}(x,t)\right)\right] = s\,T'(st)\left(\mathbf{u}_s(x) + \mathbf{w}(x,t)\right) + T(st)\frac{\partial \mathbf{w}}{\partial t} \\
&= s\,T'(0)\,T(st)\left(\mathbf{u}_s(x) + \mathbf{w}(x,t)\right) + T(st)\frac{\partial \mathbf{w}}{\partial t} \qquad \text{by (A.7)} \\
&= s\,J\,\mathcal{B}\,T(st)\left(\mathbf{u}_s(x) + \mathbf{w}(x,t)\right) + T(st)\frac{\partial \mathbf{w}}{\partial t} \qquad \text{since } J\mathcal{B} \text{ is an extension of } T'(0) \\
&= s\,J\,T^*(-st)\,\mathcal{B}\left(\mathbf{u}_s(x) + \mathbf{w}(x,t)\right) + T(st)\frac{\partial \mathbf{w}}{\partial t} \qquad \text{by } (A.9) \\
&= s\,J\,T^*(-st)\,Q'(\mathbf{u}_s) + s\,J\,T^*(-st)\,Q''(\mathbf{u}_s)\,\mathbf{w} + T(st)\frac{\partial \mathbf{w}}{\partial t} \qquad \text{by } (A.6).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
J\,E'(T(st)\left(\mathbf{u}_s + \mathbf{w}\right)) &= J\,T^*(-st)\,E'(\mathbf{u}_s + \mathbf{w}) \qquad \text{by (A.4)} \\
&= J\,T^*(-st)\,E'(\mathbf{u}_s) + J\,T^*(-st)\,E''(\mathbf{u}_s)\,\mathbf{w} + O\left(||\mathbf{w}||^2\right).
\end{aligned}$$

Since $T(st)\,\mathbf{u}_s$ is a solution of the Hamiltonian system, the terms which do not depend on $\mathbf{w}$ on each side of (A.1) written as

$$\frac{\partial}{\partial t}\left[T(st)\left(\mathbf{u}_s(x) + \mathbf{w}(x,t)\right)\right] = J\,E'(T(st)\left(\mathbf{u}_s + \mathbf{w}\right))$$

balance out, which implies

$$E'(\mathbf{u}_s) = s\,Q'(\mathbf{u}_s). \tag{A.10}$$

The rest of the equation reads, to first order in $\mathbf{w}$,

$$\begin{aligned} T(st)\frac{\partial \mathbf{w}}{\partial t} &= J\,T^*(-st)\,E''(\mathbf{u}_s)\,\mathbf{w} - s\,J\,T^*(-st)\,Q''(\mathbf{u}_s)\,\mathbf{w} \\ &= T(st)\,J\,(E''(\mathbf{u}_s) - s\,Q''(\mathbf{u}_s))\,\mathbf{w} \qquad \text{by (A.2).} \end{aligned}$$

By multiplying both sides of this equation by $T(-st)$, we obtain

$$\frac{\partial \mathbf{w}}{\partial t} = J\,(E''(\mathbf{u}_s) - s\,Q''(\mathbf{u}_s))\,\mathbf{w} \equiv J\,H_s\,\mathbf{w},$$

which is (1.3).

Finally, note that

$$E'(T(s)\,\mathbf{u}_s) - s\,Q'(T(s)\,\mathbf{u}_s) = T^*(-s)\,E'(\mathbf{u}_s) - T^*(-s)\,s\,Q'(\mathbf{u}_s) = 0$$

by (A.4), (A.8), and (A.10). Differentiation with respect to s at $s = 0$ gives

$$[E''(\mathbf{u}_s) - s\,Q''(\mathbf{u}_s)]\;T'(0)\,\mathbf{u}_s = 0,$$

i.e.,

$$H_s T'(0)\,\mathbf{u}_s = 0. \tag{A.11}$$

Acknowledgments. We are grateful to N. Ercolani for useful discussions. One of the authors, J.L., acknowledges support from the Thematic Program on Partial Differential Equations at the Fields Institute for Research in Mathematical Sciences in Toronto, during her sabbatical leave. Parts of this paper were written when J.L. was visiting the Kavli Institute for Theoretical Physics at the University of California Santa Barbara (which is supported by the National Science Foundation under grant PHY-9907949). Finally, parts of this paper were written when S.L. was a Sesqui postdoctoral fellow at the School of Mathematics and Statistics of the University of Sydney.

REFERENCES

[1] E. Kuznetsov, C. D. Levermore, and N. Ercolani, eds., *Advances in Nonlinear Mathematics and Science: A Special Issue to Honor Vladimir Zakharov*, Phys. D, 152–153 (2001).

[2] J. W. Evans, *Nerve axon equations.* IV. *The stable and unstable impulse*, Indiana Univ. Math. J., 24 (1975), pp. 1169–1190.

[3] C. K. R. T. Jones, *Stability of the travelling wave solution to the FitzHugh-Hagumo equation*, Trans. Amer. Math. Soc., 286 (1984), pp. 431–469.

[4] E. Yanagida, *Stability of fast traveling pulse solutions of the FitzHugh-Nagumo equations*, J. Math. Biol., 22 (1985), pp. 81–104.

[5] J. Alexander, R. Gardner and C. Jones, *A topological invariant arising in the stability analysis of travelling waves*, J. Reine Angew. Math., 410 (1990), pp. 167–212.

[6] R. Pego and M. Weinstein, *Eigenvalues, and instabilities of solitary waves*, Phil. Trans. R. Soc. London Ser. A, 340 (1992), pp. 47–94.

[7] R. A. Gardner and K. Zumbrun, *The gap lemma and geometric criteria for instability of viscous shock profiles*, Comm. Pure Appl. Math., 51 (1998), pp. 797–855.

[8] T. Kapitula and B. Sandstede, *Stability of bright solitary-wave solutions to perturbed nonlinear Schrödinger equations*, Phys. D, 124 (1998), pp. 58–103.

[9] B. Sandstede, *Stability of travelling waves*, in *Handbook of Dynamical Systems* II: *Towards Applications*, B. Fiedler, ed., North–Holland, Amsterdam, 2002, pp. 983–1055.

[10] T. J. Bridges, G. Derks, and G. Gottwald, *Stability and instability of solitary waves of the fifth-order KdV equation: A numerical framework*, Phys. D, 172 (2002), pp. 190–216.
[11] M. Grillakis, J. Shatah, and W. Strauss, *Stability theory of solitary waves in the presence of symmetry,* I, J. Funct. Anal., 74 (1987), pp. 160–197.
[12] M. Grillakis, J. Shatah, and W. Strauss, *Stability theory of solitary waves in the presence of symmetry,* II, J. Funct. Anal., 94 (1990), pp. 308–348.
[13] A. Goriely and M. Tabor, *Nonlinear dynamics of filaments* II: *Nonlinear analysis*, Phys. D, 105 (1997), pp. 45–61.
[14] S. Lafortune and J. Lega, *Instability of local deformations of an elastic rod*, Phys. D, 182 (2003), pp. 103–124.
[15] J. Lega and A. Goriely, *Pulses, fronts, and oscillations of an elastic rod*, Phys. D, 132 (1999), pp. 373–391.
[16] P. Olver, *Applications of Lie Groups to Differential Equations*, Springer-Verlag, New York, 1993.
[17] G. Dangelmayr and E. Knobloch, *The Takens-Bogdanov bifurcation with* O(2)*-symmetry*, Phil. Trans. Roy. Soc. London Ser. A, 322 (1987), pp. 243–279.
[18] J. Carminati and K. Vu, *Symbolic computation and differential equations: Lie symmetries*, J. Symbolic Comput., 29 (2000), pp. 95–116.
[19] E. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw–Hill, New York, 1955.
[20] M. Reed and B. Simon, *Methods of Modern Mathematical Physics.* IV. *Analysis of Operators*, Academic Press, New York, 1978.
[21] P. Lax, *Linear Algebra*, John Wiley & Sons, New York, 1996.
[22] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Springer-Verlag, New York, 1981.
[23] A. Goriely, M. Nizette, and M. Tabor, *On the dynamics of elastic strips*, J. Nonlinear Sci., 11 (2001), pp. 3–45.

SIAM J. MATH. ANAL.
Vol. 36, No. 6, pp. 1742–1771

A REACTION DISPERSION SYSTEM AND RAMAN INTERACTIONS*

MICHAEL I. WEINSTEIN† AND VADIM ZHARNITSKY‡

Abstract. We consider the problem of amplification of an optical signal wave with an optical pump wave when both are propagating in the fundamental mode of a single mode optical waveguide. We introduce a system of Ginzburg–Landau type and study the radiation loss due to the nonlinear interaction between the signal and the pump waves. The linear dynamics are dispersive, while nonlinearity governs the transfer of energy from the pump wave to the signal wave. The strength of the effect is shown to depend on a dimensionless parameter, which is given by the ratio of the diffraction length and amplification length. If this parameter is small, then the radiation loss is small. This result is established by (i) verifying the absence of resonant terms that can potentially drive the growth of radiative components and (ii) then by estimating the oscillatory (nonresonant) terms by proving the relevant PDE a priori estimates. These estimates require appropriate bounds on the solutions of the PDE, whose only conserved integral is the L^2 norm. However, the special structure of the nonlinear term, dictated by the physics of the Raman effect, implies a weak space-time bound involving the signal and pump intensities. This bound and L^2 conservation are used together with Strichartz (space-time) estimates for the Schrödinger equation to obtain control of stronger classical norms of the signal and pump fields.

Key words. Landau–Ginzburg equations, Raman interaction, nonlinear optics, optical waveguides

AMS subject classifications. 35Q55, 35Q60, 78A60

DOI. 10.1137/S0036141003428172

1. Introduction.

In this paper we study a system of nonlinear and dispersive partial differential equations, where nonlinearity is of "reaction" type, i.e., in the absence of dispersion it induces pure energy exchange between the fields. Such systems are reminiscent of reaction-diffusion systems; here the diffusive mechanism is replaced by dispersion. While the latter has been widely studied, the former has received very little attention.

Our reaction-dispersion system arises naturally in mathematical modeling of the stimulated Raman process, but will also arise in other systems (perhaps in somewhat modified form), where two dispersive waves interact nonlinearly, while other nonlinear effects (such as self-phase modulation) and diffusion are negligible. This system can be also considered as a special case of complex Ginzburg–Landau (CGL) system, which has been studied in a different parameter regime [2]. We will often refer to the systems (1.1) and (1.2) as the Raman model, due to their relation to the motivating physical context.

The Raman effect is one where light of one frequency, ω_s ("signal"), is amplified by light of a down shifted frequency, ω_p ("pump"). Taking the energy transfer characteristics of the Raman process into account as well as diffraction leads to the system

*Received by the editors May 21, 2003; accepted for publication (in revised form) August 19, 2004; published electronically May 20, 2005.

http://www.siam.org/journals/sima/36-6/42817.html

†Mathematical Sciences Research, Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974. Current address: Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY 10027 (miw2103@columbia.edu).

‡Department of Mathematics, University of Illinois at Urbana-Champaign, Urbana, IL 61801-2975 (vz@math.uiuc.edu). This author's research was supported by NSF grant DMS-0219233 and partially supported by NSF grant DMS-0073923.

of nonlinear evolution equations, discussed in greater detail in Appendix A:

$$\begin{aligned} i\partial_t u_s &= -\Delta u_s + i\epsilon|u_p|^2 u_s, \\ i\partial_t u_p &= -\Delta u_p - i\epsilon|u_s|^2 u_p. \end{aligned} \tag{1.1}$$

More generally, we must include the Kerr effect in (1.1). This would introduce cross-phase and self-phase modulation terms of the types $\alpha_s|u_p|^2u_s + \beta_s|u_s|^2u_s$ and $\alpha_p|u_s|^2u_s+\beta_p|u_p|^2u_p$ with $\alpha_{s,p}$, $\beta_{s,p}$ real on the right-hand side of (1.1). We remark on the analysis of this more complete model at the end of this section; see Remark 1.1. In a waveguide setting, which is of importance in optical communications, the equations take the form

$$\begin{aligned} i\partial_t u_s &= Hu_s + i\epsilon|u_p|^2 u_s, \\ i\partial_t u_p &= Hu_p - i\epsilon|u_s|^2 u_p, \end{aligned} \tag{1.2}$$

where

$$H = -\Delta + V.$$

Here, $u_s = u_s(x,t)$ and $u_p = u_p(x,t)$ denote, respectively, the signal and pump complex electric field envelopes. Systems (1.1) and (1.2) are valid, assuming the *paraxial approximation*. Δ denotes the Laplace operator with respect to x ($x \in \mathbb{R}^1$ or $x \in \mathbb{R}^2$). The longitudinal coordinate (z), a *time-like* variable with which propagation distance is measured, is denoted by t. In the waveguide setting, the "potential" $V(x)$ is determined by the transverse refractive index profile of the waveguide. The parameter ϵ measures the size of the nonlinear effects relative to the linear effects (e.g., diffraction, dispersion). The particular application to optical communications is discussed in Appendix A.

Our study of systems (1.2) and (1.1) is motivated by a fundamental issue arising in the modeling of the Raman interaction in a waveguide setting. In optical communication applications, the weak signal field whose envelope encodes bits of information is amplified by the strong pump field. This process takes place in an optical fiber waveguide, with one transverse localized state or "guided mode" and radiation modes. Raman amplification of the signal is based on the intended net transfer of energy from the pump to the signal. Physicists have found that good agreement with experiment is achieved by an ODE model, in which one neglects the effect of nonlinear coupling of bound states to radiation modes:

$$\begin{aligned} &\partial_t I_s = \epsilon g_s I_s I_p, \\ &\partial_t I_p = -\epsilon g_p I_s I_p, \\ &I_s \sim |u_s|^2,\ I_p \sim |u_p|^2, \end{aligned} \tag{1.3}$$

where $g_{s,p}$ are coefficients depending on the properties of the fundamental modes, e.g., on frequency and the so-called effective area.

A satisfactory explanation for the above approximation has been lacking; see, for example, [3]. This motivated us to consider the Raman energy transfer problem in the context of the model (1.2). We have found an explanation for the above statement about energy transfer using ideas and methods of resonance and averaging. In particular, in Theorem 3.1 we establish that if the initial field energy, which is not small, is in the guided mode, then this property persists with negligible error on the

length scale of physical interest, $\mathcal{O}(\epsilon^{-1})$, and therefore the model (1.3) applies. Note that on the time interval of order ϵ^{-1}, the radiation of size ϵ can grow to become of order $\mathcal{O}(1)$ (since the rate of radiation change is of order $\mathcal{O}(\epsilon)$).

The proof of Theorem 3.1 requires a good understanding of the large time dynamics of the flow defined by (1.2). Thus we consider the question of global existence of the initial value problem for such systems and we have derived results of independent interest. The standard approach to controlling the large time dynamics is to first prove local in time existence of the solutions to the initial value problem in a "natural" Banach space. A "natural" space is often one in which the physically relevant conserved integrals are defined. We formulate initial value problem for system (1.2) as a system of integral equations and prove that there are local solutions using fixed point argument; see sections 2 and 4. Global existence in time then follows from an appropriate a priori bound on the norm of this Banach space. If this norm remains bounded in terms of conserved integrals of the flow, global existence holds.

The ideas we use to prove global existence apply to both systems (1.1) and (1.2). Equations (1.1) and (1.2) have the L^2 (energy) conservation law

$$\mathcal{P}[u_s(t), u_p(t)] \equiv \int_{-\infty}^{+\infty} (|u_s|^2 + |u_p|^2)\, dx = \mathcal{P}[u_s(0), u_p(0)] \equiv \mathcal{P}_0. \tag{1.4}$$

Unfortunately, L^2 is a very weak space in which to control the nonlinear flow. Unlike the nonlinear Schrödinger equation, a Hamiltonian system, (1.2) and (1.1) do not have a second conserved integral (Hamiltonian), which controls $\|\nabla u_{s,p}\|_{L^2}$, and from which sufficient a priori control follows for global existence to hold.

We find that the key to a global existence theory is the following space-time integral a priori bound, which is a consequence of the form of the nonlinear Raman interaction terms:

$$\int_0^T dt \int |u_s|^2 |u_p|^2\, dx \le \frac{1}{2}\mathcal{P}_0. \tag{1.5}$$

Remark 1.1. We believe our theory can be extended to system (1.1) with Kerr effect included. An essential ingredient is the space-time estimate (1.5), which holds for the more general system. However, a more technical analysis is required to obtain closed space-time estimates in the presence of Kerr effect terms. This is work in progress.

Outline of the paper. The paper is structured as follows. We first consider system (1.2) in one space dimension and one time dimension. In section 2 we prove global well-posedness for the solution of the initial value problem. The key to this result are certain a priori estimates, whose point of departure is the L^2 conservation law (1.4) and the space-time a priori bound (1.5). This space-time estimate implies that (1.2) may be viewed as an inhomogeneous system of equations for u_s and u_p, with a source term, which is bounded in a space-time norm. Strichartz estimates [8] are then used to bound u_s and u_p individually in space-time norms and then finally in classical Sobolev norms. In section 3 the energy transfer from the bound mode of a single mode waveguide to radiation modes is studied by estimating nonresonant oscillating terms. Section 4 contains a theory of well-posedness in the case of two transverse dimensions. Finally, there are three appendices: one with a detailed discussion of the motivating application to optical communications, the second one in which we prove the normal form result, based on the symmetries of the system, and the third one describing numerical simulations.

2. Existence theory on $\mathbb{R}^1$. In this section we prove that system (1.2) has a unique global solution in an appropriate (physical) function space. In subsection 2.1 we provide the required operator estimates we shall require. In subsection 2.2 we derive certain a priori bounds which are satisfied by solutions of (1.2). In subsection 2.3 existence in a "weak" space is proved. In section 2.4 it is shown how to extend these results to H^s.

2.1. Estimates for the linear flow. We first introduce the fundamental solution of the Schrödinger equation.

(1) The solution of the initial value problem

$$i\partial_t u = -\partial_x^2 u, \quad u(0,x) = f \tag{2.1}$$

is denoted by $U_0(t)f$ and $U_0(t)$ is called the free propagator.

(2) The solution of the initial value problem

$$i\partial_t u = (-\partial_x^2 + V(x))u = Hu, \quad u(0,x) = f \tag{2.2}$$

is denoted by $U(t)f$.

The operator, H, may have spectrum consisting of continuous and discrete parts, with associated spectral projections P_c and $P_b = I - P_c$. We shall assume that H has finitely many point eigenvalues. Intuitively, on the range of P_c we expect $U(t)$ to behave dispersively in a manner similar to $U_0(t)$. We use dispersive estimates which involve space-time integrals, often referred to as estimates of Strichartz type; see [8] and [9]. The proofs of the space-time estimates for the free Schrödinger equation in the form we use are due to Ginibre and Velo [12] and, in the inhomogeneous case, to Yajima [10] and Cazenave and Weissler [11]. For complete proofs see, for example, Theorems 3.3 and 3.4 of [4].

We now introduce the function spaces and the notation of an *admissible pair* in terms of which the space-time estimates are expressed.

(3) For a real interval I and a Banach space X, we denote by $L^p(I,X)$ the Banach space of functions $u : I \to X$ for which $\int_I \|u(t)\|_X^p \, dt$ is finite.

(4) A pair of real numbers (q,r) is called *admissible* (for dimension $n = 1$) if

$$\frac{2}{q} = \frac{1}{2} - \frac{1}{r}, \quad r \in [2,\infty]. \tag{2.3}$$

We now state Strichartz-type estimates for $U(t)$ for the initial value problem (2.2) and the inhomogeneous problem

$$i\partial_t u = Hu + P_c g. \tag{2.4}$$

THEOREM 2.1. *Assume that V satisfies*

$$\int_{-\infty}^{+\infty} |V(x)|(1+|x|)^{5/2} \, dx < \infty. \tag{2.5}$$

Thus, H has finitely many negative eigenvalues and continuous spectrum extending from 0 to $+\infty$, with associated spectral projections P_b and P_c.[1]

[1] Finiteness of negative discrete spectrum and continuity of spectrum from $[0,\infty)$ follows from sufficient decay of the potential.

Let (q,r) be an admissible pair. For any $f \in L^2$, $U_0(t)f$ and $U(t)P_cf$ are of class $L^q(\mathbb{R}, L^r)$ and satisfy the estimates

$$\begin{aligned} \|U_0(\cdot)f\|_{L^q([0,T],L^r)} &\le C\|f\|_{L^2}, \\ \|U(\cdot)P_cf\|_{L^q([0,T],L^r)} &\le C\|f\|_{L^2}, \end{aligned} \tag{2.6}$$

where C depends only on q.

THEOREM 2.2. *Let V satisfy* (2.5) *and let (γ,ρ) be an admissible pair $(2/\gamma = 1/2 - 1/\rho)$, $f \in L^{\gamma'}([0,T], L^{\rho'})$, where (γ',ρ') is conjugate to (γ,ρ). Then for any admissible pair (q,r) $(2/q = 1/2 - 1/r)$*

$$\begin{aligned} \left\| \int_0^t U_0(t-\tau)f(\tau)d\tau \right\|_{L^q([0,T],L^r)} &\le C\|f\|_{L^{\gamma'}([0,T],L^{\rho'})}, \\ \left\| \int_0^t U(t-\tau)P_cf(\tau)d\tau \right\|_{L^q([0,T],L^r)} &\le C\|f\|_{L^{\gamma'}([0,T],L^{\rho'})}, \end{aligned} \tag{2.7}$$

where C depends only on q, γ.

The proofs rely on the $L^p - L^{p'}$ estimates for the free Schrödinger equation

$$\|U_0(t)f\|_{L^p} \le (4\pi|t|)^{-\frac{1}{2}+\frac{1}{p}} \|f\|_{L^{p'}}. \tag{2.8}$$

In the case of a Schrödinger equation with a potential in one space dimension the analogous estimate for $U(t)P_c$ was established by Weder [5]. Adapting the proofs in [4], for the free propagator $U_0(t)$, and using Weder's estimate, one obtains the Strichartz-type estimates for Schrödinger equation with a potential.

The following corollary, easily derived from the previous estimates by a change of variables, concerns the dependence of space-time estimates on a parameter, which arises when we rescale (1.2).

COROLLARY 2.3. *Consider a one-parameter family of Schrödinger initial value problems*

$$\begin{aligned} i\partial_t u &= \beta H u, \\ u(x,0) &= f(x), \end{aligned} \tag{2.9}$$

where $\beta \in [\beta_0, \infty)$ with $\beta_0 > 0$. Assume that the potential, $V(x)$, satisfies (2.5). *Then, the conclusions of Theorems* 2.1 *and* 2.2 *hold with β-dependent constants. In particular, if $U_\beta(t)f = U(\beta t)f$ denotes the solution of the initial value problem* (2.9), *then*

$$\|U_\beta(\cdot)P_cf\|_{L^q([0,T],L^r)} \le C_1(\beta,q)\|f\|_{L^2}, \tag{2.10}$$

$$\left\| \int_0^t U_\beta(t-\tau)P_cf(\tau)d\tau \right\|_{L^q([0,T],L^r)} \le C_2(\beta,q,\gamma')\|f\|_{L^{\gamma'}([0,T],L^{\rho'})}, \tag{2.11}$$

where

$$C_1(\beta,q) = C_{11}(q)\beta^{-\frac{1}{q}}, \qquad C_2(\beta,q,\gamma') = C_{22}(q,\gamma')\beta^{-1-\frac{1}{q}+\frac{1}{\gamma'}}. \tag{2.12}$$

In [6], Weder proves continuity of wave operators for (2.2). We use a special case of the main theorem from [6].

PROPOSITION 2.4. *Let V satisfy* (2.5). *Then, there exist wave operators Ω and Ω^* satisfying*

$$\Omega(I-\Delta)\Omega^* = (I+H)P_c. \tag{2.13}$$

These operators are continuous in H^1:

$$\|\Omega f\|_{H^1} \le C\|f\|_{H^1}, \qquad \|\Omega^* f\|_{H^1} \le C\|f\|_{H^1}. \tag{2.14}$$

2.2. A priori space-time estimates. Essential in the proof that system (1.2) defines the solution globally in time and that the solution does not develop singularities are a priori estimates which we now derive. For convenience, we rescale the time $\epsilon t = t_{\text{new}}$, so the uniform bound on the interval $t_{\text{new}} \in [0,T]$ will correspond to the interval $[0,T/\epsilon]$ in old time. We will continue to use t as the time variable

$$\partial_t u_s + i\beta H u_s = |u_p|^2 u_s, \tag{2.15}$$

$$\partial_t u_p + i\beta H u_p = -|u_s|^2 u_p, \tag{2.16}$$

where $\beta := \epsilon^{-1}$ is a dispersion/diffraction parameter and $\beta \in [\beta_0, \infty]$.

Multiplication of (2.15) by $\overline{u_s}$, taking the real part of the resulting equation and integrating over all gives

$$\frac{d}{dt}\int |u_s|^2\,dx = 2\int |u_s|^2|u_p|^2\,dx. \tag{2.17}$$

Similarly, multiplication of (2.16) by $\overline{u_p}$ yields

$$\frac{d}{dt}\int |u_p|^2\,dx = -2\int |u_s|^2|u_p|^2\,dx. \tag{2.18}$$

Equations (2.17) and (2.18) express the gain of signal energy at the expense of pump energy and the depletion of pump energy at the expense of signal energy; see (1.2).

Addition of (2.17) and (2.18) yields the conservation law

$$\frac{d}{dt}\int |u_s|^2 + |u_p|^2\,dx = 0 \tag{2.19}$$

or

$$\int |u_s|^2 + |u_p|^2\,dx = \int |u_s(0)|^2 + |u_p(0)|^2\,dx \equiv \mathcal{P}_0. \tag{2.20}$$

An important step in our analysis is to use the *energy dissipation identity* (2.18). Integration of (2.18) over time interval $[0,T]$ yields

$$2\int_0^T \int |u_s(x,t)|^2|u_p(x,t)|^2\,dx\,dt = \int |u_p(x,0)|^2\,dx - \int |u_p(x,t)|^2\,dx. \tag{2.21}$$

A simple consequence of (2.21) is the following *space-time bound.*

PROPOSITION 2.5 (a priori space-time estimate). *Let (u_s,u_p) denote a solution of* (2.15)–(2.16) *in the sense of Theorem* 2.9. *Then,*

$$\int_0^T \int |u_s|^2|u_p|^2\,dx\,dt \le \frac{1}{2}\mathcal{P}_0. \tag{2.22}$$

Remark 2.6. Since in Theorem 2.9 we assume that the initial conditions are merely L^2, strictly speaking the above derivation of the bound (2.22) is not valid, since the manipulations require that the solution is a classical solution of the PDE, i.e., pointwise differentiable in space and time. At the end of section 2.3, we sketch a proof of (2.22) for the case of L^2 initial data.

Equations (2.16) and (2.15) can both be viewed as inhomogeneous Schrödinger equations of the form

$$i\partial_t U = \beta H U + g \tag{2.23}$$

with a source term g given by

$$g = |u_p|^2 u_s \quad \text{or} \quad g = -|u_s|^2 u_p. \tag{2.24}$$

We next show that the a priori estimate (2.22) implies bounds on the source terms (2.24) which are suitable for application of the inhomogeneous Strichartz estimate of Theorem 2.2.

PROPOSITION 2.7. *Let g denote either term in* (2.24). *Then, for any $T > 0$ and any $\kappa \in [0, 2]$,*

$$\int_0^T \left| \int |g|\, dx \right|^\kappa \, dt \le 2^{-\frac{\kappa}{2}} \mathcal{P}_0^\kappa T^{\frac{2-\kappa}{2}}. \tag{2.25}$$

In particular, for $\kappa \in [1, 2]$ and $g \in L^\kappa([0, T], L^1)$,

$$\|g\|_{L^\kappa([0,T],L^1)} \le 2^{-\frac{1}{2}} \mathcal{P}_0 T^{\frac{2-\kappa}{2k}}. \tag{2.26}$$

To prove Proposition 2.7, let $g = -|u_s|^2 u_p$. The proof for $g = |u_p|^2 u_s$ is analogous. By the Cauchy–Schwarz inequality,

$$\left| \int |u_s|^2 |u_p|\, dx \right| \le \left(\int |u_s|^2\, dx \right)^{\frac{1}{2}} \left(\int |u_s|^2 |u_p|^2\, dx \right)^{\frac{1}{2}}. \tag{2.27}$$

Squaring this inequality and integrating the result over the time interval $[0, T]$ yields, after using that the L^2 norm of u_s is bounded by $\mathcal{P}_0$, that

$$\int_0^T \left| \int |u_s|^2 |u_p|\, dx \right|^2 \, dt \le \frac{1}{2} \mathcal{P}_0^2. \tag{2.28}$$

This handles the case $\kappa = 2$. For $\kappa = 0$ the trivial bound of T holds. The result follows by interpolation.

Using the a priori bounds of Proposition 2.7 we can now estimate the solution in $L^q([0, T], L^r)$ spaces, for admissible (q, r).

THEOREM 2.8 (a priori bounds in $L^q([0, T], L^r)$). *Let (q, r) be an admissible pair; see* (2.3). *Then, any solution (u_s, u_p) of system* (2.15)–(2.16) *for $0 \le t \le T$ satisfies the bounds*

$$\begin{aligned} \|u_s\|_{L^q([0,T],L^r)} &\le C \left(\mathcal{P}_0 + \mathcal{P}_0^{\frac{1}{2}} \right) (T + 1), \\ \|u_p\|_{L^q([0,T],L^r)} &\le C \left(\mathcal{P}_0 + \mathcal{P}_0^{\frac{1}{2}} \right) (T + 1), \end{aligned} \tag{2.29}$$

where C depends only on q, γ, and β_0.

Proof of Theorem 2.8. We estimate $\|u_s\|_{L^q([0,T],L^r)}$. The corresponding estimate for u_p is similar. Equation (2.15) can be rewritten as an equivalent integral equation:

$$u_s(t) = U_\beta(t)u_s(0) + \int_0^t U_\beta(t-\tau)|u_p(\tau)|^2 u_s(\tau)d\tau. \tag{2.30}$$

To estimate the space-time norm of u_s, we apply Corollary 2.3 to the continuous spectral part and estimate the finite-dimensional (bound state) part of u_s separately. For ease of presentation we assume that H has only one spatial localized bound state solution, $\phi(x)$; the proof is the same for any finite number of bound states. Estimation of (2.30) using Corollary 2.3 gives

$$\begin{aligned}
&\|u_s\|_{L^q([0,T],L^r)} \\
&\quad\le C\left\|U_\beta(t)u_s(0) + \int_0^t U_\beta(t-\tau)|u_p(\tau)|^2 u_s(\tau)d\tau\right\|_{L^q([0,T],L^r)} \\
&\quad\le C_1(\|u_s(0)\|_{L^2} + \langle u_s(0),\phi\rangle\|\phi\|_{L^q([0,T],L^r)}) + C_2\|u_p^2 u_s\|_{L^{\gamma'}([0,T],L^{\rho'})} \\
&\qquad + \left\|\int_0^t U_\beta(t-\tau)\langle|u_p(\tau)|^2 u_s(\tau),\phi\rangle\phi d\tau\right\|_{L^q([0,T],L^r)} \\
&\quad\le C_1(1+T^{q^{-1}}\|\phi\|_{L^2}\|\phi\|_{L^r})\|u_s(0)\|_{L^2} + C_2\|u_p^2 u_s\|_{L^{\gamma'}([0,T],L^{\rho'})} \\
&\qquad + \left\|\int_0^t e^{-i\lambda(t-\tau)}\langle|u_p(\tau)|^2 u_s(\tau),\phi\rangle\phi d\tau\right\|_{L^q([0,T],L^r)}.
\end{aligned} \tag{2.31}$$

The last integral is estimated as follows: using that

$$\begin{aligned}
\left|\int_0^t \langle|u_p(\tau)|^2 u_s(\tau),\phi\rangle\phi d\tau\right| &\le \left|\int_0^t |\phi(x)|\int|\phi(x)|\cdot|u_p^2(x,\tau)u_s(x,\tau)|\,dxd\tau\right| \\
&\le |\phi(x)|\int_0^t \|\phi\|_{L^\rho}\cdot\|u_p^2 u_s\|_{L^{\rho'}}d\tau \\
&\le |\phi(x)|\cdot\|\phi\|_{L^\gamma([0,T],L^\rho)}\cdot\|u_p^2 u_s\|_{L^{\gamma'}([0,T],L^{\rho'})}
\end{aligned}$$

we have

$$\begin{aligned}
&\left\|\int_0^t e^{-i\lambda(t-\tau)}\langle|u_p(\tau)|^2 u_s(\tau),\phi\rangle\phi d\tau\right\|_{L^q([0,T],L^r)} \\
&\quad\le \|\phi\|_{L^\gamma([0,T],L^\rho)}\cdot\|\phi\|_{L^q([0,T],L^r)}\cdot\|u_p^2 u_s\|_{L^{\gamma'}([0,T],L^{\rho'})} \\
&\quad= \|\phi\|_{L^\rho}\cdot\|\phi\|_{L^r}\cdot\|u_p^2 u_s\|_{L^{\gamma'}([0,T],L^{\rho'})}\cdot T^{\gamma^{-1}+q^{-1}}.
\end{aligned}$$

Now Proposition 2.7 implies a bound on $\|u_p^2 u_s\|_{L^{\gamma'}([0,T],L^{\rho'})}$, where $\rho' = 1$ and $\gamma' \in [0,2]$. Note that the exponents γ and ρ, dual to γ' and $\rho' = 1$, form an admissible pair provided $\gamma = 4$ and $\gamma' = 4/3$.[2]

Setting $\gamma' = 4/3$ and $\rho' = 1$ in (2.31) and applying Proposition 2.7 with $\kappa = \gamma' = 4/3$ implies

$$\begin{aligned}
\|u_s\|_{L^q([0,T],L^r)} &\le C_1(1+T^{1/q}\|\phi\|_{L^2}\|\phi\|_{L^r})\|u_s(0)\|_{L^2} \\
&\quad + C_2\|u_p^2 u_s\|_{L^{\frac{4}{3}}([0,T],L^1)} + T^{1/4+1/q}\|\phi\|_{L^\infty}\|\phi\|_{L^r}\|u_p^2 u_s\|_{L^{\frac{4}{3}}([0,T],L^1)} \\
&\le C_1(1+T^{\frac{1}{q}})\mathcal{P}_0^{\frac{1}{2}} + C_2\mathcal{P}_0 T^{\frac{1}{4}} + C_3\mathcal{P}_0 T^{\frac{1}{2}+\frac{1}{q}} \le C(\mathcal{P}_0+\mathcal{P}_0^{\frac{1}{2}})(T+1),
\end{aligned} \tag{2.32}$$

[2]Indeed, since $1/\rho' + 1/\rho = 1$, $1/\gamma + 1/\gamma' = 1$, and $2/\gamma = 1/2 - 1/\rho$, we have $\rho = \infty$, $\gamma = 4$, and $\gamma' = 4/3$.

where C_1, C_2, C_3 depend on the corresponding norms of ϕ and we used that $4 \le q \le \infty$. This completes the proof of Theorem 2.8. □

2.3. Existence in $L^8(\mathbb{R}_+, L^4) \cap L^\infty(\mathbb{R}_+, L^2)$. In this subsection we prove existence of solutions in a function space $\mathcal{X}(T)$ defined by

$$\mathcal{X}(T) = L^8([0,T], L^4) \cap L^\infty([0,T], L^2) \tag{2.33}$$

with the norm

$$\|u\|_{\mathcal{X}(T)} = \|u\|_{L^8([0,T],L^4)} + \|u\|_{L^\infty([0,T],L^2)} \tag{2.34}$$

$$= \|u\|_{8,4} + \|u\|_{\infty,2}, \tag{2.35}$$

the latter being written when there is no ambiguity. For the two-dimensional field (u_s, u_p), we naturally define the norm

$$\|u_s, u_p\|_{\mathcal{X}(T)} = \|u_s\|_{\mathcal{X}(T)} + \|u_p\|_{\mathcal{X}(T)} = \|u_s\|_{8,4} + \|u_s\|_{\infty,2} + \|u_p\|_{8,4} + \|u_p\|_{\infty,2}. \tag{2.36}$$

Since Theorem 2.8 gives a priori control of solutions in $L^q([0,T], L^r)$ spaces for any admissible pairs (q,r), it is natural to obtain a local existence theorem in a space, where the maximal time of existence depends only on $L^q([0,T], L^r)$ bounds. Then, global existence follows from Theorem 2.8; see the discussion below.

Define the mapping

$$(u_s, u_p) \mapsto A_\beta(u_s, u_p) \equiv \left(A_\beta^{(s)}(u_s, u_p), A_\beta^{(p)}(u_s, u_p) \right), \tag{2.37}$$

where

$$A_\beta^{(s)}(u_s, u_p) = U_\beta(t) u_s(0) + \int_0^t U_\beta(t-\tau) |u_p(\tau)|^2 u_s(\tau) d\tau, \tag{2.38}$$

$$A_\beta^{(p)}(u_s, u_p) = U_\beta(t) u_p(0) - \int_0^t U_\beta(t-\tau) |u_s(\tau)|^2 u_p(\tau) d\tau. \tag{2.39}$$

Then, the above evolution equation has the equivalent formulation as a fixed point problem.

For initial data $(u_s(0), u_p(0)) \in L^2$, find $(u_s, u_p) \in \mathcal{X}(T)$ for some $T > 0$ such that

$$(u_s, u_p) = A_\beta(u_s, u_p). \tag{2.40}$$

Our local existence theorem is the following.

THEOREM 2.9 (local existence).

(1) *Given initial data $(u_s(0), u_p(0)) \in L^2$, there exist a $T > 0$ and a unique solution $(u_s, u_p) \in \mathcal{X}(T)$ of* (2.40). *This local solution satisfies the a priori estimate* (2.22).

(2) *Let $T_{\max} > 0$ denote the maximal time of existence. Either $T_{\max} = \infty$ (global existence in time) or*

$$T_{\max} < \infty \quad \text{and} \quad \limsup_{T \to T_{\max}} \|(u_s, u_p)\|_{\mathcal{X}(T)} = \infty. \tag{2.41}$$

Using the local existence theory of Theorem 2.9 and the a priori bounds of Theorem 2.8, we have $T_{\max} = \infty$. Therefore, the following result holds.

THEOREM 2.10 (global existence). *For any initial data* $(u_s(0), u_p(0))$ *in* L^2, (2.40) *has a unique global solution of class* $L^8(\mathbb{R}_+, L^4) \cap L^\infty(\mathbb{R}_+, L^2)$.

We need only prove the local existence Theorem 2.9. The proof follows from the next two propositions in which we establish that for L^2 initial conditions and T sufficiently small,

(i) the transformation A_β maps a specified ball $\mathcal{B}(T)$ in $\mathcal{X}(T)$ into itself and that

(ii) A_β is a contraction mapping on $\mathcal{B}(T)$.

PROPOSITION 2.11. *Let* $(u_s(0), u_p(0))$ *be in* L^2. *Define the ball in* $\mathcal{X}(T)$

$$\mathcal{B}(T) = \{(u_s, u_p) \in \mathcal{X}(T) : \|(u_s, u_p)\|_{\mathcal{X}(T)} \le 2C(\|u_s(0)\|_2 + \|u_p(0)\|_2)\}, \tag{2.42}$$

where C *is found in the proof below. There exists* $T_0 > 0$ *such that for any* $T < T_0$, *the ball is mapped into itself, i.e.,* $A_\beta(\mathcal{B}(T)) \subset \mathcal{B}(T)$ *for any* $\beta \in [\beta_0, \infty]$, *with* T_0 *depending on* β_0.

Proof of Proposition 2.11. We estimate the action of $A_\beta^{(s)}$. The estimation for $A_\beta^{(p)}$ is similar.

Following the proof of Theorem 2.8, we obtain a similar inequality

$$\begin{aligned}\|A_\beta^{(s)}(u_s, u_p)\|_{q,r} &\le \|U_\beta(t)u_s(0)\|_{q,r} + \left\|\int_0^t U_\beta(t-\tau)|u_p|^2 u_s\right\|_{q,r} \\ &\le C_1(1+T)\|u_s(0)\|_2 + C_2(1+T)\|u_p^2 u_s\|_{\gamma',\rho'},\end{aligned}$$

where C_1, C_2 depend on ϕ.

Estimation of the cubic term proceeds as follows. By the Cauchy–Schwarz inequality,

$$\begin{aligned}\|u_p^2 u_s\|_{\gamma',\rho'} &= \left[\int_0^T \left(\int |u_p^2 u_s|^{\rho'}\,dx\right)^{\gamma'/\rho'} dt\right]^{1/\gamma'} \\ &\le \left[\int_0^T \left(\int |u_p|^{4\rho'}\,dx\right)^{\gamma'/\rho'} dt\right]^{1/2\gamma'} \left[\int_0^T \left(\int |u_s|^{2\rho'}\,dx\right)^{\gamma'/\rho'} dt\right]^{1/2\gamma'}.\end{aligned}$$

Set $\rho' = 1$ and therefore $\gamma' = 4/3$. Then the last expression becomes

$$= \left[\int_0^T \left(\int |u_p|^4\,dx\right)^{4/3} dt\right]^{3/8} \left[\int_0^T \left(\int |u_s|^2\,dx\right)^{4/3} dt\right]^{3/8} \tag{2.43}$$

and by Hölder's inequality, applied to each factor, we have the bound

$$\begin{aligned}&\le \left[\left[\int_0^T \left(\int |u_p|^4\,dx\right)^2 dt\right]^{(2/3)\cdot(3/8)} \left[\int_0^T 1^3\,dt\right]^{(1/3)}\right]^{3/8} \left[\sup_t \|u_s(t)\|_2^{2\cdot(4/3)} T\right]^{3/8} \\ &\quad \le \|u_p\|_{8,4}^2 \|u_s\|_{\infty,2} T^{1/2}.\end{aligned}$$

Adding up all the terms, we obtain

$$\begin{aligned}\left\|A_\beta^{(s)}(u_s, u_p)\right\|_{\mathcal{X}(T)} &\le C(\|u_s(0)\|_2 + \|u_p\|_{8,4}^2 \|u_s\|_{\infty,2} T^{\frac{1}{2}})(1+T), \\ \left\|A_\beta^{(p)}(u_s, u_p)\right\|_{\mathcal{X}(T)} &\le C(\|u_p(0)\|_2 + \|u_s\|_{8,4}^2 \|u_p\|_{\infty,2} T^{\frac{1}{2}})(1+T),\end{aligned}$$

where $C = \max\{C_1, C_2\}$. Finally, combining the last two terms, we have

$$\|A_\beta(u_s, u_p)\|_{\mathcal{X}(T)} \le C(\|u_s(0)\|_2 + \|u_p(0)\|_2 + \|(u_s, u_p)\|^3_{\mathcal{X}(T)} T^{\frac{1}{2}})(1+T). \tag{2.44}$$

Assume now that $\|(u_s, u_p)\|_{\mathcal{X}(T)} \le 2C(\|u_s(0)\|_2 + \|u_p(0)\|_2)$ and that T is sufficiently small; then A_β maps $\mathcal{B}(T)$ into itself. This completes the proof of Proposition 2.11. ∎

PROPOSITION 2.12. *For $T < T_1 \le T_0$ sufficiently small, the transformation, A_β, is a contraction on $\mathcal{B}(T)$. That is,*

$$\|A_\beta(u_s, u_p) - A_\beta(v_s, v_p)\|_{\mathcal{X}(T)} \le q\|(u_s - v_s, u_p - v_p)\|_{\mathcal{X}(T)}, \tag{2.45}$$

where $0 < q < 1$.

Proof of Proposition 2.12. Consider the first component of the map. By Corollary 2.3,

$$\begin{aligned}\|A^{(s)}_\beta(u_s, u_p) - A^{(s)}_\beta(v_s, v_p)\|_{q,r} &\le C_2\||u_p|^2 u_s - |v_p|^2 v_s\|_{\frac{4}{3},1} \\ &\le \|u_p^2(u_s - v_s)\|_{\frac{4}{3},1} + \|u_p v_s(u_p - v_p)\|_{\frac{4}{3},1} \\ &\quad + \|v_p v_s(u_p - v_p)\|_{\frac{4}{3},1}.\end{aligned}$$

These terms are all estimated in a similar manner. We focus on the second term. First, by the Cauchy–Schwarz inequality,

$$\begin{aligned}\|u_p v_s(u_p - v_p)\|_{\frac{4}{3},1} &\le \left[\int_0^T \left(\int |u_p|^2 |v_s|^2\, dx\right)^{4/3} dt\right]^{3/8} \left[\int_0^T \left(\int |u_p - v_p|^2 dx\right)^{4/3} dt\right]^{3/8} \\ &\le \left[\int_0^T \left(\int |u_p|^2 |v_s|^2\, dx\right)^{4/3} dt\right]^{3/8} T^{3/8}\|u_p - v_p\|_{\infty,2}.\end{aligned}$$

Another application of the Cauchy–Schwarz inequality to the spatial integral in the first factor in the previous expression and then Hölder's inequality to the time integral gives

$$\begin{aligned}&\left[\int_0^T \left(\int |u_p|^4\, dx\right)^{4/3} dt\right]^{3/16} \left[\int_0^T \left(\int |v_s|^4\, dx\right)^{4/3} dt\right]^{3/16} T^{3/8}\|u_p - v_p\|_{\infty,2} \\ &\quad \le \|u_p\|_{8,4}\|v_s\|_{8,4} T^{\frac{2}{16}} T^{\frac{3}{8}} \|u_p - v_p\|_{\infty,2} = \|u_p\|_{8,4}\|v_s\|_{8,4} T^{1/2}\|u_p - v_p\|_{\infty,2} \\ &\quad \le \|u_p\|_{8,4}\|v_s\|_{8,4} T^{1/2}\|(u_s - v_s, u_p - v_p)\|_{\mathcal{X}(T)}.\end{aligned}$$

Adding the estimates for

$$\left\|A^{(s)}_\beta(u_s, u_p) - A^{(s)}_\beta(v_s, v_p)\right\|_{\mathcal{X}(T)} \quad \text{and} \quad \left\|A^{(p)}_\beta(u_s, u_p) - A^{(p)}_\beta(v_s, v_p)\right\|_{\mathcal{X}(T)} \tag{2.46}$$

and choosing, if necessary, $T_1 < T_0$, we obtain the contraction estimate. This completes the proof. ∎

Remark 2.13. Finally, we give a proof of the space-time bound for solutions with data in L^2.

(1) *Existence of solutions for very regular data.* Using that H^s is an algebra for $s > \frac{1}{2}$, it is standard to prove, by a contraction mapping argument, that for

data in H^s with $s \geq s_0 \geq 2$ there is a unique classical solution. However, this argument requires differentiation of the original system and as a consequence $V(x)$. This requires imposing unnecessary smoothness assumptions on V. To avoid such restrictions on V, we observe that the norms $\|\cdot\|_{H^2} = \|(I+H)P_c\cdot\|_{L^2}$ and $\|(I-\Delta)\cdot\|_{L^2}$ are equivalent, by Proposition 2.4; see also Proposition 2.15. Therefore, applying $(I + H)P_c$, which commutes with $U_\beta(\cdot)$, to the integral equation for (u_s, u_p), we can use standard estimates to obtain a classical solution. An argument of this type is implemented in section 2.4. Therefore, by the computation of section 2.2, this classical solution satisfies (2.22).

(2) *Continuity of solutions with respect to variations in the initial data.* Let (u_s, u_p) denote the solution corresponding to data $(u_s(0), u_p(0))$ and (v_s, v_p) denote the solution corresponding to data $(v_s(0), v_p(0))$. Both of these are fixed points of the operator A_β (see (2.37)) with the corresponding data. By the same estimate as in the proof of Proposition 2.12 we have (2.45) plus an additional data term on the right-hand side: $\|(u_s(0) - v_s(0), u_p(0) - v_p(0))\|_{L^2}$, where the Strichartz estimate for the free propagator is applied to the difference of initial conditions. In other words,

$$\|(u_s, u_p) - (v_s, v_p)\|_{q,r} \leq C\|(u_s(0) - v_s(0), u_p(0) - v_p(0))\|_{L^2}. \tag{2.47}$$

(3) *Convergence.* Finally, take a sequence of initial data in $H^s, s \geq s_0 \geq 2$, which converges in L^2 to a limit. For each member of this sequence, the solution satisfies the space-time bound (2.22). The right-hand side of (2.22) converges by convergence in L^2 of the data and the left-hand side of (2.22) converges by (2.47). Therefore, (2.22) holds on the interval of existence for any solution with L^2 data.

2.4. Existence in $H^1(\mathbb{R}^1_+)$. We consider the existence theory in H^1. In this section we prove the following theorem.

THEOREM 2.14. *Let $(u_s(0), u_p(0))$ be in H^1 and let the potential V satisfy* (2.5). *Then there exists a unique global solution for system* (1.2) *in $L^\infty(\mathbb{R}^1_+, H^1)$.*

We first observe that our proof of local existence, via the contraction mapping principle, extends to the space

$$\mathcal{X}_1(T) \equiv C([0,T], H^1) \cap \mathcal{X}(T). \tag{2.48}$$

In particular, one needs only to prove that A_β maps a ball to a ball in this smaller space and it is a contraction mapping there, for $T < T_2$, where $T_2 \leq T_1 \leq T_0$. This can be proven by applying $(H+I)^{\frac{1}{2}}P_c$ to the equations, using equivalence of norms (in the appropriate spaces): $\|(H+I)^{\frac{1}{2}}P_c\cdot\|_{L^2}$ and $\|\cdot\|_{H^1}$ and carrying out the standard energy estimates. We use $(H+I)P_c$ rather than $I-\Delta$ because functions of H commute with H and thus we avoid differentiation of the potential $V(x)$. Otherwise, we would require bounds on norms of $\partial_x V$.

If $T^*_{\max}$ denotes the maximal time of existence for the solution in $\mathcal{X}_1(T)$, then in view of the a priori estimates in $\mathcal{X}(T)$, global existence ($T^*_{\max} = \infty$) will follow from a priori bounds on (u_s, u_p) in H^1.

Let

$$\mathcal{A}_c \equiv (I + H)P_c. \tag{2.49}$$

Applying to system (2.15)–(2.16) operator $\mathcal{A}_c^{1/2}$, we obtain the inequality

$$\frac{\partial}{\partial t}\int (|\mathcal{A}_c^{1/2}u_s|^2 + |\mathcal{A}_c^{1/2}u_s|^2)dx \tag{2.50}$$

$$\leq 2\left|\int \left(\overline{\mathcal{A}_c^{1/2}u_s}\mathcal{A}_c^{1/2}(|u_p|^2u_s) - \overline{\mathcal{A}_c^{1/2}u_p}\mathcal{A}_c^{1/2}(|u_s|^2u_p)\right)dx\right|. \tag{2.51}$$

PROPOSITION 2.15. *Assume that V satisfies condition* (2.5). *Then the operator $\mathcal{A}_c^{1/2}(I-\Delta)^{-1/2}$ and its inverse are bounded in L^2; that is, for any $f \in L^2$ both $\mathcal{A}_c^{1/2}(I-\Delta)^{-1/2}f$ and $(I-\Delta)^{1/2}\mathcal{A}_c^{-1/2}f$ are bounded in L^2 and*

$$\|\mathcal{A}_c^{1/2}(I-\Delta)^{-1/2}f\|_{L^2} \le C\|f\|_{L^2}, \tag{2.52}$$

$$\|(I-\Delta)^{1/2}\mathcal{A}_c^{-1/2}f\|_{L^2} \le C\|f\|_{L^2}. \tag{2.53}$$

Proof. This proposition states that the $\|(I-\Delta)^{\frac{1}{2}}\|_{L^2}$ norm and $\|\mathcal{A}_c^{1/2}\cdot\|_{L^2}$ are equivalent. Then our strategy will be similar to the proof in the potential-free case (as if $\mathcal{A}_c^{1/2}$ were ∂_x).

We prove the proposition using Weder's result on the continuity of wave operators [6]. Under the conditions stated above, Weder proves that there exists wave operator Ω such that

$$\Omega(I-\Delta)\Omega^* = \mathcal{A}_c,$$

where Ω is a bounded continuous operator on H^1. Then, taking the square root, we obtain

$$\Omega(I-\Delta)^{-1/2}\Omega^* = \mathcal{A}_c^{-1/2}. \tag{2.54}$$

The square root exists since $\mathcal{A}_c = (I+H)P_c$ is a positive operator on the subspace corresponding to the continuous spectrum.

Now it is easy to verify (2.52):

$$\begin{aligned}\|\mathcal{A}_c^{1/2}(I-\Delta)^{-1/2}f\|_{L^2} &= \|\Omega(I-\Delta)^{1/2}\Omega^*(I-\Delta)^{-1/2}f\|_{L^2}\\ &= \|(I-\Delta)^{1/2}\Omega^*(I-\Delta)^{-1/2}f\|_{L^2} \le C\|\Omega^*(I-\Delta)^{-1/2}f\|_{H^1}\\ &\le C\|(I-\Delta)^{-1/2}f\|_{H^1} \le C\|f\|_{L^2},\end{aligned}$$

where we have used that Ω, Ω^* are isometries in L^2 and continuous in H^1. The other inequality (2.53) can be proved similarly. □

COROLLARY 2.16. *Let $f \in H^1 \cap \mathrm{Range}(P_c)$. Then*

$$\|\mathcal{A}_c^{1/2}f\|_2 \le C\|(I-\Delta)^{1/2}f\|_2, \tag{2.55}$$

$$\|(I-\Delta)^{1/2}f\|_2 \le \|\mathcal{A}_c^{1/2}f\|_2. \tag{2.56}$$

Proof. Let $f = (I-\Delta)^{1/2}g$ in (2.52), with $g \in H^1 \cap \mathrm{Range}(P_c)$. Then we have

$$\|\mathcal{A}_c^{1/2}g\|_{L^2} \le \|(I-\Delta)^{1/2}g\|_{L^2}.$$

To prove the other inequality (2.56), we write

$$\|(I-\Delta)^{1/2}f\|_{L^2} = \|(I-\Delta)^{1/2}\mathcal{A}_c^{-1/2}\mathcal{A}_c^{1/2}f\|_{L^2} \le \|\mathcal{A}_c^{1/2}f\|_{L^2}. \qquad \square$$

Proof of Theorem 2.14. First, using the above proposition, we estimate

$$\begin{aligned}\left|\int \left(\overline{\mathcal{A}_c^{1/2}u_s}\mathcal{A}_c^{1/2}(|u_p|^2u_s)\right)dx\right| &\le \|\mathcal{A}_c^{1/2}u_s\|_2 \cdot \|\mathcal{A}_c^{1/2}(|u_p|^2u_s)\|_2\\ &= \|\mathcal{A}_c^{1/2}u_s\|_2 \cdot \|\mathcal{A}_c^{1/2}(I-\Delta)^{-1/2}(I-\Delta)^{1/2}(|u_p|^2u_s)\|_2\\ &\le \|\mathcal{A}_c^{1/2}u_s\|_2 \cdot \|\mathcal{A}_c^{1/2}(I-\Delta)^{-1/2}\|_{\mathcal{B}(L^2,L^2)} \cdot \|(I-\Delta)^{1/2}(|u_p|^2u_s)\|_2.\end{aligned}$$

Using Leibnitz formula for the fraction, see [7],

$$\|(I-\Delta)^{1/2}(fg)\|_2 \le \|f\|_\infty \cdot \|(I-\Delta)^{1/2}g\|_2 + \|(I-\Delta)^{1/2}f\|_2 \cdot \|g\|_\infty,$$

we obtain

$$\|(I-\Delta)^{1/2}(|u_p|^2u_s)\|_2 \le \|u_p\|_\infty^2\|(I-\Delta)^{1/2}u_s\|_2 + 2\|u_p\|_\infty\|u_s\|_\infty\|(I-\Delta)^{1/2}u_p\|_2.$$

Combining the last two estimates, we obtain

$$\begin{aligned}
&\left|\int \left(\overline{\mathcal{A}_c^{1/2}u_s}\mathcal{A}_c^{1/2}(|u_p|^2u_s)\right)\,dx\right| \\
&\quad\le C\|\mathcal{A}_c^{1/2}u_s\|_2 \cdot (\|u_p\|_\infty^2 + \|u_s\|_\infty^2)(\|(I-\Delta)^{1/2}u_p\|_2 + \|(I-\Delta)^{1/2}u_s\|_2) \\
&\quad\le C\|\mathcal{A}_c^{1/2}u_s\|_2 \cdot (\|u_p\|_\infty^2 + \|u_s\|_\infty^2)(\|(I-\Delta)^{1/2}P_cu_p\|_2 + \|(I-\Delta)^{1/2}\langle u_p,\phi\rangle\phi\|_2 \\
&\qquad\qquad +\|(I-\Delta)^{1/2}P_cu_s\|_2 + \|(I-\Delta)^{1/2}\langle u_s,\phi\rangle\phi\|_2) \\
&\quad\le C(\|\mathcal{A}_c^{1/2}u_s\|_2^2 + \|\mathcal{A}_c^{1/2}u_p\|_2^2 + 1)\cdot(\|u_p\|_\infty^2 + \|u_s\|_\infty^2).
\end{aligned}$$

Finally, adding the s and p components of the differential inequalities, we obtain

$$\partial_t(\|\mathcal{A}_c^{1/2}u_s\|_2^2 + \|\mathcal{A}_c^{1/2}u_p\|_2^2) \le C(\|u_p\|_\infty^2 + \|u_s\|_\infty^2)\cdot(\|\mathcal{A}_c^{1/2}u_s\|_2^2 + \|\mathcal{A}_c^{1/2}u_p\|_2^2 + 1)$$

which implies that

$$\begin{aligned}
&\|\mathcal{A}_c^{1/2}u_s\|_2^2 + \|\mathcal{A}_c^{1/2}u_p\|_2^2 \\
&\qquad\le \exp\left(C\int_0^T [\|u_s\|_{L^\infty}^2 + \|u_p\|_{L^\infty}^2]\right)(\|\mathcal{A}_c^{1/2}u_s(0)\|_2^2 + \|\mathcal{A}_c^{1/2}u_p(0)\|_2^2 + 1).
\end{aligned}$$

Applying Hölder's inequality to the time integral in the exponent we have

$$\begin{aligned}
\int_0^T [\|u_s\|_{L^\infty}^2 + \|u_p\|_{L^\infty}^2]\,dt &\le C\|u_s\|_{L^4([0,T],L^\infty)}^2T^{\frac12} + C\|u_p\|_{L^4([0,T],L^\infty)}^2T^{\frac12} \\
&= C(\|u_s\|_{L^4([0,T],L^\infty)}^2 + \|u_p\|_{L^4([0,T],L^\infty)}^2)T^{\frac12} \\
&\le C(\mathcal{P}_0 + \mathcal{P}_0^{\frac12})(T^{\frac14} + T^{\frac34})T^{\frac12}.
\end{aligned}$$

The last inequality follows from the a priori space-time estimate of Theorem 2.8 and the fact that $(4,\infty)$ is an admissible pair. We, thus, establish the boundedness of (u_s,u_p) in $\|\mathcal{A}_c(\cdot)\|_2$ norm:

$$\|\mathcal{A}_c^{1/2}u_s\|_2^2 + \|\mathcal{A}_c^{1/2}u_p\|_2^2 \le K_1e^{K_2(\mathcal{P}_0+1)(T^2+1)}(\|\mathcal{A}_c^{1/2}u_s(0)\|_2^2 + \|\mathcal{A}_c^{1/2}u_p(0)\|_2^2 + 1).$$

Therefore, using the equivalence of norms, see Corollary 2.16, we obtain

$$\|u_s(t)\|_{H^1} + \|u_p(t)\|_{H^1} \le \tilde{K}_1e^{K_2(\mathcal{P}_0+1)(T^2+1)}(\|u_s(0)\|_{H^1} + \|u_p(0)\|_{H^1} + 1) \tag{2.57}$$

for some $K_1, K_2 > 0$. This completes the proof of global existence in H^1. ∎

3. Energy transfer from the guided mode to radiation modes. In this section, we prove that, over time scales of interest ($t \leq \mathcal{O}(\epsilon^{-1})$), radiation terms remain small during the amplification process and the finite-dimensional model (1.3) is a valid approximation. In the discussion of this section we return to the time-scale, where nonlinear terms are of order ϵ:

$$i\partial_t u_s - Hu_s = i\epsilon|u_p|^2 u_s, \tag{3.1}$$

$$i\partial_t u_p - Hu_p = -i\epsilon|u_s|^2 u_p. \tag{3.2}$$

For this system, we are going to show that radiation is indeed bounded by $C\epsilon$ on a time scale of order $1/\epsilon$.

To proceed, we first orthogonally decompose a solution of (3.1)–(3.2) into its bound state and continuous spectral (radiative) parts:

$$u_s(x,t) = a_s(t)\phi(x) + U_s(x,t), \tag{3.3}$$

$$u_p(x,t) = a_p(t)\phi(x) + U_p(x,t). \tag{3.4}$$

We prove the following theorem.

THEOREM 3.1. *Let $(u_s(0), u_p(0)) \in H^1$ and $P_c u_s(0) = P_c u_p(0) = 0$. Assume that $0 < \epsilon < \epsilon_0 < \infty$ and that V satisfies (2.5). Then, for any $T > 0$ there exists $C(T, \epsilon_0)$ so that*

$$\max\{\|U_s(t)\|_{H^1}, \|U_p(t)\|_{H^1}\} \leq C(T,\epsilon_0)\epsilon \tag{3.5}$$

on the interval $t \in [0, T/\epsilon]$.

We begin the proof with the following proposition, which follows from the ϵ-independent bounds $\|u_{s,p}\|_{H^1} \leq C(T,\epsilon_0)$, (2.57).

PROPOSITION 3.2. *Let $0 < \epsilon < \epsilon_0 < \infty$. Then for any $T > 0$ there exists $C(T,\epsilon_0)$ such that*

$$\|U_s\|_{H^1}, \|U_p\|_{H^1}, |a_s|, |a_p| \leq C \tag{3.6}$$

on the interval $t \in [0, T/\epsilon]$.

Substitution of (3.3)–(3.4) into (3.1)–(3.2) and projection onto ϕ and the range of P_c gives

$$\begin{aligned} i\partial_t a_s - \lambda a_s = i\epsilon[&|a_p|^2 a_s\langle\phi^3|\phi\rangle + \overline{a_p}a_s\langle\phi^3|U_p\rangle + a_p a_s\langle\phi^3|\overline{U_p}\rangle \\ &+ |a_p|^2\langle\phi^3|U_s\rangle + \cdots + \langle|U_p|^2 U_s|\phi\rangle], \end{aligned} \tag{3.7}$$

$$\begin{aligned} i\partial_t a_p - \lambda a_p = -i\epsilon[&|a_s|^2 a_p\langle\phi^3|\phi\rangle + \overline{a_s}a_p\langle\phi^3|U_s\rangle + a_p a_s\langle\phi^3|\overline{U_s}\rangle \\ &+ |a_s|^2\langle\phi^3|U_p\rangle + \cdots + \langle|U_s|^2 U_p|\phi\rangle], \end{aligned} \tag{3.8}$$

$$\begin{aligned} i\partial_t U_s - HU_s = i\epsilon[&|a_s||a_p|^2 P_c\phi^3 + \overline{a_p}a_s P_c\phi^2 U_p + a_p a_s P_c\phi^2\overline{U_p} \\ &+ |a_p|^2 P_c\phi^2 U_s + \cdots + P_c|U_p|^2 U_s], \end{aligned} \tag{3.9}$$

$$\begin{aligned} i\partial_t U_p - HU_p = -i\epsilon[&a_p|a_s|^2 P_c\phi^3 + \overline{a_s}a_p P_c\phi^2 U_s + a_p a_s P_c\phi^2\overline{U_s} \\ &+ |a_s|^2 P_c\phi^2 U_p + \cdots + P_c|U_s|^2 U_p], \end{aligned} \tag{3.10}$$

where $H = -\partial_x^2 + V(x)$ and $H\phi = \lambda\phi$.

COROLLARY 3.3. *The fundamental modes are slowly varying with the rate ϵ*

$$|\partial_t|a_s\||, |\partial_t|a_p\|| \leq C(T,\epsilon_0)\epsilon.$$

Proof. This follows from Proposition 3.2 and (3.7)-(3.8). Indeed,

$$|\partial_t|a_s\|| = |\partial_t|ia_s e^{it\lambda}\|| \le |\partial_t i(a_s e^{it\lambda})| = |(i\partial_t - \lambda)a_s| \le C(T,\epsilon_0)\epsilon.$$

The same argument leads to a similar estimate for $|\partial_t|a_p\||$. This ends the proof of the corollary. □

The following estimates are used in the proof of the theorem and can be easily verified.

LEMMA 3.4.

$$\begin{aligned}
&\|P_c f\|_2 \le \|f\|_2,\\
&\|P_c f\|_\infty \le \|f\|_\infty(1+\|\phi\|_1\cdot\|\phi\|_\infty),\\
&\|(H-\lambda)^{-1}e^{i(H-\lambda)t}P_c f\|_2 \le C\|f\|_2,\\
&\|(H-\lambda)^{-1}P_c\|_{H^1} \le \frac{C}{\operatorname{dist}(H_c,\lambda)} \le \frac{C}{|\lambda|}.
\end{aligned} \tag{3.11}$$

Proof of Theorem 3.1. We now make transformations $a_s = e^{-i\lambda t}A_s$ and $U_s = \epsilon e^{-iHt}W_s$ to remove rapid oscillations and explicitly show the smallness of radiation. By hypothesis of Theorem 3.1, we have $\|W_{s,p}(0)\|_{H^1} \le C$. Note that by the bounds of Proposition 3.2 we have $\|W_{s,p}\|_{H^1} \le C(T,\epsilon_0)/\epsilon$ and $|A_{s,p}| \le C(T,\epsilon_0)$.

The slowly varying amplitudes A_s, A_p satisfy

$$\begin{aligned}
\partial_t A_s = {}& \epsilon|A_p|^2 A_s\langle\phi^3|\phi\rangle + \epsilon^2\overline{A_p}A_s\langle\phi^3|e^{-i(H-\lambda)t}W_p\rangle + \epsilon^2 A_p A_s\langle\phi^3|e^{i(H-\lambda)t}\overline{W_p}\rangle\\
&+\epsilon^2|A_p|^2\langle\phi^3|e^{-i(H-\lambda)t}W_s|\phi\rangle + \cdots + \epsilon^4\langle|e^{-iHt}W_p|^2 e^{-i(H-\lambda)t}W_s\rangle,
\end{aligned} \tag{3.12}$$

$$\begin{aligned}
\partial_t A_p = {}& -\epsilon|A_s|^2 A_p\langle\phi^3|\phi\rangle - \epsilon^2\overline{A_s}A_p\langle\phi^3|e^{-i(H-\lambda)t}W_s\rangle - \epsilon^2 A_p A_s\langle\phi^3|e^{i(H-\lambda)t}\overline{W_s}\rangle\\
&-\epsilon^2|A_s|^2\langle\phi^3|e^{-i(H-\lambda)t}W_p\rangle - \cdots - \epsilon^4\langle|e^{-iHt}W_s|^2 e^{-i(H-\lambda)t}W_p|\phi\rangle.
\end{aligned} \tag{3.13}$$

Further, $W_{s,p}$ satisfy

$$\begin{aligned}
\partial_t W_s = {}& e^{i(H-\lambda)t}A_s|A_p|^2 P_c\phi^3\\
&+e^{iHt}[\epsilon\overline{A_p}A_s P_c\phi^2 e^{-iHt}W_p + \epsilon A_p A_s P_c\phi^2 e^{iHt}\overline{W_p}\\
&\quad+\epsilon|A_p|^2 P_c\phi^2 e^{-iHt}W_s + \cdots + \epsilon^3 P_c|e^{-iHt}W_p|^2 e^{-iHt}W_s],
\end{aligned} \tag{3.14}$$

$$\begin{aligned}
\partial_t W_p = {}& e^{i(H-\lambda)t}A_p|A_s|^2 P_c\phi^3\\
&+e^{iHt}[\epsilon\overline{A_s}A_p P_c\phi^2 e^{-iHt}W_s + \epsilon A_p A_s P_c\phi^2 e^{iHt}\overline{W_s}\\
&\quad+\epsilon|A_s|^2 P_c\phi^2 e^{-iHt}W_p + \cdots + \epsilon^3 P_c|e^{-iHt}W_s|^2 e^{-iHt}W_p].
\end{aligned} \tag{3.15}$$

The goal is now to show that given initial data where W is $\mathcal{O}(1)$ (which corresponds to radiation $\mathcal{O}(\epsilon)$) during the evolution W will remain $\mathcal{O}(1)$ on time interval $\mathcal{O}(1/\epsilon)$.

In order to do this we integrate the above equations:

$$W_s(t) = W_s(0) + \int_0^t e^{i(H-\lambda)s}A_s|A_p|^2 P_c\phi^3 ds + \epsilon\int_0^t R_s ds, \tag{3.16}$$

where ϵR_s is the ϵ-order part in (3.14), i.e., the second and the third lines. Integrating by parts

$$\begin{aligned}
W_s(t) = {}& W_s(0) + \frac{e^{i(H-\lambda)t}-1}{i(H-\lambda)}A_s|A_p|^2 P_c\phi^3\\
&-\int_0^t \frac{e^{i(H-\lambda)s}}{i(H-\lambda)}\partial_s(A_s|A_p|^2)P_c\phi^3 ds + \epsilon\int_0^t R_s ds
\end{aligned} \tag{3.17}$$

and applying $\|\mathcal{A}_c^{\frac{1}{2}} \cdot\|_{L^2}$, we obtain

$$\|\mathcal{A}_c^{\frac{1}{2}} W_s(t)\|_{L^2} \le \|\mathcal{A}_c^{\frac{1}{2}} W_s(0)\|_{L^2} + \left\| \frac{e^{i(H-\lambda)t}-1}{i(H-\lambda)} A_s |A_p|^2 \mathcal{A}_c^{\frac{1}{2}} P_c \phi^3 \right\|_{L^2} \tag{3.18}$$

$$+ C(T,\epsilon_0)\epsilon \int_0^t \left\| \frac{e^{i(H-\lambda)t}-1}{i(H-\lambda)} \mathcal{A}_c^{\frac{1}{2}} P_c \phi^3 \right\|_{L^2} \tag{3.19}$$

$$+ \epsilon \int_0^t \|\mathcal{A}_c^{\frac{1}{2}} R_s\|_{L^2} ds.$$

Therefore, we have

$$\|\mathcal{A}_c^{\frac{1}{2}} W_s(t)\|_{L^2} \le \|\mathcal{A}_c^{\frac{1}{2}} W_s(0)\|_{L^2} + C\|(H-\lambda)^{-1} \mathcal{A}_c^{\frac{1}{2}} P_c \phi^3\|_{L^2} \tag{3.20}$$

$$+ \epsilon \int_0^t \Big(\|\mathcal{A}_c^{\frac{1}{2}} P_c \phi^2 W_p\|_{L^2} \tag{3.21}$$

$$+ \cdots + \epsilon^2 \|\mathcal{A}_c^{\frac{1}{2}} P_c |e^{-iHt} W_p|^2 e^{-iHt} W_s\|_{L^2} \Big) ds.$$

The terms on the right-hand side in the first line are bounded by a constant. To estimate the other terms we use the above properties of $\mathcal{A}_c$, $(H-\lambda)^{-1}$, etc. We illustrate how one proceeds with the estimates using the last term:

$$\epsilon^2 \|\mathcal{A}_c^{\frac{1}{2}} P_c |e^{-iHt} W_p|^2 e^{-iHt} W_s\|_{L^2} \tag{3.22}$$

$$\le \epsilon^2 \|(I-\Delta)^{\frac{1}{2}} |e^{-iHt} W_p|^2 e^{-iHt} W_s\|_{L^2}$$

$$\le \epsilon^2 \|(I-\Delta)^{\frac{1}{2}} e^{-iHt} W_p\|_{L^2}^2 \cdot \|(I-\Delta)^{\frac{1}{2}} e^{-iHt} W_s\|_{L^2} \tag{3.23}$$

$$\le \epsilon^2 \|\mathcal{A}_c^{\frac{1}{2}} e^{-iHt} W_p\|_{L^2}^2 \cdot \|\mathcal{A}_c^{\frac{1}{2}} e^{-iHt} W_s\|_{L^2}$$

$$\le \epsilon^2 \|\mathcal{A}_c^{\frac{1}{2}} W_p\|_{L^2}^2 \cdot \|\mathcal{A}_c^{\frac{1}{2}} W_s\|_{L^2} \tag{3.24}$$

$$\le \epsilon^2 \|W_p\|_{H^1}^2 \cdot \|\mathcal{A}_c^{\frac{1}{2}} W_s\|_{L^2} \le C \|\mathcal{A}_c^{\frac{1}{2}} W_s\|_{L^2},$$

where we used equivalence of norms, Leibnitz rule, and the uniform bound $\|W_{s,p}\|_{H^1} \le C/\epsilon$. Thus, the inequality takes the form

$$\|\mathcal{A}_c^{\frac{1}{2}} W_s(t)\|_{L^2} \le B + \epsilon K \int_0^t (\|\mathcal{A}_c^{\frac{1}{2}} W_p\|_{L^2} + \cdots + \|\mathcal{A}_c^{\frac{1}{2}} W_s\|_{L^2}) ds, \tag{3.25}$$

where B and K do not depend on $\epsilon < \epsilon_0$. Adding the last inequality with the similar one for the p-component, and then using the notation $z(t) = \|\mathcal{A}_c^{\frac{1}{2}} W_s(t)\|_{L^2} + \|\mathcal{A}_c^{\frac{1}{2}} W_p(t)\|_{L^2}$, we obtain the inequality with modified B and K (but still independent of ϵ):

$$z(t) \le B + \epsilon K \int_0^t z(s) ds. \tag{3.26}$$

Using the standard Gronwall's result, we find

$$z(t) \le B e^{\epsilon K t} \Rightarrow z(t) \le B e^{KT}, \tag{3.27}$$

which proves the bound $\|W_s\|_{H^1} \le C(T,\epsilon_0)$. □

Remark 3.5. Using dispersive properties of e^{itH}, it is possible to establish smallness of radiation in weaker spaces, namely, in $\|\cdot\|_{L^\infty}$ norm. Taking (3.9)–(3.10) for $U_{s,p}$ and rewriting them in the integral form, we are led to estimate the terms

$$\epsilon \int_0^t e^{itH} a_s |a_p|^2 P_c \phi^3 ds \tag{3.28}$$

and

$$\epsilon \int_0^t e^{itH} R_s ds. \tag{3.29}$$

After some changes of variables with the aid of standard L^∞ decay estimates for the Schrödinger evolution, one obtains that

$$\|U_{s,p}\|_{L^\infty} \le C\sqrt{\epsilon}. \tag{3.30}$$

This argument also extends to the two-dimensional case with even better decay in ϵ (see the end of section 4.4).

4. Two-dimensional problem. We now consider the Raman system in the case of two transverse spatial dimensions

$$i\partial_t u_s - \beta H u_s = i|u_p|^2 u_s, \tag{4.1}$$

$$i\partial_t u_p - \beta H u_p = -i|u_s|^2 u_p, \tag{4.2}$$

where $H = -\Delta + V(x,y)$ and we prove analogous existence results and energy transfer estimates. Our strategy in the two-dimensional case is similar to the one-dimensional case; therefore, we omit some calculations which can be found in the previous sections.

In the two-dimensional case we require stronger conditions on the potential.

Assumption 4.1. The potential $V(x,y)$ is twice differentiable and

$$|D^\alpha V| \le C_\alpha (1 + x^2 + y^2)^{-a},$$

where $a > 6$ and $|\alpha| \le 2$.

Assumption 4.2. We assume that potential $V(x,y)$ has no zero energy eigenvalues or resonances.[3]

These assumptions are required to obtain space-time estimates in the next section.

4.1. Space-time estimates for the propagator. The definition of admissible pair is modified: (q,r) is admissible (in dimension $n=2$) if

$$\frac{1}{q} = \frac{1}{2} - \frac{1}{r}, \quad r \in [2,\infty].$$

THEOREM 4.3. *Assume that the potential V satisfies both assumptions and let (q,r) be an admissible pair. Then for any $f \in L^2$ we have that $U_0(t)f$ and $U(t)P_c f$ are in $L^q(\mathbb{R}, L^r)$ and*

$$\|U_0(\cdot)f\|_{L^q([0,T],L^r)} \le C\|f\|_{L^2}, \tag{4.3}$$
$$\|U(\cdot)P_c f\|_{L^q([0,T],L^r)} \le C\|f\|_{L^2},$$

[3]Zero eigenvalues and resonances are obstructions to the optimal time-decay estimates for e^{-iHt}. Their absence holds generically; see, for example, [10].

where C depends only on q.

THEOREM 4.4. *Assume that the potential V satisfies both assumptions, let (γ, ρ) be an admissible pair, and let $f \in L^{\gamma'}([0,T], L^{\rho'})$, where (γ', ρ') is conjugate to (γ, ρ). Then for any admissible pair (q, r)*

$$\left\| \int_0^t U_0(t-\tau) f(\tau) d\tau \right\|_{L^q([0,T],L^r)} \le C \|f\|_{L^{\gamma'}([0,T],L^{\rho'})},$$

$$\left\| \int_0^t U(t-\tau) P_c f(\tau) d\tau \right\|_{L^q([0,T],L^r)} \le C \|f\|_{L^{\gamma'}([0,T],L^{\rho'})}, \tag{4.4}$$

where C depends only on q, γ.

Both Theorems 4.3 and 4.4 are proven by using Yajima's results [10] on $W^{k,p}$ continuity of wave operators. The argument follows the proof of Proposition 2.15 and is omitted here. Finally, Corollary 2.3 is valid in the current setting without any changes as the scaling is independent of the space dimension.

Remark 4.5. The application of Yajima's results on $W^{k,p}$ continuity of wave operators is the origin of the more restrictive smoothness assumptions on the potential $V(x)$. In one space dimension smoothness of $V(x)$ is not required [6].

4.2. A priori space-time estimates. The same argument as in the one-dimensional case applies here and we obtain the a priori bound of Proposition 2.5, as well as the bound (2.28) on nonlinear terms.

THEOREM 4.6 (a priori bounds in $L^q([0,T], L^r)$). *Let (q, r) be an admissible pair. Then any solution (u_s, u_p) satisfies the bounds*

$$\|u_s\|_{L^q([0,T],L^r)} \le C(\mathcal{P}_0 + 1)(T+1), \tag{4.5}$$

$$\|u_p\|_{L^q([0,T],L^r)} \le C(\mathcal{P}_0 + 1)(T+1). \tag{4.6}$$

Proof. As in the one-dimensional case, this estimate is proved by a straightforward application of space-time estimate for Schrödinger equation with a potential and using a priori estimate on nonlinear terms. Following the proof of Theorem 2.8, we obtain the same bounds with different ρ, ρ', γ, γ'. We must impose $\rho' = 1$ with $\rho = \infty$, but $\gamma = 2$ (since in $1/\gamma = 1/2 - 1/\rho$) with $\gamma' = 2$. This results in

$$\begin{aligned} \|u_s\|_{L^q([0,T],L^r)} &\le C_1 \|u_s(0)\|_{L^2} + C_2 \|u_p^2 u_s\|_{L^2([0,T],L^1)} \\ &\quad + \|\phi\|_{L^\infty} \|\phi\|_{L^r} T^{1/2+1/q} \|u_p^2 u_s\|_{L^2([0,T],L^1)} \\ &\le C_1 \mathcal{P}_0^{\frac{1}{2}} + C_2 \mathcal{P}_0 + C_3 \mathcal{P}_0 (T + T^{\frac{1}{2}}), \end{aligned}$$

since $q \ge 2$. ∎

4.3. Local existence in $H^2(\mathbb{R}^2)$. We now prove local existence of solutions in H^2 and will extend it to a global solution using our space-time estimates.

THEOREM 4.7 (local existence).

(1) *Given initial data $(u_s(0), u_p(0)) \in H^2(\mathbb{R}^2)$, there exist $T > 0$ and a unique solution $(u_s, u_p) \in L^\infty([0,T], H^2(\mathbb{R}^2))$.*

(2) *Let $T_{\max} > 0$ denote the maximal time of existence. Then, either $T_{\max} = \infty$ (global existence in time) or*

$$T_{\max} < \infty \quad \text{and} \quad \limsup_{t \to T_{\max}} \|(u_s, u_p)\|_{L^\infty([0,T],H^2(\mathbb{R}^2))} = \infty.$$

To prove the local existence theorem we have to show that a ball in H^2 is mapped into itself and that the mapping is a contraction.

Consider the same mapping as in the one-dimensional case (2.38)–(2.39). We will first show that it maps a ball into a ball:

$$\|A_\beta^{(s)}(u_s,u_p)\|_{H^2} \le \|U_\beta(t)u_s(0)\|_{H^2} + \left\|\int_0^t U_\beta(t-\tau)|u_p|^2u_s\right\|_{H^2}. \tag{4.7}$$

The first term is estimated as follows:

$$\begin{aligned}\|U_\beta(t)u_s(0)\|_{H^2} &= \|(I-\Delta)U_\beta(t)u_s(0)\|_{L^2} = \|(I-\Delta)(H+i)(H+i)^{-1}U_\beta(t)u_s(0)\|_{L^2}\\ &\le \|(I-\Delta)(H+i)^{-1}\|_{\mathcal{B}(L^2,L^2)}\|(H+i)U_\beta(t)u_s(0)\|_{L^2}.\end{aligned}$$

Note that the operator $(I-\Delta)(H+i)^{-1}$ is bounded in L^2-operator norm. This follows from the identity

$$(I-\Delta)(H+i)^{-1} = -I + (I+i-V)(H+i)^{-1}$$

and the boundedness of V in L^∞ and of $(H+i)^{-1}$ in L^2.

Next, we have to establish the bound for

$$\begin{aligned}\|(H+i)U_\beta(t)u_s(0)\|_{L^2} &= \|(H+i)u_s(0)\|_{L^2} = \|(H+i)(I-\Delta)^{-1}(I-\Delta)u_s(0)\|_{L^2}\\ &\le \|(H+i)(I-\Delta)^{-1}\|_{\mathcal{B}(L^2,L^2)}\|(I-\Delta)u_s(0)\|_{L^2} \le C\|u_s(0)\|_{H^2},\end{aligned}$$

where the operator $(H+i)(I-\Delta)^{-1}$ is bounded by similar calculations as for $(I-\Delta)(H+i)^{-1}$.

Now, we estimate the second term in (4.7):

$$\left\|\int_0^t U_\beta(t-\tau)|u_p|^2u_s\,d\tau\right\|_{H^2} \le C\left\|\int_0^t (H+i)U_\beta(t-\tau)|u_p|^2u_s\,d\tau\right\|_{L^2}.$$

To bound this term we write it in the form

$$\int dx\left\{\int_0^t\int_0^t (H+i)U_\beta(t-\tau_1)|u_p(\tau_1)|^2u_s(\tau_1)(H-i)\overline{U_\beta}(t-\tau_2)|u_p(\tau_2)|^2\overline{u_s}(\tau_2)d\tau_1d\tau_2\right\}$$

(using Hölder inequality and isometry of U_β in L^2)

$$\begin{aligned}&\le \int_0^t\int_0^t d\tau_1d\tau_2\left|\int |(H+i)|u_p(\tau_1)|^2u_s(\tau_1)|^2\,dx\right|^{\frac12}\left|\int |(H+i)|u_p(\tau_2)|^2u_s(\tau_2)|^2\,dx\right|^{\frac12}\\ &\quad\le t^2\sup_{\tau\in[0,t]}\|(H+i)u_p^2(\tau)u_s(\tau)\|_2 \le Ct^2\sup_{\tau\in[0,t]}\|u_p^2(\tau)u_s(\tau)\|_{H^2}^2\\ &\quad\le Ct^2\sup_{\tau\in[0,t]}\|u_p(\tau)\|_{H^2}^4\|u_s(\tau)\|_{H^2}^2.\end{aligned}$$

Therefore, we finally obtain the bound on the s-part of the map

$$\|A_\beta^{(s)}(u_s,u_p)\|_{H_2} \le C\|u_s(0)\|_{H^2} + Ct\sup_{\tau\in[0,t]}\|u_p(\tau)\|_{H^2}^2\|u_s(\tau)\|_{H^2}$$

and the full map

$$\begin{aligned}\|A_\beta(u_s,u_p)\|_{L^\infty([0,t],H^2)} &\le C\left(\|u_s(0)\|_{H^2} + \|u_p(0)\|_{H^2}\right)\\ &\quad + Ct\big(\|u_p\|_{L^\infty([0,t],H^2)}^2\|u_s\|_{L^\infty([0,t],H^2)}\\ &\qquad + \|u_p\|_{L^\infty([0,t],H^2)}\|u_s\|_{L^\infty([0,t],H^2)}^2\big).\end{aligned}$$

With the obtained inequality it is easy to establish the following proposition.

PROPOSITION 4.8. *Let $(u_s(0), u_p(0))$ be in H^2. Define the ball in $L^\infty([0,T], H^2)$ as*

$$\mathcal{B}(T) = \big\{(u_s, u_p) \in L^\infty([0,T], H^2) : \|(u_s, u_p)\|_{L^\infty([0,T],H^2)} \\ \le 2C(\|u_s(0)\|_{H^2} + \|u_p(0)\|_{H^2})\big\}.$$

Then there exists $T_0 > 0$ such that for any $T < T_0$, the ball is mapped into itself, i.e., $A_\beta(\mathcal{B}(T)) \subset \mathcal{B}(T)$ with T_0 independent of β.

Next, we have to show that the mapping is a contraction in $L^\infty([0,T], H^2)$. Consider the first component of the map applied to two different pairs (u_s, u_p) and (v_s, v_p) with the same initial data $(u_s(0), u_p(0)) = (v_s(0), v_p(0))$:

$$\begin{aligned}
&\|A^{(s)}_\beta(u_s, u_p) - A^{(s)}_\beta(v_s, v_p)\|_{L^\infty([0,t],H^2)} \\
&\quad \le CT \sup_{t\in[0,T]} \||u_p(t)|^2 u_s(t) - |v_p(t)|^2 v_s(t)\|_{H^2} \\
&\quad \le CT \sup_{t\in[0,T]} (\|u_p^2(u_s - v_s)\|_{H^2} + \|u_p v_s(u_p - v_p)\|_{H^2} + \|v_p v_s(u_p - v_p)\|_{H^2}) \\
&\quad \le C(\|u_p, u_s, v_p, v_s\|_{L^\infty([0,T],H^2)})T\|(u_p - v_p, u_s - v_s)\|_{L^\infty([0,T],H^2)}.
\end{aligned}$$

Adding both s and p components of the map we obtain

$$\begin{aligned}
&\|A_\beta(u_p, u_s) - A_\beta(v_p, v_s)\|_{L^\infty([0,T],H^2)} \\
&\qquad \le TC(\|u_p, u_s, v_p, v_s\|_{L^\infty([0,T],H^2)})\|(u_p - v_p, u_s - v_s)\|_{L^\infty([0,T],H^2)}.
\end{aligned}$$

This inequality implies the following proposition.

PROPOSITION 4.9. *There exists $T_1 : 0 < T_1 < T_0$ sufficiently small such that the map A_β is a contraction in the ball $\mathcal{B}(T)$ for any $T : 0 < T < T_1$:*

$$\|A_\beta(u_p, u_s) - A_\beta(v_p, v_s)\|_{L^\infty([0,T],H^2)} \le q\|(u_p - v_p, u_s - v_s)\|_{L^\infty([0,T],H^2)},$$

where $q < 1$.

Remark 4.10. Whenever there exists a local solution on $t \in [0,T]$, it is bounded in $L^q([0,T], L^r(\mathbb{R}^2))$. Indeed, a solution in $L^\infty([0,T], H^2)$ is also in $L^q([0,T], L^r(\mathbb{R}^2))$ and the earlier obtained a priori bounds apply.

4.4. Global existence in $H^2(\mathbb{R}^2)$. In this section we will show that the local solution obtained via the contraction mapping principle in the previous section can be extended to a global solution using space-time estimates. We start with establishing uniform bound in H^1 space.

A priori estimates in H^1. Proceeding as in the one-dimensional problem (2.4), we obtain H^1 bound. Since $(\infty, 2)$ is an admissible pair, we obtain that the solutions are even uniformly bounded in time on the interval of existence of a local solution:

$$\|(u_s, u_p)\|_{L^\infty([0,T],H^1)} \le C(\mathcal{P}_0, T)\|(u_s(0), u_p(0))\|_{H^1}. \tag{4.8}$$

Continuation to a global solution using Theorem 4.7 requires an H^2 estimate, which we now derive.

Global existence in H^2. Having established uniform bound in H^1, we are ready to obtain a bound in H^2, which will rule out the first alternative in Theorem 4.7 (on local existence) and, thus, leave the global existence as the only possibility.

Applying $H+I$ to the s component of (4.2) and using energy estimates, we obtain

$$\begin{aligned}\partial_t \int |(H+I)u_s|^2 &\le \int |(H+I)u_s| \cdot |(H+I)(|u_p|^2 u_s)| \\ &\le \|(H+I)u_s\|_2 \|(H+I)(|u_p|^2 u_s)\|_2. \end{aligned} \tag{4.9}$$

To estimate the last term, we write

$$\begin{aligned}\int |\Delta(|u_p|^2 u_s)|^2 &= \int |u_p|^4 |\Delta u_s|^2 + 2\int |\nabla u_p|^2 |\nabla u_s|^2 |u_p|^2 + \int |\nabla u_p|^4 |u_s|^2 \\ &\le C(\|u_p\|_\infty^4 \cdot \|u_s\|_{H^2}^2 + \|u_s\|_\infty^2 \cdot \|u_p\|_{H^2}^2 \cdot \|u_p\|_{H^1}^2 \\ &\quad + \|u_p\|_\infty^2 \cdot \|u_s\|_{H^2}^2 \cdot \|u_s\|_{H^1}^2 + \|u_p\|_\infty^2 \cdot \|u_p\|_{H^2}^2 \cdot \|u_p\|_{H^1}^2)\end{aligned}$$

and therefore we have

$$\int |\Delta(|u_p|^2 u_s)|^2 \le C\left(1 + \|u_p\|_\infty^4 + \|u_s\|_\infty^4\right) \|(u_s, u_p)\|_{H^1}^2 \cdot \|(u_s, u_p)\|_{H^2}^2.$$

Now, adding the s and p components, we obtain

$$\begin{aligned}&\partial_t \int (|(H+I)u_s|^2 + |(H+I)u_s|^2) \\ &\quad \le C(1 + \|u_p\|_\infty^2 + \|u_s\|_\infty^2)(\|(H+I)u_s\|_{H^2}^2 + \|(H+I)u_p\|_{H^2}^2),\end{aligned}$$

where we have used that $\|(u_s, u_p)\|_{H^1}$ is bounded (see previous paragraph on H^1 estimates).

Since $(2, \infty)$ is an admissible pair, both $\int \|u_p\|_\infty^2 \, dt$ and $\int \|u_s\|_\infty^2 \, dt$ are bounded, and we obtain the required bound:

$$\|(H+I)u_s\|_2^2 + \|(H+I)u_p\|_2^2 \le C(\mathcal{P}_0, T) e^{kT} (\|(H+I)u_s(0)\|_2^2 + \|(H+I)u_p(0)\|_2^2)$$

on the interval $t \in [0, T]$. Because of the norm equivalence, we also obtain

$$\|u_s(t)\|_{H^2}^2 + \|u_p(t)\|_{H^2}^2 \le C(\mathcal{P}_0, T) e^{kT} (\|u_s(0)\|_{H^2}^2 + \|u_p(0)\|_{H^2}^2).$$

This bound implies the global existence of solutions in H^2.

Radiation losses in two-dimensional amplification model. The analogue of Theorem 3.1 on the boundedness of radiation in H^2 holds with the proof carrying over from the one-dimensional case. One can also obtain boundedness in L^∞ as described in Remark 3.5 with even faster decay: $\epsilon \log \epsilon$ rather than $\sqrt{\epsilon}$.

Appendix A. Application to optical communications. In this section we provide the details on the Raman model in optical communication systems as well as the derivation of the reaction-dispersion system. In modern long-haul optical communication systems the signal propagates in the fundamental mode of a single mode fiber. In an "ideal" lossless optical fiber waveguide the transverse shape of the wave envelope does not change and there is no transfer of energy from the fundamental mode to radiation modes. However, in the real systems Rayleigh scattering causes

attenuation of signal power, thus requiring periodic amplification [1]. A current strategy for amplification of a signal is based on the stimulated Raman effect. Here, light of a second *pump* wavelength is co- or counter-propagated in the medium. The stimulated Raman effect is a parametric process in which light of the pump frequency is transferred to that of the signal frequency. This amplification process is inherently nonlinear and therefore is expected to cause deformation of the transverse mode shape. There are other linear and nonlinear effects which may need to be taken into account such as group velocity dispersion, self-phase modulation, and four wave mixing. Also, refractive index depends on the frequency shift between the pump and signal frequencies.

Regarding linear effects, like group velocity dispersion, in practice they are weaker compared to the Raman effect assuming that pulses are not too short. However, in our case, we consider a model problem with both pump and signal being continuous (constant amplitude) waves. Then dispersion just vanishes.

The refractive index depends on the light frequency. As a result fundamental modes would be slightly different for pump and signal waves. Here, we assume that the modes are the same as it simplifies the exposition. All our results can be obtained for the frequency-dependent dispersion/diffraction coefficients with minimal modifications.

Approximate equations for the Raman interaction of signal and pump in the waveguide have been derived in [3]. These authors derived a pair of coupled ODEs for the signal and pump intensities, based on the assumption that all energy is contained in the fundamental modes. This model compares well with experiment [3] (see also [1] and the references therein). The authors [3] also discussed why the approximation was so accurate. They suggested that radiative losses (energy transfer from bound to radiation modes) is negligible due to the fact that "the wave-guiding action of the fiber reforms the pump and Stokes waves so that they always have intensities distributions which are close approximations to those which would exist in the absence of Raman interaction." However, this explanation has some limitations as the energy transfer could occur adiabatically (e.g., like the ionization of an atom). In other words, a weak process may lead to non-negligible changes after sufficiently long time. In particular, one might expect that the modes would undergo continuous deformation while also shedding radiation, so that after the full energy exchange a non-negligible amount of energy would accumulate in radiative modes and would constitute a significant loss.

To understand these effects, equations which take into account the effects of diffraction, wave-guiding and amplification should be studied. Naturally, the model will contain a small parameter: the ratio of diffraction and amplification lengths.

Raman stimulated emission describes the amplification of signal photons (with frequency ω_s) with Stokes down shifted pump photons (with frequency ω_p) and is governed by [3]

$$\frac{\partial n_s}{\partial z} = g(\omega_p - \omega_s) n_s n_p, \tag{A.1}$$

$$\frac{\partial n_p}{\partial z} = -g(\omega_p - \omega_s) n_s n_p, \tag{A.2}$$

where n_s, n_p are the number densities of signal and pump photons, respectively, and the total number of photons per unit volume is conserved $n_s + n_p = N$. Since we wish to focus on the effects due to the resonant coupling of the two wave fields (pump and signal), we ignore other effects, such as amplified spontaneous emission [1] which is always present, though a small effect.

Introducing the intensities

$$I_s = h\omega_s n_s, \qquad I_p = h\omega_p n_p, \tag{A.3}$$

where h denotes Planck's constant, we obtain the corresponding equations

$$\frac{\partial I_s}{\partial z} = \frac{g(\omega_p - \omega_s)}{h\omega_p} I_s I_p, \tag{A.4}$$

$$\frac{\partial I_p}{\partial z} = -\frac{g(\omega_p - \omega_s)}{h\omega_s} I_s I_p. \tag{A.5}$$

These equations satisfy the photon number conservation relation

$$\frac{I_s}{\omega_s} + \frac{I_p}{\omega_p} = \text{constant}. \tag{A.6}$$

In the case of radiative loss, this conservation law would be violated, since some photons would be lost from the bound waveguide mode to radiation modes. Equations (A.4)–(A.5) describe the plane wave Raman interaction.

Consider now the propagation of light in a dielectric cylinder waveguide with longitudinal coordinate, z. Maxwell's equations [1] imply

$$\Delta \mathbf{E} - \frac{1}{c^2}\mathbf{E}_{tt} - \nabla(\nabla \cdot \mathbf{E}) = \frac{1}{c^2}[\chi^{(1)}(\mathbf{r}, t) * \mathbf{E}_{tt}]_{tt} = 0,$$

where $\mathbf{E} \in \mathbb{R}^3$ is the electric field and $\chi^{(1)}(\mathbf{r}, t)$ is the linear susceptibility. Neglecting vector effects (see, e.g., [1]), we find that the time Fourier transform of $\mathbf{E}$, $\hat{\mathbf{E}}$, satisfies

$$\Delta \hat{\mathbf{E}} + \frac{\omega^2 n^2(\mathbf{x}_\perp, \omega)}{c^2}\hat{\mathbf{E}} = 0.$$

Each component of $\mathbf{E}$ satisfies

$$E_{zz} + \Delta_\perp E + \frac{\omega^2 n^2(\mathbf{x}_\perp, \omega)}{c^2} E = 0. \tag{A.7}$$

Next, we introduce the paraxial approximation. Let δ be a small parameter, and assume the following structure for the refraction index dependence on $\mathbf{x}_\perp$:

$$n^2(\mathbf{x}_\perp, \omega) = n_0^2(\omega) + \delta^2 n_1^2(\delta \mathbf{x}_\perp/\lambda_0, \omega). \tag{A.8}$$

We also seek E, in the form

$$E = A((\delta/\lambda_0)\mathbf{x}_\perp, (\delta^2/\lambda_0)z)e^{ikz}, \tag{A.9}$$

where λ_0 is the light wavelength and

$$\frac{2\pi}{\lambda_0} = k = \omega n_0(\omega)/c.$$

Thus, E varies more rapidly in the transverse than longitudinal directions.

Substituting (A.8), (A.9) into (A.7) and multiplying by $\lambda_0^2 \delta^{-4}$ we obtain

$$A_{ZZ} + 2ik\lambda_0\delta^{-2}A_Z + \delta^{-2}\Delta_\perp A + \delta^{-2}\lambda_0^2(\omega^2/c^2)n_1^2(\mathbf{X}_\perp) = 0,$$

where $Z = (\delta^2/\lambda_0)z$, $\mathbf{X}_\perp = (\delta/\lambda_0)\mathbf{x}_\perp$, and $(*)_\perp$ denotes differentiation with respect to $\mathbf{X}_\perp$. For δ small, we keep the dominant terms, those of order δ^{-2} and obtain, after using the relation between k and λ_0,

$$i4\pi A_Z + \Delta_\perp A + \frac{4\pi^2 n_1(\mathbf{X}_\perp, \omega)}{n_0^2} A = 0. \tag{A.10}$$

Equation (A.10) governs the linear propagation of any light field (signal or pump) in the paraxial approximation. To obtain a model governing the Raman interaction of signal and pump fields, u_s and u_p, we argue as follows. The signal field envelope propagates through a medium with refractive index (A.8) corrected by an imaginary term proportional to $i|u_p|^2$ corresponding to the Raman amplification by pump. The pump field envelope, u_p, propagates through a medium with refractive index (A.8) corrected by an imaginary term proportional to $-i|u_s|^2$ corresponding to pump depletion by the signal. The coupled signal and pump envelopes are then taken to satisfy the system

$$i\partial_z u_s + \Delta_\perp u_s - V(\omega_s, \mathbf{x}_\perp)u_s = i\epsilon_s|u_p|^2 u_s, \tag{A.11}$$

$$i\partial_z u_p + \Delta_\perp u_p - V(\omega_p, \mathbf{x}_\perp)u_p = -i\epsilon_p|u_s|^2 u_p, \tag{A.12}$$

where $\epsilon_{p,s}$ is the parameter which measures the ratio of the diffraction and nonlinear lengths. Usually, $\epsilon_{p,s}$ is very small [1]. We further assume[4] $\epsilon = \epsilon_s = \epsilon_p$ and neglect the dependence of the refractive index on frequency, i.e., $V(x_\perp, \omega) = V(x_\perp)$.

System (A.12) models the Raman energy exchange between the two continuous waves. We have not included the effects of losses due to the Rayleigh scattering in order not to burden the exposition. In reality, the Raman amplification length might be comparable to the effective (loss) length (20 km).

Thus, we have

$$\begin{aligned} i\partial_t u_s - Hu_s &= i\epsilon|u_p|^2 u_s, \\ i\partial_t u_p - Hu_p &= -i\epsilon|u_s|^2 u_p, \end{aligned} \tag{A.13}$$

where we use "t" to denote the "time-like" direction, z, $x_\perp = x$, and

$$H = -\Delta + V(x). \tag{A.14}$$

We study system (A.13) in the case where H has spectrum consisting of one point eigenvalue, $\lambda < 0$, with corresponding eigenfunction ϕ, $\|\phi\|_{L^2} = 1$. The components of u_s and u_p, which are orthogonal to ϕ, are called radiative components. Our goal is to prove if for $t = 0$ the order-one energy is concentrated in ϕ alone, then on time scales of order ϵ^{-1} the energy in radiative components is at most of order ϵ.

Appendix B. Normal form theorem. In this section we state a normal form result on the absence of the terms driving the radiation to any order of the perturbation theory. While in the main part of the paper we have used only the fact that these terms can be removed at the first order, we find this result important and potentially useful in the search for sharper estimates of radiation growth. The results in this section are formal in the sense that we do not verify the validity of transformations and of obtained systems.

[4]Nonlinear coefficient ϵ is then the same for both fields. Therefore, in this system the energy is conserved rather than photon number. This is done to simplify the presentation. All results hold for the "true" model as well.

For the Raman energy exchange system

$$i\partial_t u_s - Hu_s = i\epsilon|u_p|^2 u_s, \tag{B.1}$$

$$i\partial_t u_p - Hu_p = -i\epsilon|u_s|^2 u_p, \tag{B.2}$$

we now use representation

$$u_s = u_0^s\phi_0^s(x) + \int_0^\infty u_\lambda^s\phi_\lambda^s(x)d\lambda,$$
$$u_p = u_0^p\phi_0^p(x) + \int_0^\infty u_\lambda^p\phi_\lambda^p(x)d\lambda,$$

where H becomes diagonal

$$H\phi_0 = \lambda_0\phi_0,$$
$$H\phi_\lambda = \lambda\phi_\lambda, \qquad \langle u_\lambda, \phi_\lambda\rangle = u_\lambda,$$

where $\lambda_0 < 0$ corresponds to the fundamental mode and the remaining part of the spectrum ($\lambda > 0$) corresponds to the continuous spectrum. In this representation the equations take the form

$$\begin{aligned}
i\partial_t u_0^s - \lambda_0 u_0^s &= i\epsilon C_{00}^{00} u_0^p\overline{u}_0^p u_0^s + i\epsilon\int_0^\infty C_{\lambda_1 0}^{00} u_0^p\overline{u}_0^p u_{\lambda_1}^s d\lambda_1 \\
&\quad + \cdots + i\epsilon\int_0^\infty\int_0^\infty\int_0^\infty C_{\lambda_1 0}^{\mu_1\mu_2} u_{\mu_1}^p\overline{u}_{\mu_2}^p u_{\lambda_1}^s d\mu_1 d\mu_2 d\lambda_1, \\
i\partial_t u_0^p - \mu_0 u_0^p &= i\epsilon C_{00}^{00} u_0^s\overline{u}_0^s u_0^p + i\epsilon\int_0^\infty C_{00}^{\mu_1 0} u_0^s\overline{u}_0^s u_{\mu_1}^p d\mu_1 \\
&\quad + \cdots + i\epsilon\int_0^\infty\int_0^\infty\int_0^\infty C_{\lambda_1\lambda_2}^{\mu_1 0} u_{\lambda_1}^s\overline{u}_{\lambda_2}^s u_{\mu_1}^p d\lambda_1 d\lambda_2 d\mu_1, \\
i\partial_t u_\lambda^s - \lambda u_\lambda^s &= i\epsilon C_{0\lambda}^{00} u_0^p\overline{u}_0^p u_0^s + i\epsilon\int_0^\infty C_{\lambda_1\lambda}^{00} u_0^p\overline{u}_0^p u_{\lambda_1}^s d\lambda_1 \\
&\quad + \cdots + i\epsilon\int_0^\infty\int_0^\infty\int_0^\infty C_{\lambda_1\lambda}^{\mu_1\mu_2} u_{\mu_1}^p\overline{u}_{\mu_2}^p u_{\lambda_1}^s d\mu_1 d\mu_2 d\lambda_1, \\
i\partial_t u_\mu^p - \mu u_\mu^p &= i\epsilon C_{00}^{0\mu} u_0^s\overline{u}_0^s u_0^p + i\epsilon\int_0^\infty C_{00}^{\mu_1\mu} u_0^s\overline{u}_0^s u_{\mu_1}^p d\mu_1 \\
&\quad + \cdots + i\epsilon\int_0^\infty\int_0^\infty\int_0^\infty C_{\lambda_1\lambda_2}^{\mu_1\mu} u_{\lambda_1}^s\overline{u}_{\lambda_2}^s u_{\mu_1}^p d\lambda_1 d\lambda_2 d\mu_1,
\end{aligned}$$

where

$$C_{\lambda_1\lambda_2}^{\mu_1\mu_2} = \int_{-\infty}^{+\infty}\phi_{\mu_1}^p\overline{\phi}_{\mu_2}^p\phi_{\lambda_1}^s\overline{\phi}_{\lambda_2}^s\,dx.$$

The natural consequence of the stimulated emission process is the invariance with respect to the phase shifts of both the signal and the pump modes:

$$u_{\lambda_i}^s, u_{\mu_j}^p \to u_{\lambda_i}^s e^{i\psi_s}, u_{\mu_j}^p e^{i\psi_p}.$$

This torus action will be called G_{sp}-action below.

Consider the class of polynomial vector fields which stay invariant under this group action. We are going to consider near-identity transformations, which commute with the torus action. Therefore these transformations map a G_{sp}-invariant vector field to another G_{sp}-invariant vector field. We will now invoke these transformations to remove nonresonant terms from the equations. We first observe that a polynomial vector field that is invariant with respect to the G_{sp}-action is generated by the monomials of this form

$$\mathbf{e}^{\mathbf{s}}_{\lambda} u^p_{\mu_1} \overline{u}^p_{\mu_2} \cdots u^p_{\mu_{m-1}} \overline{u}^p_{\mu_m} u^s_{\lambda_1} \overline{u}^s_{\lambda_2} \cdots u^s_{\lambda_{k-1}} \overline{u}^s_{\lambda_k} u^s_{\lambda_{k+1}}, \tag{B.3}$$

$$\mathbf{e}^{\mathbf{p}}_{\mu} u^s_{\lambda_1} \overline{u}^s_{\lambda_2} \cdots u^s_{\lambda_{m-1}} \overline{u}^s_{\lambda_m} u^p_{\mu_1} \overline{u}^p_{\mu_2} \cdots u^p_{\mu_{k-1}} \overline{u}^p_{\mu_k} u^p_{\mu_{k+1}}, \tag{B.4}$$

where m and k are even numbers.

DEFINITION. *The monomial of type* (B.3) *is called resonant if*

$$\mu_1 - \mu_2 + \cdots + \mu_{m-1} - \mu_m + \lambda_1 - \lambda_2 + \cdots + \lambda_{k-1} - \lambda_k + \lambda_{k+1} - \lambda = 0$$

and the monomial of type (B.4) *is called resonant if*

$$\lambda_1 - \lambda_2 + \cdots + \lambda_{m-1} - \lambda_m + \mu_1 - \mu_2 + \cdots + \mu_{k-1} - \mu_k + \mu_{k+1} - \mu = 0.$$

If in a G_{sp}-invariant vector field initially all the energy is concentrated in fundamental modes, then the radiative modes can be excited only through the terms

$$\mathbf{e}^{\mathbf{s}}_{\lambda} |u^p_0|^k |u^s_0|^l u^s_0 \quad \text{and} \quad \mathbf{e}^{\mathbf{p}}_{\mu} |u^s_0|^k |u^p_0|^l u^p_0,$$

while no other radiation driving terms can appear after application of a G_{sp}-invariant transformation. These terms are nonresonant, since the corresponding arithmetic combinations are $\lambda - \lambda_0$ and $\mu - \lambda_0$, where $\lambda, \mu > 0$ and $\lambda_0 < 0$. Therefore, $|\lambda - \lambda_0| > |\lambda_0| > 0$ and $|\mu - \lambda_0| > |\lambda_0| > 0$.

According to the standard normal form procedure, these terms can be removed by employing the transformations of the form[5]

$$u^s_{\lambda} = U^s_{\lambda} + \epsilon^{\frac{k+l}{2}} C^s_{\lambda} |U^p_0|^k |U^s_0|^l U^s_0, \tag{B.5}$$

$$u^p_{\mu} = U^p_{\mu} + \epsilon^{\frac{k+l}{2}} C^p_{\mu} |U^s_0|^k |U^p_0|^l U^p_0. \tag{B.6}$$

Now we formulate the normal form theorem.

THEOREM B.1. *For any $N \geq 1$ there exists a sequence of transformations of the form* (B.5)–(B.6)*, which bring the system to the form*

$$\begin{aligned}
i\partial_t U^s_0 - \lambda_0 U^s_0 &= i \left[\epsilon C^{00}_{00} U^p_0 \overline{U}^p_0 U^s_0 + \sum_{n_1,n_2>0}^{n_1+n_2<N+1} \epsilon^{\frac{n_1+n_2}{2}} C^s_{n1,n_2} |U^p_0|^{n_1} |U^s_0|^{n_2} U^s_0 \right] \\
&\quad + \epsilon R^s_0(U,\epsilon) + O(\epsilon^{N+1}), \\
i\partial_t U^p_0 - \lambda_0 U^p_0 &= -i \left[\epsilon C^{00}_{00} U^s_0 \overline{U}^s_0 U^p_0 + \sum_{n_1,n_2>0}^{n_1+n_2<N+1} \epsilon^{\frac{n_1+n_2}{2}} C^p_{n1,n_2} |U^p_0|^{n_1} |U^s_0|^{n_2} U^p_0 \right] \\
&\quad + \epsilon R^p_0(U,\epsilon) + O(\epsilon^{N+1}), \\
i\partial_t U^s_{\lambda} - \lambda U^s_{\lambda} &= \epsilon R^s_{\lambda}(U,\epsilon) + O(\epsilon^{N+1}), \\
i\partial_t U^p_{\mu} - \mu U^p_{\mu} &= \epsilon R^p_{\mu}(U,\epsilon) + O(\epsilon^{N+1}),
\end{aligned}$$

[5]Capitals denote new variables while small ones denote old variables.

where $R_*^{s,p}(U,\epsilon)$ are the terms which vanish if there is no energy in radiative modes ($U_\lambda^s = U_\mu^p = 0$ for all $\lambda, \mu > 0 \Rightarrow R = 0$).

Proof. We will prove the theorem by applying a series of nearly identical transformations. We start with the transformation

$$u_\lambda^s = U_\lambda^s + \epsilon C_\lambda^s U_0^p \overline{U_0^p} U_0^s.$$

Straightforward calculations show that in order to remove the corresponding radiation driving terms, the coefficient C_λ^s must be of the form

$$C_\lambda^s = \frac{C_{0\lambda}^{00}}{i(\lambda_0 - \lambda)}$$

and similarly to remove pump radiation driving terms, we apply

$$u_\mu^p = U_\mu^p + \epsilon C_\mu^p U_0^s \overline{U_0^s} U_0^p$$

with

$$C_\mu^p = \frac{C_{0\mu}^{00}}{i(\mu_0 - \mu)}.$$

Our results from the previous sections indicate that the transformations are valid and the new system is well defined. Indeed, with the first-order radiation driving terms removed, the obtained system is equivalent to (3.14)–(3.15).

Next, we formally remove quadratic radiation driving terms. We observe that all transformations are near-identity ones differing only by fundamental mode amplitudes. No small denominators arise due to the gap in the spectrum (between the eigenvalue and the continuum spectrum). Continuing these transformations, we remove higher-order radiation driving terms to order N. ∎

Appendix C. Numerical simulations. We verify some of the results obtained in this paper by carrying out numerical simulations. We simulate system (1.2) in one dimension, where the potential is chosen to be $V = \mathrm{sech}^2(x)$, so that the fundamental mode can be explicitly calculated. We use the initial data with all the power (L^2 norm) contained in fundamental modes.

The numerical simulation uses a Fourier split-step scheme, where evolutions due to dispersive, potential, and nonlinear interactions are calculated separately. Nonlinear interaction is solved exactly using the standard solution of the corresponding ODE [1]. Time step is chosen to be $\Delta t = 0.01$ and there are 2^{12} Fourier modes. In Figure 1, the L^2 norm of total radiation is calculated after sufficiently long evolution, so that power exchange between the fields is almost complete. This is done for $\epsilon \in [0.005, 0.05]$. One can see that the time interval is sufficiently long from Figure 2, where for the smallest $\epsilon = 0.005$, the power of both fields contained in the fundamental modes is computed. Even in this case with the smallest ϵ (so that the energy exchange takes longer) there is enough time for the pump field to transfer almost all the power to the signal field. It appears from Figure 1 that losses due to radiation (L^2 norm) scale linearly with nonlinearity strength ϵ, as predicted by our analysis.

Acknowledgment. Part of this research was carried out while V. Zharnitsky was visiting the Program in Applied and Computational Mathematics at Princeton University. He would like to thank Ingrid Daubechies for her hospitality and for providing stimulating research environment.

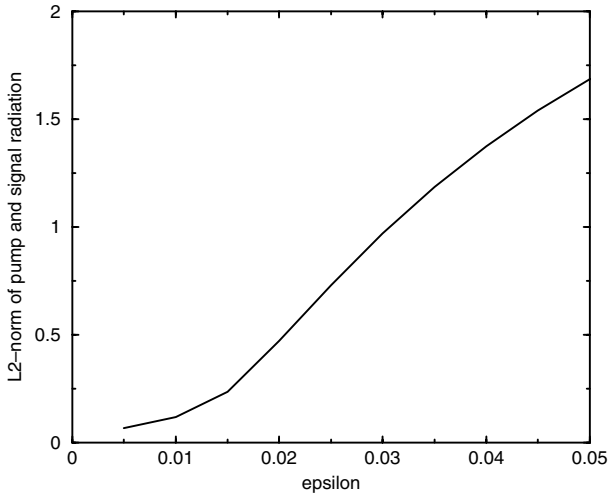

FIG. 1. *Dependence of radiation power on time. The radiation is "measured" after the evolution for* 75 *units of dimensionless time.*

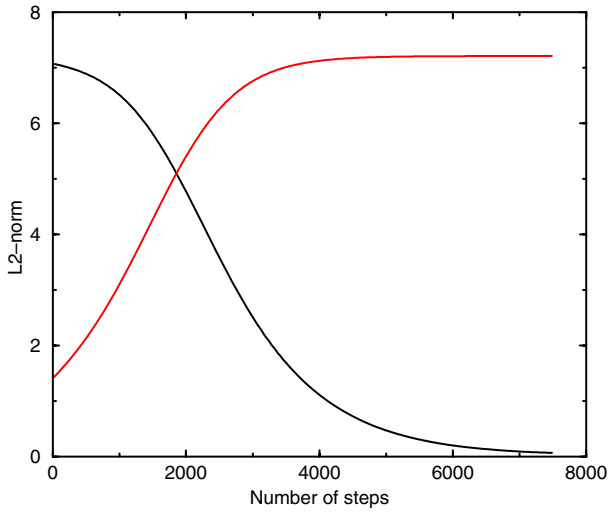

FIG. 2. *Power exchange between fundamental modes for the smallest* $\epsilon = 0.005$ *after the evolution for the same time* $T = 75$ *(it corresponds to* 7500 *steps). Note that the pump field is almost completely "depleted." This indicates that time interval* $T = 75$ *is of sufficient length for the power exchange to take place.*

REFERENCES

[1] G. P. AGRAWAL, *Nonlinear Fiber Optics*, Academic Press, San Diego, 1995.

[2] C. D. LEVERMORE AND M. OLIVER, *The complex Ginzburg-Landau equation as a model problem*, in Dynamical Systems and Probabilistic Methods in Partial Differential Equations, Berkeley, CA, 1994, Lectures in Appl. Math. 31, AMS, Providence, RI, 1996, pp. 141–190.

[3] W. P. URQUHART AND P. J. LAYBOURN, *Effective core area for stimulated Raman scattering in single-mode optical fibers*, IEEE Proc., 132 (1985), pp. 201–204.

[4] C. SULEM AND P.-L. SULEM, *The Nonlinear Schrödinger Equation*, Springer, New York, 1999.

[5] R. WEDER, $L^p - L^{p'}$ *estimates for Schrödinger equation on the line and inverse scattering for*

the nonlinear Schrödinger equation with a potential, J. Funct. Anal., 170 (2000), pp. 37–68.

[6] R. Weder, *The $W^{k,p}$-continuity of the Schrödinger wave operators on the line*, Comm. Math. Phys., 208 (1999), pp. 507–520.

[7] T. Kato and G. Ponce, *Commutator estimates and the Euler Navier-Stokes equations*, Comm. Pure Appl. Math., 41 (1988), pp. 891–907.

[8] R. S. Strichartz, *Restrictions of Fourier transforms to quadratic surfaces and decay of solutions of wave equations*, Duke Math. J., 44 (1977), pp. 705–714.

[9] I. E. Segal, *Space-time decay for solutions of wave equations*, Adv. Math., 22 (1976), pp. 305–311.

[10] K. Yajima, *L^p-boundedness of wave operators for two dimensional Schrödinger operators*, Comm. Math. Phys., 208 (1999), pp. 125–152.

[11] T. Cazenave and F. Weissler, *The Cauchy problem for the nonlinear Schrödinger equation in H^1*, Manuscripta Math., 61 (1988), pp. 477–494.

[12] J. Ginibre and G. Velo, *The global Cauchy problem for the nonlinear Schrödinger equation revisited*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 2 (1985), pp. 309–327.

SIAM J. MATH. ANAL.
Vol. 36, No. 6, pp. 1772–1806

BLOW-UP BEHAVIOR OF PLANAR HARMONIC FUNCTIONS SATISFYING A CERTAIN EXPONENTIAL NEUMANN BOUNDARY CONDITION*

KAI MEDVILLE† AND MICHAEL S. VOGELIUS†

Abstract. In this paper we provide a detailed analysis of the limiting behavior of some very general families of solutions to the boundary value problem $\Delta v = 0$ in Ω, $\partial v/\partial \mathbf{n} = \lambda \sinh(v)$ on $\partial\Omega$, as $\lambda \to 0^+$. The existence of countably many of these families has already been established in [*Quart. Appl. Math.*, 60 (2002), pp. 675–694] and [*Proc. Roy. Soc. Edinburgh Sect. A*, 133 (2003), pp. 119–149].

Key words. harmonic functions, exponential boundary conditions, blow-up

AMS subject classifications. 35B30, 35J65, 78A35

DOI. 10.1137/S0036141003436090

1. Introduction.

In this paper we give a detailed analysis of the limiting "blow-up" behavior of certain solutions to the two-dimensional, nonlinear elliptic boundary value problem

$$\begin{aligned} &\Delta v_\lambda = 0 \quad \text{in } \Omega, \\ &\frac{\partial v_\lambda}{\partial \mathbf{n}} = \lambda \sinh(v_\lambda) \quad \text{on } \partial\Omega. \end{aligned} \tag{1.1}$$

This is a simplified model problem, the likes of which frequently show up in connection with corrosion/oxidation modeling. For a brief discussion of some practical aspects of this problem, and some references to the applied literature, we refer the reader to [5] and [12]. For $\lambda < 0$ the solution structure of (1.1) is trivial: zero is the only solution. For $\lambda = 0$ the only solutions are constants. Our focus is thus on certain nontrivial (nonzero) solutions corresponding to $\lambda > 0$, and in particular on the asymptotic behavior of these solutions as λ approaches zero. In the case when $\Omega = D$ is a disk (e.g., the unit disk) it is possible to find explicit formulas for a countable set of families of solutions to the problem (1.1) (cf. [4]). To be precise, let $\{\mathbf{x}_i\}_{i=0}^{2k-1}$ be a set of $2k$ equispaced points on the unit circle, and set

$$v_{2k,\lambda}(x) = \sum_{i=0}^{2k-1} (-1)^i K(x - \mu_k(\lambda)\mathbf{x}_i),$$

with $K(x) = \log|x|^2$ and $\mu_k(\lambda) = [(k+\lambda)/(k-\lambda)]^{1/2k}$. For any positive integer k and any $0 < \lambda < k$ the functions $v_{2k,\lambda}$ are solutions to (1.1). A simple computation shows that

$$\|\nabla v_{2k,\lambda}\|^2_{L^2(D)} = 8k\pi \log(1/\lambda) + O(1),$$

*Received by the editors October 14, 2003; accepted for publication (in revised form) March 15, 2004; published electronically May 20, 2005. This research was partially supported by NSF grants DMS-0307119 and INT-0003788.

http://www.siam.org/journals/sima/36-6/43609.html

†Department of Mathematics, Rutgers University, New Brunswick, NJ 08903 (medville@math.rutgers.edu, vogelius@math.rutgers.edu).

as $\lambda \to 0^+$. It is also easy to see that

$$\frac{\partial v_{2k,\lambda}}{\partial \mathbf{n}} \to 2\pi \sum_{i=0}^{2k-1} (-1)^{i+1} \delta_{\mathbf{x}_i}$$

in the sense of measures on ∂D, as $\lambda \to 0^+$.

In [6] it was shown that similar families of solutions whose gradient L^2 norms blow up like $\sqrt{\log(1/\lambda)}$, as $\lambda \to 0^+$, continue to exist for arbitrary smooth domains, Ω. These solutions were characterized variationally, and it was also shown that the corresponding boundary currents, $\partial v_\lambda/\partial \mathbf{n} = \lambda \sinh(v_\lambda)$, stay bounded in $L^1(\partial\Omega)$ (the bound depending on the particular family). We were thus able to conclude that there exist appropriately normalized convergent subsequences of solutions, and we showed that each subsequence has a finite, nonempty set of "blow-up" points, which also happen to be the points at which the limit of the absolute value of the flux, $|\partial v_\lambda/\partial \mathbf{n}|$, has nonzero point masses. As was pointed out in [6] these solutions do not necessarily represent all solutions. For instance, for certain nonsimply connected domains it is not hard to construct additional families of solutions whose gradient L^2 norms blow up faster than $\sqrt{\log(1/\lambda)}$ (and which themselves blow up everywhere, except on a set of measure 0).

In this paper we provide a detailed characterization of the elliptic boundary value problem satisfied by an arbitrary ($\lambda \to 0^+$) limit point v_0, of the normalized functions $v_\lambda^0 = v_\lambda - \int_{\partial\Omega} v_\lambda/|\partial\Omega|$, coming from a family of solutions to (1.1) whose boundary currents stay bounded in $L^1(\partial\Omega)$. In particular we show (Theorem 3.1) that the limiting boundary flux, in addition to a nontrivial finite sum of point masses, may contain a regular part that is proportional to e^{v_0} or e^{-v_0}. The possible presence of a regular part of the limiting boundary flux represents a sharp contrast to the corresponding situation for the problem $\Delta v_\lambda = -\lambda e^{v_\lambda}$ (with Dirichlet boundary conditions) where only a pure sum of (negative) point masses occur. It will frequently happen that the constant in front of the exponential term is zero (so that the regular part vanishes)—this is, for instance, the case whenever the solutions have an odd symmetry, or in general whenever the boundary averages of the solutions are small in the sense that $\lambda \exp[|\int_{\partial\Omega} v_\lambda|/|\partial\Omega|] \to 0$. We do, however, also provide very convincing numerical evidence that nonvanishing regular parts do indeed occur: For a family of domains that are simple conformal images (using the maps $z \to e^{\gamma z}$) of the unit disk, our computations clearly document how, for certain values of the parameter γ, a regular part seems to emerge.

For simply connected domains, we derive necessary conditions for the weights and locations of the point masses. The weights are always larger than or equal to 2π in absolute value, and generically take values $\pm 2\pi$ (Theorem 4.1). The conditions concerning locations are derived under the assumption that the limit flux be a pure sum of point masses (no regular part). These conditions express the fact that the tangential derivative of the regular part of the limiting solution, v_0, vanishes at all potential point mass locations (Theorem 4.6). They may be seen as the analogues of the conditions derived in [11] (see also [10] and [13]) for the weights and the singularity locations of limits of solutions to the equation $\Delta v_\lambda = -\lambda e^{v_\lambda}$, with Dirichlet boundary conditions.

2. The particular solutions. Let $\lambda_n \to 0$ be a sequence of positive real numbers, and let v_{λ_n} be solutions of

$$\begin{aligned} &\Delta v_{\lambda_n} = 0 \quad \text{in } \Omega, \\ &\frac{\partial v_{\lambda_n}}{\partial \mathbf{n}} = \lambda_n \sinh(v_{\lambda_n}) \quad \text{on } \partial\Omega, \end{aligned} \tag{2.1}$$

where Ω is a bounded, smooth (C^∞) domain in $\mathbb{R}^2$. Whenever in this paper we talk about solutions to (2.1), we mean $v_{\lambda_n} \in H^1(\Omega)$ that satisfy the standard weak formulation of this nonlinear Neumann problem. Due to elliptic regularity theory it is well known that any such solution is also a classical $C^\infty(\overline{\Omega})$ solution. Let $E_\lambda(\cdot)$ denote the energy

$$E_\lambda(v) = \frac{1}{2}\int_\Omega |\nabla v|^2 \, dx - \lambda \int_{\partial\Omega} (\cosh(v) - 1) \, d\sigma.$$

Suppose that, for some positive constants a_i, b_i, $i = 0, 1$,

$$a_0 \log\left(\frac{1}{\lambda_n}\right) - b_0 \le E_{\lambda_n}(v_{\lambda_n}) \le a_1 \log\left(\frac{1}{\lambda_n}\right) + b_1. \tag{2.2}$$

Since $\cosh(x) - 1 \le \epsilon x \sinh(x) + C_\epsilon$ (for any $0 < \epsilon$), it follows that, for solutions to (2.1),

$$\begin{aligned} E_{\lambda_n}(v_{\lambda_n}) &\le \frac{1}{2}\int_\Omega |\nabla v_{\lambda_n}|^2 \, dx \\ &= E_{\lambda_n}(v_{\lambda_n}) + \lambda_n \int_{\partial\Omega} (\cosh(v_{\lambda_n}) - 1) \, d\sigma \\ &\le E_{\lambda_n}(v_{\lambda_n}) + \epsilon\lambda_n \int_{\partial\Omega} v_{\lambda_n} \sinh(v_{\lambda_n}) \, d\sigma + C_\epsilon \lambda_n \\ &= E_{\lambda_n}(v_{\lambda_n}) + \epsilon \int_\Omega |\nabla v_{\lambda_n}|^2 \, dx + C_\epsilon \lambda_n, \end{aligned}$$

and so the assumption (2.2) is also (for $0 < \lambda_n < C$) equivalent to

$$a_0 \log\left(\frac{1}{\lambda_n}\right) - b_0 \le \int_\Omega |\nabla v_{\lambda_n}|^2 \, dx \le a_1 \log\left(\frac{1}{\lambda_n}\right) + b_1 \tag{2.3}$$

for some positive constants a_i, b_i, $i = 0, 1$. The existence of infinitely (countably) many families of solutions to (2.1), that satisfy (2.2) (or (2.3)), was already established in [6]. These solutions were characterized variationally. To be more precise, the upper bound in (2.2) (or (2.3)) is a consequence of the particular construction we perform in [6]. The lower bounds, however, hold for any nontrivial solution, as asserted by the following lemma.

LEMMA 2.1. *Suppose v_λ, $0 < \lambda$, is a solution to* (1.1) *which is not identically zero. There exist constants $a, b > 0$, independent of λ and v_λ such that*

$$a \log\left(\frac{1}{\lambda}\right) - b \le E_\lambda(v_\lambda) \quad \textit{and} \quad a \log\left(\frac{1}{\lambda}\right) - b \le \int_\Omega |\nabla v_\lambda|^2 \, dx.$$

Proof. This is a restatement of Lemma 3.2 in [6]. We refer to that paper for the proof. ∎

A consequence of the upper bound in (2.2) (or (2.3)) is that

$$\|\lambda_n \sinh(v_{\lambda_n})\|_{L^1(\partial\Omega)} = \lambda_n \int_{\partial\Omega} |\sinh(v_{\lambda_n})|\, d\sigma \le C. \tag{2.4}$$

The verification of this relies on the following real analysis lemma.

LEMMA 2.2. *Let a and b be two given positive constants, and let w be a continuous function such that for a certain $\lambda \in (0,1)$, one has*

$$\int_{\partial\Omega} |w| e^{|w|}\, d\sigma \le \frac{a}{\lambda} \log\left(\frac{1}{\lambda}\right) + b.$$

There exists a positive constant C, depending only on a, b, and $|\partial\Omega|$ such that

$$\int_{\partial\Omega} e^{|w|}\, dx \le \frac{C}{\lambda}.$$

Proof. Let f denote the function $f(x) = x \log x$. A simple computation shows that f is convex and monotonically increasing on the half-line $[1,\infty)$. An application of Jensen's inequality gives

$$\begin{aligned} f\left(\int_{\partial\Omega} e^{|w|} \frac{d\sigma}{|\partial\Omega|}\right) &\le \int_{\partial\Omega} f\big(e^{|w|}\big) \frac{d\sigma}{|\partial\Omega|} \\ &= \int_{\partial\Omega} e^{|w|} |w| \frac{d\sigma}{|\partial\Omega|} \\ &\le \frac{a}{|\partial\Omega|\lambda} \log\left(\frac{1}{\lambda}\right) + \frac{b}{|\partial\Omega|} \\ &\le \frac{C}{\lambda} \log \frac{C}{\lambda} = f\left(\frac{C}{\lambda}\right), \end{aligned}$$

with $C \ge 1$ depending only on a, b, and $|\partial\Omega|$. The monotonicity of f now yields the desired estimate. ∎

To arrive at (2.4) from the upper bound in (2.3), simply note that

$$\begin{aligned} \int_{\partial\Omega} |v_{\lambda_n}| e^{|v_{\lambda_n}|}\, d\sigma &\le 2 \int_{\partial\Omega} v_{\lambda_n} \sinh(v_{\lambda_n})\, d\sigma + |\partial\Omega| e^{-1} \\ &= 2\lambda_n^{-1} \int_{\Omega} |\nabla v_{\lambda_n}|^2\, dx + |\partial\Omega| e^{-1} \\ &\le \frac{a}{\lambda_n} \log\left(\frac{1}{\lambda_n}\right) + b, \end{aligned}$$

and then use Lemma 2.2.

We shall also make use of the decomposition $v_{\lambda_n} = v^0_{\lambda_n} + s_{\lambda_n}$, with $s_{\lambda_n} = \int_{\partial\Omega} v_{\lambda_n}\, d\sigma/|\partial\Omega|$ (and thus $\int_{\partial\Omega} v^0_{\lambda_n}\, d\sigma = 0$). By a combination of Jensen's inequality for the exponential function and the estimate (2.4), it follows immediately that

$$\begin{aligned} |s_{\lambda_n}| &\le \log\left(\exp\left(\int_{\partial\Omega} |v_{\lambda_n}| \frac{d\sigma}{|\partial\Omega|}\right)\right) \le \log\left(\int_{\partial\Omega} \exp(|v_{\lambda_n}|) \frac{d\sigma}{|\partial\Omega|}\right) \\ &\le \log\left(\int_{\partial\Omega} \frac{2}{|\partial\Omega|} |\sinh(v_{\lambda_n})|\, d\sigma + 1\right) \le \log\left(\frac{C}{\lambda_n} + 1\right) \\ &\le \log \frac{1}{\lambda_n} + D. \end{aligned} \tag{2.5}$$

This estimate was used in the "blow-up" analysis in [6]. We note that as an immediate consequence of (2.5) we get the estimate $\lambda_n e^{|s_{\lambda_n}|} \le e^D$.

3. The limiting behavior. We shall now study in more detail the limiting behavior of solutions v_{λ_n} to (2.1) that satisfy the L^1 boundary flux estimate (2.4). For that purpose it is useful to introduce the Neumann function $N(x,y)$. For fixed $y \in \Omega$ this solves

$$\begin{aligned} \Delta_x N(x,y) &= \delta_y \quad \text{in } \Omega, \\ \frac{\partial N}{\partial \mathbf{n}_x}(x,y) &= \frac{1}{|\partial\Omega|}, \end{aligned}$$

with the normalization $\int_{\partial\Omega} N(x,y)\, d\sigma_x = 0$. It is well known that $N(x,y)$ may be smoothly extended to $\overline{\Omega \times \Omega} \setminus \{x = y\}$ and that $N(x,y) = N(y,x)$. For fixed $y \in \partial\Omega$ the function $N(x,y)$ satisfies

$$\begin{aligned} \Delta_x N(x,y) &= 0 \quad \text{in } \Omega, \\ \frac{\partial N}{\partial \mathbf{n}_x}(x,y) &= -\delta_y + \frac{1}{|\partial\Omega|}. \end{aligned}$$

For fixed $y \in \partial\Omega$, $N(x,y)$ therefore allows the decomposition

$$N(y,x) = N(x,y) = \frac{1}{\pi} \log|x-y| + H_y(x), \tag{3.1}$$

where the function $H_y(\cdot) \in C^\infty(\overline{\Omega})$ is the classical solution to

$$\begin{aligned} &\Delta_x H_y(x) = 0 \quad \text{in } \Omega, \qquad \frac{\partial H_y}{\partial \mathbf{n}_x} = -\frac{1}{\pi} \frac{(x-y)\cdot \mathbf{n}_x}{|x-y|^2} + \frac{1}{|\partial\Omega|} \quad \text{on } \partial\Omega, \\ &\qquad \text{with } \int_{\partial\Omega} H_y \, d\sigma_x = -\int_{\partial\Omega} \frac{1}{\pi} \log|x-y| \, d\sigma_x. \end{aligned} \tag{3.2}$$

In terms of $N(x,y)$ the functions $v^0_{\lambda_n}$ may be represented as follows:

$$v^0_{\lambda_n}(y) = -\int_{\partial\Omega} N(x,y) \frac{\partial v_{\lambda_n}}{\partial \mathbf{n}} \, d\sigma_x = -\int_{\partial\Omega} N(x,y) \lambda_n \sinh(v_{\lambda_n}) \, d\sigma_x. \tag{3.3}$$

Given any function f, let $f^+ \geq 0$ and $f^- \geq 0$ denote its positive and negative part, respectively, i.e., let $f^+ = \max\{f, 0\}$ and $f^- = -\min\{f, 0\}$. With this definition $f = f^+ - f^-$ and $|f| = f^+ + f^-$. Some of our main results are contained in the following theorem.

THEOREM 3.1. *Let $\Omega \subset \mathbb{R}^2$ be a bounded smooth (C^∞) domain, and let $v_{\lambda_n} \in H^1(\Omega)$, $\lambda_n \to 0^+$, be a sequence of nontrivial, i.e., not identically vanishing, solutions to the nonlinear elliptic Neumann problem* (2.1) *that additionally satisfy the boundary flux estimate* (2.4). *Decompose v_{λ_n} as $v_{\lambda_n} = v^0_{\lambda_n} + s_{\lambda_n}$, with $s_{\lambda_n} = \int_{\partial\Omega} v_{\lambda_n} \, d\sigma / |\partial\Omega|$. There exists a subsequence, for simplicity also denoted v_{λ_n}; two positive, regular Borel measures μ_+ and μ_-; and two nonnegative constants d_+ and d_- such that*

$$\lambda_n \sinh(v^+_{\lambda_n}) = \lambda_n \sinh(v_{\lambda_n})^+ \to \mu_+, \qquad \lambda_n \sinh(v^-_{\lambda_n}) = \lambda_n \sinh(v_{\lambda_n})^- \to \mu_-$$

in the sense of measures on $\partial\Omega$ (i.e., in the weak* topology of the dual of $C^0(\partial\Omega)$) and

$$\lambda_n e^{s_{\lambda_n}} \to d_+, \qquad \lambda_n e^{-s_{\lambda_n}} \to d_-.$$

At least one of the constants d_+ and d_- is zero, i.e., there are two possible scenarios:

$$(d_+, d_-) = (d_+, 0) \quad or \quad (d_+, d_-) = (0, d_-).$$

The subsequence $v^0_{\lambda_n}$ converges in $H^t(\Omega)$ for any $t < 1$; the limit, v_0, is the solution to

$$\Delta v_0 = 0 \quad in\ \Omega, \qquad \frac{\partial v_0}{\partial \mathbf{n}} = \mu_+ - \mu_- \quad on\ \partial\Omega, \quad \int_{\partial\Omega} v_0\ d\sigma = 0,$$

in the sense that

$$v_0(y) = -\int_{\partial\Omega} N(x,y)\ d(\mu_+ - \mu_-)_x, \quad y \in \Omega.$$

There exist two finite sets of points $\{\mathbf{x}_i^+\}_{i=1}^M$ and $\{\mathbf{x}_i^-\}_{i=1}^N \subset \partial\Omega$, and two sets of positive weights $\{\alpha_i^+\}_{i=1}^M$ and $\{\alpha_i^-\}_{i=1}^N$ such that

$$\mu_+ = \sum_{i=1}^M \alpha_i^+ \delta_{\mathbf{x}_i^+} + \frac{d_+}{2} e^{v_0}, \qquad \mu_- = \sum_{i=1}^N \alpha_i^- \delta_{\mathbf{x}_i^-} + \frac{d_-}{2} e^{-v_0}. \tag{3.4}$$

The combined set $S = \{\mathbf{x}_i^+\}_{i=1}^M \cup \{\mathbf{x}_i^-\}_{i=1}^N$ is nonempty. The function v_0 is infinitely smooth away from S, i.e., $v_0 \in C^\infty(\overline{\Omega}\backslash S)$, and the convergence of $v^0_{\lambda_n}$ toward v_0 takes place in $C^\infty(K)$ for any compact set $K \subset \overline{\Omega}\backslash S$. The functions $\frac{d_\pm}{2} e^{\pm v_0}$ of the limiting boundary fluxes (3.4) *are in $L^1(\partial\Omega)$. The set $S = \{\mathbf{x}_i^+\}_{i=1}^M \cup \{\mathbf{x}_i^-\}_{i=1}^N$ represents exactly the locations of the point masses of the measure $\mu_+ + \mu_- = \lim_{\lambda_n \to 0^+} \lambda_n |\sinh(v_{\lambda_n})|$. Furthermore, this set also represents the "blow-up" points for the subsequence $v^0_{\lambda_n}$, in the sense that*

$$S = \{x \in \bar{\Omega} : \exists x_n \in \bar{\Omega},\ with\ x_n \to x,\ such\ that\ |v^0_{\lambda_n}(x_n)| \to \infty\}.$$

Remark 3.2. As stated in the theorem, the locations of the point masses for the measure $\lim_{\lambda_n \to 0^+} \lambda_n |\sinh(v_{\lambda_n})| = \mu_+ + \mu_-$ are exactly the set $\{\mathbf{x}_i^+\}_{i=1}^M \cup \{\mathbf{x}_i^-\}_{i=1}^N$. We cannot exclude some overlap between the points $\mathbf{x}_i^+$ of the measure μ_+ and the points $\mathbf{x}_i^-$ of the measure μ_-. In the case of common points it might at first seem possible that the corresponding coefficients α^+ and α^- are equal. In other words it might at first seem possible that the locations of the nonzero point masses for the measure $\lim_{\lambda_n \to 0^+} \lambda_n \sinh(v_{\lambda_n}) = \mu_+ - \mu_-$ are a strict subset of $\{\mathbf{x}_i^+\}_{i=1}^M \cup \{\mathbf{x}_i^-\}_{i=1}^N$. However, a closer analysis shows that this is never the case; in Theorem 4.1 we prove that for Ω simply connected $|\alpha^+ - \alpha^-|$ (as well as $\alpha^+ + \alpha^-$) is always greater than or equal to 2π. In particular, it follows that $\mu_+ - \mu_-$ and $\mu_+ + \mu_-$ have the exact same nonzero point mass locations. See also Remark 4.2, following the statement of Theorem 4.1.

Remark 3.3. In the case when Ω is a disk we have constructed countably many families of explicit solutions satisfying boundary flux bound (2.4) (cf. [4]). These solutions all have $s_\lambda = \int_{\partial\Omega} v_\lambda\ d\sigma/|\partial\Omega| = 0$, so that $d_+ = d_- = 0$, and thus the corresponding limiting problems have boundary fluxes consisting of point masses only. In section 5 of this paper we provide numerical examples of families of solutions for which the limiting boundary fluxes have point masses, as well as nonzero regular parts. For these examples the domain, Ω, is an exponential image of a disk.

Remark 3.4. If v_{λ_n}, $\lambda_n \to 0^+$, is a sequence of solutions for which the boundary flux estimate (2.4) does not hold, then we may extract a subsequence such that

$$\alpha_n = \|\lambda_n \sinh(v_{\lambda_n})\|_{L^1(\partial\Omega)} \to \infty.$$

Consider now $w_{\lambda_n} = v_{\lambda_n}/\alpha_n$. By extraction of a subsequence we may obtain that

$$\frac{\partial w_{\lambda_n}}{\partial \mathbf{n}} = \lambda_n \sinh(v_{\lambda_n})/\alpha_n \to \tilde{\mu},$$

in the sense of measures on $\partial\Omega$. If we *assume* that the limiting measure $\tilde{\mu}$ is *not* identically zero, then it is easy to show that

$$w^0_{\lambda_n}(y) = -\int_{\partial\Omega} N(x,y)\frac{\partial w_{\lambda_n}}{\partial \mathbf{n}}\, d\sigma_x \to -\int_{\partial\Omega} N(x,y)\, d\tilde{\mu}_x = w_0(y), \quad y \in \Omega,$$

with $w_0(y)$ being different from zero almost everywhere in Ω. It follows immediately that $v^0_{\lambda_n} = \alpha_n w^0_{\lambda_n}$ converges to $\pm\infty$ almost everywhere in Ω. As indicated by this simple argument, "blow-up" almost everywhere in Ω appears as a highly probable alternative to the finite (boundary) point "blow-up" described by Theorem 3.1. However, we do want to emphasize that here, unlike in the case of the boundary value problem $\Delta u_\lambda = -\lambda e^{u_\lambda}$ in Ω, $u_\lambda = 0$ on $\partial\Omega$ (cf. [11]), this is *not* the only alternative for a sequence v_{λ_n}, with $\|\lambda_n \sinh(v_{\lambda_n})\|_{L^1(\partial\Omega)} \to \infty$. For instance, it is very easy to select, among the explicit solutions we constructed in [4], a sequence whose elements (as $\lambda_n \to 0^+$) come from "higher and higher" branches in such a way that $\|\lambda_n \sinh(v_{\lambda_n})\|_{L^1(\partial\Omega)} \to \infty$, but at the same time $v^0_{\lambda_n}(y) = v_{\lambda_n}(y)$ has a finite limit (zero) at any point y inside the unit disk.

Proof of Theorem 3.1. Due to the L^1 bound (2.4) on $\lambda_n \sinh(v_{\lambda_n})$ it follows that $\lambda_n \sinh(v_{\lambda_n})^{\pm}$ are bounded in $L^1(\partial\Omega)$, and therefore norm bounded in the dual of $C^0(\partial\Omega)$. From the bound $|s_{\lambda_n}| \le \log\frac{1}{\lambda_n} + D$ (see (2.5)) we get that $\lambda_n e^{|s_{\lambda_n}|} \le e^D$. These bounds (and the compactness) imply the existence of a subsequence (also denoted λ_n) two nonnegative, regular Borel measures μ_+, μ_- and two nonnegative constants d_+, d_-, so that

$$\lambda_n \sinh(v_{\lambda_n})^+ \to \mu_+, \qquad \lambda_n \sinh(v_{\lambda_n})^- \to \mu_- \tag{3.5}$$

in the sense of measures—that is, in the weak* topology on the dual of $C^0(\partial\Omega)$—and

$$\lambda_n e^{s_{\lambda_n}} \to d_+, \qquad \lambda_n e^{-s_{\lambda_n}} \to d_-.$$

If $d_+ > 0$, then $e^{s_{\lambda_n}} \to \infty$, and thus $e^{-s_{\lambda_n}} \to 0$, so that $d_- = 0$; similarly, if $d_- > 0$, then we may conclude that $d_+ = 0$. In summary, at least one of the constants d_+ and d_- is zero.

Due to the fact that $\int_{\partial\Omega} \lambda_n \sinh(v_{\lambda_n})\, d\sigma = 0$ we conclude that $(\mu_+ - \mu_-)(\partial\Omega) = \lim_{\lambda_n\to 0}\int_{\partial\Omega} \lambda_n \sinh(v_{\lambda_n})\, d\sigma = 0$, or $\mu_+(\partial\Omega) = \mu_-(\partial\Omega)$. The $L^1(\partial\Omega)$ bound on $\lambda_n \sinh(v_{\lambda_n})$ in combination with Sobolev's imbedding theorem implies that

$$\begin{aligned}\|\lambda_n \sinh(v_{\lambda_n})\|_{H^{-s}(\partial\Omega)} &= \sup_{\|w\|_{H^s(\partial\Omega)}\le 1}\int_{\partial\Omega} \lambda_n \sinh(v_{\lambda_n}) w\, d\sigma \\ &\le \|\lambda_n \sinh(v_{\lambda_n})\|_{L^1(\partial\Omega)} \sup_{\|w\|_{H^s(\partial\Omega)}\le 1} \|w\|_{L^\infty(\partial\Omega)} \le C\end{aligned}$$

for any $s > 1/2$. Duality and elliptic estimates for solutions to the boundary value problem $\Delta w = f$ in Ω, $\partial w/\partial\mathbf{n} = \text{const}$ on $\partial\Omega$ now yield

$$\|v^0_{\lambda_n}\|_{H^{\frac{3}{2}-s}(\Omega)} \le C\|\lambda_n \sinh(v_{\lambda_n})\|_{H^{-s}(\partial\Omega)} \le C \tag{3.6}$$

for arbitrary $\frac{3}{2} - s < 1$. By compactness we may extract a subsequence, also referred to as $v^0_{\lambda_n}$, so that $v^0_{\lambda_n}$ converges, in say $L^2(\Omega)$, to a limit v_0. By compactness (and uniqueness of the limit) this subsequence will now actually converge to v_0 in $H^t(\Omega)$ for any $t < 1$. Since

$$v^0_{\lambda_n}(y) = -\int_{\partial\Omega} N(x,y)\frac{\partial v_{\lambda_n}}{\partial \mathbf{n}}\, d\sigma_x = -\int_{\partial\Omega} N(x,y)\lambda_n \sinh(v_{\lambda_n})\, d\sigma_x, \quad y \in \Omega,$$

it follows immediately from (3.5) that $v^0_{\lambda_n}$ converges to $-\int_{\partial\Omega} N(x,y)\ d(\mu_+ - \mu_-)_x$ pointwise in Ω. By uniqueness of the limit we thus get

$$v_0(y) = -\int_{\partial\Omega} N(x,y)\ d(\mu_+ - \mu_-)_x. \tag{3.7}$$

Let ν denote the nonnegative measure $\nu = \mu_+ + \mu_-$. Following [6] (and [3]) we call a point $x_0 \in \partial\Omega$ *regular* if there exists a continuous function $0 \le \psi \le 1$, with $\psi \equiv 1$ in a neighborhood of x_0 such that $\int_{\partial\Omega} \psi d\nu < \pi/2$. Lemma 4.5 of [6] shows that given any regular point x_0 there exists a neighborhood $B_{r_0}(x_0) \cap \partial\Omega$ of x_0, and a constant C such that

$$\|v^0_{\lambda_n}\|_{L^\infty(B_{r_0}(x_0)\cap\partial\Omega)} \le C.$$

The proof of this estimate relies crucially on (an appropriate adaptation of) an inequality due to Brezis and Merle (cf. [3] and [6]). Following [6] we call a point $x_0 \in \partial\Omega$ *singular* if it is not regular in the above sense. A point $x_0 \in \partial\Omega$ is thus singular if for any continuous $0 \le \psi \le 1$, with $\psi \equiv 1$ in a neighborhood of x_0, we have $\int_{\partial\Omega} \psi d\nu \ge \pi/2$; as a consequence, $\nu(\{x_0\}) \ge \pi/2$ for any singular point x_0. Let S denote the set of singular points. We immediately conclude that S must consist of finitely many points and that

$$\#S \le \frac{\int_{\partial\Omega} d\nu}{\inf_{x_0\in S}\ \nu(\{x_0\})} \le 2\frac{\int_{\partial\Omega} d\nu}{\pi}.$$

This estimate is part of Lemma 4.7 of [6]. That same lemma further establishes that S is nonempty by showing that otherwise $E_{\lambda_n}(v_{\lambda_n}) \to 0$ as $\lambda_n \to 0^+$, which obviously contradicts the lower bound for nontrivial solutions (cf. Lemma 2.1). Thus $\pi/2 \le \nu(\partial\Omega) = \mu_+(\partial\Omega) + \mu_-(\partial\Omega)$, and so μ_+, μ_-, and ν are indeed positive measures.

Since $v^0_{\lambda_n}|_{\partial\Omega}$ is bounded in L^∞ near any regular point it follows that

$$\frac{\partial v^0_{\lambda_n}}{\partial \mathbf{n}} = \frac{\lambda_n}{2}e^{v_{\lambda_n}} - \frac{\lambda_n}{2}e^{-v_{\lambda_n}} = \frac{\lambda_n}{2}e^{s_{\lambda_n}}e^{v^0_{\lambda_n}} - \frac{\lambda_n}{2}e^{-s_{\lambda_n}}e^{-v^0_{\lambda_n}}$$

is bounded in L^∞, and thus in L^2, near any regular point. By elliptic regularity (and the estimate (3.6)) it now follows that for any regular point, x_0,

$$\|v^0_{\lambda_n}\|_{H^{3/2}(B_{r_1}(x_0)\cap\Omega)} \le C$$

for some $r_1 > 0$. Consequently

$$\|v^0_{\lambda_n}\|_{H^1(B_{r_1}(x_0)\cap\partial\Omega)} \le C, \qquad \|e^{\pm v^0_{\lambda_n}}\|_{H^1(B_{r_1}(x_0)\cap\partial\Omega)} \le C,$$
$$\text{and} \quad \left\|\frac{\partial v^0_{\lambda_n}}{\partial \mathbf{n}}\right\|_{H^1(B_{r_1}(x_0)\cap\partial\Omega)} \le C.$$

By repeated use of elliptic estimates (induction) we may conclude that there exists $r_1 > 0$ such that

$$\|v^0_{\lambda_n}\|_{H^k(B_{r_1}(x_0)\cap\partial\Omega)} \le C_k, \qquad \|e^{\pm v^0_{\lambda_n}}\|_{H^k(B_{r_1}(x_0)\cap\partial\Omega)} \le C_k,$$
$$\text{and} \quad \left\| \frac{\partial v^0_{\lambda_n}}{\partial \mathbf{n}} \right\|_{H^k(B_{r_1}(x_0)\cap\partial\Omega)} \le C_k \quad \text{for any } k \ge 1.$$

Therefore

$$\|v^0_{\lambda_n}\|_{H^s(B_{r_1}(x_0)\cap\Omega)} \le C_s \quad \text{for any } s \ge 1.$$

In combination with a compactness argument and interior elliptic regularity results this yields

$$\|v^0_{\lambda_n}\|_{H^s(K)} \le C_{s,K} \quad \text{for any index } s \tag{3.8}$$

and any compact set $K \subset \overline{\Omega} \setminus S$. Since we already know that $v^0_{\lambda_n}$ converges to v_0 in $H^t(\Omega)$, $t < 1$, it follows from (3.8), compactness, and the uniqueness of the limit that v_0 lies in $C^\infty(\overline{\Omega} \setminus S)$ and that $v^0_{\lambda_n}$ converges to v_0 in $C^\infty(K)$ for any compact set $K \subset \overline{\Omega} \setminus S$ (i.e., $v^0_{\lambda_n}$ converges to v_0 in $C^\infty(\overline{\Omega} \setminus S)$). In particular,

$$\begin{aligned} &v^0_{\lambda_n} \text{ converges to } v_0 \text{ uniformly with all derivatives} \\ &\text{in a neighborhood of any regular point } x_0 \in \partial\Omega. \end{aligned}$$

It follows immediately that

$$\begin{aligned} &\lambda_n \sinh(v_{\lambda_n}) = \frac{\lambda_n}{2}\left(e^{s_{\lambda_n}} e^{v^0_{\lambda_n}} - e^{-s_{\lambda_n}} e^{-v^0_{\lambda_n}}\right) \text{ converges to} \\ &\frac{d_+}{2} e^{v_0} - \frac{d_-}{2} e^{-v_0} \text{ uniformly with all derivatives in a} \\ &\text{neighborhood of any regular point } x_0 \in \partial\Omega \end{aligned}$$

and that

$$\begin{aligned} &\lambda_n |\sinh(v_{\lambda_n})| = \left| \frac{\lambda_n}{2}\left(e^{s_{\lambda_n}} e^{v^0_{\lambda_n}} - e^{-s_{\lambda_n}} e^{-v^0_{\lambda_n}}\right) \right| \text{ converges to} \\ &\left| \frac{d_+}{2} e^{v_0} - \frac{d_-}{2} e^{-v_0} \right| = \frac{d_+}{2} e^{v_0} + \frac{d_-}{2} e^{-v_0} \text{ uniformly in a} \\ &\text{neighborhood of any regular point } x_0 \in \partial\Omega. \end{aligned}$$

For the last identity we used the fact that $d_\pm$ are nonnegative, with at least one being zero. As a consequence

$$\begin{aligned} &\lambda_n \sinh(v_{\lambda_n})^+ = \frac{\lambda_n}{2}(|\sinh(v_{\lambda_n})| + \sinh(v_{\lambda_n})) \text{ converges to } \frac{d_+}{2} e^{v_0} \\ &\text{uniformly in a neighborhood of any regular point } x_0 \in \partial\Omega \end{aligned} \tag{3.9}$$

and

$$\begin{aligned} &\lambda_n \sinh(v_{\lambda_n})^- = \frac{\lambda_n}{2}(|\sinh(v_{\lambda_n})| - \sinh(v_{\lambda_n})) \text{ converges to } \frac{d_-}{2} e^{-v_0} \\ &\text{uniformly in a neighborhood of any regular point } x_0 \in \partial\Omega. \end{aligned} \tag{3.10}$$

From the representation formula (3.7) and the just-established regularity of the measures $\mu_\pm$ at regular points, we conclude that v_0 satisfies the boundary condition

$$\frac{\partial v_0}{\partial \mathbf{n}}(y) = \frac{d_+}{2} e^{v_0(y)} - \frac{d_-}{2} e^{-v_0(y)}$$

in a classical sense at points $y \in \partial\Omega \setminus S$. For any compact set $K \subset \partial\Omega \setminus S$

$$\begin{aligned}\frac{d_\pm}{2} \int_K e^{\pm v_0}\, d\sigma &= \lim_{\lambda_n \to 0} \int_K \lambda_n \sinh(v_{\lambda_n})^\pm \, d\sigma \\ &\le \lim_{\lambda_n \to 0} \int_{\partial\Omega} \lambda_n |\sinh(v_{\lambda_n})| \, d\sigma \\ &= (\mu_+ + \mu_-)(\Omega) = \nu(\Omega).\end{aligned}$$

From Lebesgue's monotone convergence theorem it therefore follows that $\frac{d_\pm}{2} e^{\pm v_0}$ are in $L^1(\partial\Omega)$ with

$$\left\| \frac{d_\pm}{2} e^{\pm v_0} \right\|_{L^1(\partial\Omega)} \le \nu(\Omega).$$

Suppose $S = \cup_{i=1}^L \{\mathbf{x}_i\}$ and let $\phi_i \in C^0(\partial\Omega)$, $i = 1, \dots, L$, be a fixed set of functions with $0 \le \phi_i \le 1$, with $\phi_i(\mathbf{x}_j) = 0$, $j \ne i$, and with $\phi_i \equiv 1$ in a neighborhood of $\mathbf{x}_i$. Given any $\phi \in C^0(\partial\Omega)$ we may now write

$$\phi = \phi_0 + \sum_{i=1}^L \phi(\mathbf{x}_i)\phi_i,$$

where $\phi_0 \in C^0(\partial\Omega)$ vanishes at all points of S. Given any $\epsilon > 0$ we may find $\phi_{0,\epsilon} \in C^0(\partial\Omega)$, with compact support $K \subset \partial\Omega \setminus S$, such that

$$\|\phi_0 - \phi_{0,\epsilon}\|_{C^0(\partial\Omega)} \le \epsilon. \tag{3.11}$$

Using (3.9), (3.10), and compactness we obtain

$$\int_{\partial\Omega} \lambda_n \sinh(v_{\lambda_n})^\pm \phi_{0,\epsilon} \, d\sigma \to \frac{d_\pm}{2} \int_{\partial\Omega} e^{\pm v_0} \phi_{0,\epsilon} \, d\sigma \tag{3.12}$$

as $\lambda_n \to 0$. We also have

$$\begin{aligned}&\left| \int_{\partial\Omega} \lambda_n \sinh(v_{\lambda_n})^\pm \phi \, d\sigma - \sum_{i=1}^L \phi(\mathbf{x}_i) \int_{\partial\Omega} \lambda_n \sinh(v_{\lambda_n})^\pm \phi_i \, d\sigma \right. \\ &\qquad \left. - \int_{\partial\Omega} \lambda_n \sinh(v_{\lambda_n})^\pm \phi_{0,\epsilon} \, d\sigma \right| \\ &= \left| \int_{\partial\Omega} \lambda_n \sinh(v_{\lambda_n})^\pm (\phi_0 - \phi_{0,\epsilon}) \, d\sigma \right| \\ &\le \epsilon \|\lambda_n \sinh(v_{\lambda_n})\|_{L^1(\partial\Omega)}.\end{aligned}$$

After passage to the limit $\lambda_n \to 0$, and combination with (3.12), this yields

$$\left| \int_{\partial\Omega} \phi \, d\mu_\pm - \sum_{i=1}^L \beta_i^\pm \phi(\mathbf{x}_i) - \frac{d_\pm}{2} \int_{\partial\Omega} e^{\pm v_0} \phi_{0,\epsilon} \, d\sigma \right| \le \epsilon \nu(\partial\Omega),$$

with

$$\beta_i^{\pm} = \lim_{\lambda_n \to 0} \int_{\partial\Omega} \lambda_n \sinh(v_{\lambda_n})^{\pm} \phi_i \, d\sigma = \int_{\partial\Omega} \phi_i \, d\mu_{\pm}.$$

Thus

$$\begin{aligned}
&\left| \int_{\partial\Omega} \phi \, d\mu_{\pm} - \sum_{i=1}^{L} \beta_i^{\pm} \phi(\mathbf{x}_i) - \frac{d_{\pm}}{2} \int_{\partial\Omega} e^{\pm v_0} \phi_0 \, d\sigma \right| \\
&\quad \leq \epsilon \nu(\partial\Omega) + \left| \int_{\partial\Omega} \frac{d_{\pm}}{2} e^{\pm v_0} (\phi_0 - \phi_{0,\epsilon}) \, d\sigma \right| \\
&\quad \leq \epsilon \nu(\partial\Omega) + \epsilon \left\| \frac{d_{\pm}}{2} e^{\pm v_0} \right\|_{L^1(\partial\Omega)} \\
&\quad \leq 2\epsilon \nu(\partial\Omega).
\end{aligned}$$

By introducing

$$\alpha_i^{\pm} = \beta_i^{\pm} - \frac{d_{\pm}}{2} \int_{\partial\Omega} e^{\pm v_0} \phi_i \, d\sigma,$$

we may rewrite this latter inequality as

$$\left| \int_{\partial\Omega} \phi \, d\mu_{\pm} - \sum_{i=1}^{L} \alpha_i^{\pm} \phi(\mathbf{x}_i) - \frac{d_{\pm}}{2} \int_{\partial\Omega} e^{\pm v_0} \phi \, d\sigma \right| \leq 2\epsilon \nu(\partial\Omega) \quad \text{for any } \epsilon > 0.$$

We therefore conclude that

$$\int_{\partial\Omega} \phi \, d\mu_{\pm} = \sum_{i=1}^{L} \alpha_i^{\pm} \phi(\mathbf{x}_i) + \frac{d_{\pm}}{2} \int_{\partial\Omega} e^{\pm v_0} \phi \, d\sigma$$

or

$$\mu_{\pm} = \sum_{i=1}^{L} \alpha_i^{\pm} \delta_{\mathbf{x}_i} + \frac{d_{\pm}}{2} e^{\pm v_0}. \tag{3.13}$$

Since $\mu_{\pm}$ are positive it follows that $\alpha_i^{\pm} \geq 0$, $1 \leq i \leq L$. We now let $\{\mathbf{x}_i^+\}_{i=1}^M \subset \{\mathbf{x}_i\}_{i=1}^L$ denote those points for which the corresponding coefficients α_i^+ are strictly positive, and similarly we let $\{\mathbf{x}_i^-\}_{i=1}^N \subset \{\mathbf{x}_i\}_{i=1}^L$ denote those points for which the corresponding coefficients α_i^- are strictly positive. It is obvious that $\{\mathbf{x}_i^+\}_{i=1}^M \cup \{\mathbf{x}_i^-\}_{i=1}^N$ are exactly the locations at which the measure $\nu = \mu_+ + \mu_-$ has nonzero point masses. According to Lemma 4.8 of [6] (see also the corrigendum [7]) the set of singular points, S, is likewise characterized as the set of points at which the measure $\nu = \mu_+ + \mu_-$ has point masses. Consequently $S = \{\mathbf{x}_i\}_{i=1}^L = \{\mathbf{x}_i^+\}_{i=1}^M \cup \{\mathbf{x}_i^-\}_{i=1}^N$ and $N + M \geq L$. Since S is nonempty (i.e., $L \geq 1$) at least one of the sets $\{\mathbf{x}_i^+\}_{i=1}^M$ and $\{\mathbf{x}_i^-\}_{i=1}^N$ must be nonempty (i.e., $M \geq 1$ or $N \geq 1$). From (3.13) and the definition of the points $\{\mathbf{x}_i^+\}_{i=1}^M$ and $\{\mathbf{x}_i^-\}_{i=1}^N$ it now follows (after renumbering the $\alpha_i^{\pm}$) that

$$\mu_+ = \sum_{i=1}^{M} \alpha_i^+ \delta_{\mathbf{x}_i^+} + \frac{d_+}{2} e^{v_0} \quad \text{and} \quad \mu_- = \sum_{i=1}^{N} \alpha_i^- \delta_{\mathbf{x}_i^-} + \frac{d_-}{2} e^{-v_0},$$

with $\alpha_i^{\pm} > 0$, and $d_{\pm} \geq 0$. We recall that at least one of the coefficients d_+ or d_- is zero. This is exactly the desired representation formula. Finally, Lemma 4.8 of [6] asserts that S also equals the set of "blow-up" points for the subsequence $v_{\lambda_n}^0$ in the sense introduced here. This completes the proof of Theorem 3.1. ∎

4. Singularity weights and locations. The aim of this section is to deduce more specific information about the "blow-up" behavior of the subsequence of solutions to (2.1), extracted in Theorem 3.1. We shall do this by uncovering more specific information about the form of the limiting measures μ_+ and μ_-.

THEOREM 4.1. *Suppose the domain $\Omega \subset \mathbb{R}^2$ is smooth, bounded, and simply connected. Let μ_+ and μ_- be the limiting measures from Theorem* 3.1*, i.e.,*

$$\mu_+ = \lim \lambda_n \sinh(v_{\lambda_n})^+ \quad \textit{and} \quad \mu_- = \lim \lambda_n \sinh(v_{\lambda_n})^-.$$

Let $\{\mathbf{x}_i^+\}_{i=1}^M$ and $\{\mathbf{x}_i^-\}_{i=1}^N$ be the locations of the nonzero point masses of μ_+ and μ_-, respectively, and let $S = \{\mathbf{x}_i^+\}_{i=1}^M \cup \{\mathbf{x}_i^-\}_{i=1}^N$. Then

$$(\mu_+ + \mu_-)(\{\mathbf{x}^*\}) \geq |(\mu_+ - \mu_-)(\{\mathbf{x}^*\})| \geq 2\pi \quad \forall \mathbf{x}^* \in S,$$

with

$$(\mu_+ + \mu_-)(\{\mathbf{x}^*\}) = |(\mu_+ - \mu_-)(\{\mathbf{x}^*\})| = 2\pi$$

for all $\mathbf{x}^* \in S \setminus (\{\mathbf{x}_i^+\}_{i=1}^M \cap \{\mathbf{x}_i^-\}_{i=1}^N)$. Furthermore, if the regular part of $\mu_+ - \mu_-$ is strictly positive, i.e., if $d_+ > 0$, then $(\mu_+ - \mu_-)(\{\mathbf{x}^*\})$ is negative for all $\mathbf{x}^* \in S$, whereas if the regular part of $\mu_+ - \mu_-$ is strictly negative, i.e., if $d_- > 0$, then $(\mu_+ - \mu_-)(\{\mathbf{x}^*\})$ is positive for all $\mathbf{x}^* \in S$. In particular, the measures $\mu_+ + \mu_-$ and $\mu_+ - \mu_-$ have the exact same set of locations with nonzero point masses. This set coincides with the "blow-up" points for the sequence $v_{\lambda_n}^0$.

Remark 4.2. In Theorem 4.1 of [6] it is stated that the set of point mass locations of $\mu = \mu_+ - \mu_-$ is finite and nonempty and that it equals the set of "blow-up" points for the sequence $v_{\lambda_n}^0$. As pointed out in the subsequent corrigendum this is not quite the statement proven in [6]. What was indeed proven was that the set of point mass locations of $\nu = \mu_+ + \mu_-$ is finite and nonempty and that this set equals the set of "blow-up" points for the sequence $v_{\lambda_n}^0$ (this statement is also included as part of Theorem 3.1 of the present paper). By showing, as we have done here, that the point mass locations of the measures $\mu_+ + \mu_-$ and $\mu_+ - \mu_-$ agree, we have indeed established the validity of the original formulation of Theorem 4.1 in [6] for simply connected domains.

Before proceeding to the proof of Theorem 4.1, we establish three lemmas which will be used in that proof as well as in the proof of our second theorem in this section (Theorem 4.6).

LEMMA 4.3. *Let v_{λ_n} be the subsequence extracted in Theorem* 3.1*. Then*

$$\lambda_n e^{v_{\lambda_n}} \to 2\mu_+ \quad \textit{and} \quad \lambda_n e^{-v_{\lambda_n}} \to 2\mu_-$$

in the sense of measures on $\partial\Omega$. The convergence takes place in $L^\infty(K)$ for any compact set $K \subset \partial\Omega \setminus S = \partial\Omega \setminus (\{\mathbf{x}_i^+\}_{i=1}^M \cup \{\mathbf{x}_i^-\}_{i=1}^N)$.

Proof. From Theorem 3.1 we know that

$$\lambda_n \sinh(v_{\lambda_n}) \to \mu_+ - \mu_- \quad \text{and} \quad \lambda_n |\sinh(v_{\lambda_n})| \to \mu_+ + \mu_- \tag{4.1}$$

in the sense of measures on $\partial\Omega$. From Theorem 3.1 we also know that the convergence takes place in $L^\infty(K)$ for any compact set $K \subset \partial\Omega \setminus S$. The identity $\cosh(x) = |\sinh(x)| + e^{-|x|}$ and the fact that $|\lambda_n e^{-|v_{\lambda_n}|}| \leq \lambda_n$ now imply that

$$\lambda_n \cosh(v_{\lambda_n}) = \lambda_n |\sinh(v_{\lambda_n})| + \lambda_n e^{-|v_{\lambda_n}|} \to \mu_+ + \mu_- \tag{4.2}$$

in the sense of measures on $\partial\Omega$. It also follows that this convergence takes place in $L^\infty(K)$ for any compact set $K \subset \partial\Omega \setminus S$. A combination of the first statement in (4.1) and the statement (4.2) immediately leads to the conclusion of this lemma. ∎

LEMMA 4.4. *Let $\mathbb{H}$ denote the half-plane $\mathbb{H} = \{(y_1, y_2) : y_2 > 0\}$, and for fixed real $\gamma \neq 0$ and β, let $l_{\gamma,\beta}$ denote the half-line $l_{\gamma,\beta} = \{y_2 = \gamma y_1 + \beta\} \cap \mathbb{H}$. Suppose F is Lebesgue integrable on $\mathbb{R}$, that is, suppose F is in $L^1(\mathbb{R})$. Then we have the following asymptotic statements:*

$$\int_{\mathbb{R}} F(z_1) \frac{y_1 - z_1}{(y_1 - z_1)^2 + y_2^2}\, dz_1 = o(1/y_2), \qquad \int_{\mathbb{R}} F(z_1) \frac{y_2}{(y_1 - z_1)^2 + y_2^2}\, dz_1 = o(1/y_2),$$
$$\text{and} \qquad \int_{\mathbb{R}} F(z_1) \frac{(y_1 - z_1) y_2}{[(y_1 - z_1)^2 + y_2^2]^2}\, dz_1 = o(1/y_2^2)$$

as $(y_1, y_2) \in \mathbb{H}$ approaches the point $(-\beta/\gamma, 0) \in \partial\mathbb{H}$ along the half-line $l_{\gamma,\beta}$.

Proof. We shall prove the first of these three statements. The proof of the other two proceed in a similar fashion but are left to the reader. Simple calculations give that for $(y_1, y_2) \in l_{\gamma,\beta}$

$$\begin{aligned}
\int_{\mathbb{R}} F(z_1) \frac{y_1 - z_1}{(y_1 - z_1)^2 + y_2^2}\, dz_1 &= \int_{\mathbb{R}} F(z_1) \frac{\frac{y_2}{\gamma} - \left(z_1 + \frac{\beta}{\gamma}\right)}{\left(\frac{y_2}{\gamma} - \left(z_1 + \frac{\beta}{\gamma}\right)\right)^2 + y_2^2}\, dz_1 \\
&= \int_{\mathbb{R}} F\left(z_1 - \frac{\beta}{\gamma}\right) \frac{\frac{y_2}{\gamma} - z_1}{\left(\frac{y_2}{\gamma} - z_1\right)^2 + y_2^2}\, dz_1 \\
&= \frac{\gamma}{y_2} \int_{\mathbb{R}} F\left(z_1 - \frac{\beta}{\gamma}\right) \frac{1 - \frac{\gamma z_1}{y_2}}{\left(1 - \frac{\gamma z_1}{y_2}\right)^2 + \gamma^2}\, dz_1.
\end{aligned} \tag{4.3}$$

From the inequality $|s/(s^2 + \gamma^2)| \leq 1/2\gamma$ it follows that

$$\left| F\left(z_1 - \frac{\beta}{\gamma}\right) \frac{1 - \frac{\gamma z_1}{y_2}}{\left(1 - \frac{\gamma z_1}{y_2}\right)^2 + \gamma^2} \right| \leq \frac{1}{2\gamma} \left| F\left(z_1 - \frac{\beta}{\gamma}\right) \right|.$$

Since the right-hand side is an integrable function, and since

$$F\left(z_1 - \frac{\beta}{\gamma}\right) \frac{1 - \frac{\gamma z_1}{y_2}}{\left(1 - \frac{\gamma z_1}{y_2}\right)^2 + \gamma^2} \to 0 \quad \text{a.e. in } z_1$$

as $y_2 \to 0$, it now follows from Lebesgue's dominated convergence theorem that

$$\int_{\mathbb{R}} F\left(z_1 - \frac{\beta}{\gamma}\right) \frac{1 - \frac{\gamma z_1}{y_2}}{\left(1 - \frac{\gamma z_1}{y_2}\right)^2 + \gamma^2}\, dz_1 \to 0 \quad \text{as } y_2 \to 0.$$

By a combination with (4.3) we immediately get the first asymptotic statement of this lemma. ∎

At several points in this section we shall use the notion of conformal equivalence. We shall call a smooth mapping $\Phi : \Omega \to \mathbb{R}^2$ a conformal equivalence if and only if the following four conditions are satisfied:

(a) $D\Phi(x)$ is a similarity, i.e., $|D\Phi(x)\xi| = k(x)|\xi|$, $k(x) > 0$, $\forall\, \xi \in \mathbb{R}^2$, and $x \in \Omega$,
(b) $k(x) = |\det(D\Phi(x))|^{1/2} \geq k_0 > 0\ \forall x \in \Omega$,
(c) the mapping Φ is injective and may be extended as an injective mapping: $\overline{\Omega} \to \mathbb{R}^2 \cup \{\infty\}$, and
(d) the extended mapping Φ is either smooth or there exists a point z^* (in $\mathbb{R}^2 \setminus \Phi(\overline{\Omega})$) such that the mapping $\frac{(\cdot - z^*)}{|\cdot - z^*|^2} \circ \Phi(x)$ has a smooth extension to $\overline{\Omega}$.

We note that conditions (a)–(d) imply that $\langle D\Phi(x)\xi, D\Phi(x)\eta\rangle = k^2(x)\langle\xi,\eta\rangle = |\det(D\Phi(x))|\langle\xi,\eta\rangle\ \forall\, \xi,\ \eta$ in $\mathbb{R}^2$ and all x in $\overline{\Omega} \setminus \{\Phi^{-1}(\infty)\}$. We also note that for any smooth, bounded, simply connected domain we may construct a conformal equivalence of Ω onto the upper half-plane. The point $\Phi^{-1}(\infty)$ may be picked arbitrarily on $\partial\Omega$. This follows from Riemann's mapping theorem and subsequent composition with a linear fractional transformation.

The following representation result will prove extremely useful.

LEMMA 4.5. *Let Φ be a conformal equivalence of the smooth, bounded, simply connected domain Ω onto the half-plane $\mathbb{H} = \{(y_1, y_2) : y_2 > 0\}$, constructed so that the point $\Phi^{-1}(\infty)$ lies in $\partial\Omega \setminus (\{\mathbf{x}_i^+\}_{i=1}^M \cup \{\mathbf{x}_i^-\}_{i=1}^N)$, where $\mathbf{x}_i^\pm$ are the "blow-up" points from Theorem* 3.1. *Let v_0 denote the limit from Theorem* 3.1, *i.e.,*

$$v_0 = \lim v_{\lambda_n}^0 = \lim \left(v_{\lambda_n} - \frac{1}{|\partial\Omega|} \int_{\partial\Omega} v_{\lambda_n}\ d\sigma \right),$$

and define $u_0 := v_0 \circ \Phi^{-1}$. Let F denote the function

$$F(y_1) = \left(\frac{d_+}{2} e^{u_0(y_1,0)} - \frac{d_-}{2} e^{-u_0(y_1,0)} \right) h(y_1, 0),$$

with

$$h(y) = |\det(D\Phi(\Phi^{-1}(y)))|^{-1/2}$$

and $d_\pm$ as in Theorem 3.1. *Then $F \in L^1(\mathbb{R}, \log(|x|+2)dx)$, and the function u_0 and its derivatives have the representation formulas*

$$\begin{aligned} u_0(y) = v_0(\Phi^{-1}(\infty)) &- \frac{1}{\pi} \sum_{i=1}^M \alpha_i^+ \log|y - \mathbf{y}_i^+| + \frac{1}{\pi} \sum_{i=1}^N \alpha_i^- \log|y - \mathbf{y}_i^-| \\ &- \frac{1}{\pi} \int_{\mathbb{R}} F(z_1) \log|y - (z_1, 0)|\ dz_1, \quad y \in \mathbb{H}, \\ \frac{\partial u_0}{\partial y_j}(y) = &- \frac{1}{\pi} \sum_{i=1}^M \alpha_i^+ \frac{(y - \mathbf{y}_i^+)_j}{|y - \mathbf{y}_i^+|^2} + \frac{1}{\pi} \sum_{i=1}^N \alpha_i^- \frac{(y - \mathbf{y}_i^-)_j}{|y - \mathbf{y}_i^-|^2} \\ &- \frac{1}{\pi} \int_{\mathbb{R}} F(z_1) \frac{(y - (z_1, 0))_j}{|y - (z_1, 0)|^2}\ dz_1, \quad y \in \mathbb{H},\ j = 1, 2. \end{aligned}$$

The coefficients $\alpha_i^\pm$ are as in Theorem 3.1, *and $\mathbf{y}_i^\pm = \Phi(\mathbf{x}_i^\pm)$.*

Proof. For any $z \in \mathbb{H}$ let $\overline{z}$ denote the "reflection in $\partial\mathbb{H}$," i.e., $\overline{z} = \overline{(z_1, z_2)} = (z_1, -z_2)$. A simple calculation shows that the function $G(x, w)$, $x, w \in \Omega$, defined by

$$G(x, w) = \frac{1}{2\pi} \log|\Phi(x) - \Phi(w)| + \frac{1}{2\pi} \log|\Phi(x) - \overline{\Phi(w)}|$$

is indeed the solution to

$$\Delta_x G(x,w) = \delta_w \quad \text{in } \Omega, \qquad \frac{\partial}{\partial \mathbf{n}_x} G(x,w) = \delta_{w^*} \quad \text{on } \partial\Omega,$$

with $w^* = \Phi^{-1}(\infty)$. Using the representation formula for v_0 from Theorem 3.1 we get

$$\int_{\partial\Omega} G(x,w)\, d(\mu_+ - \mu_-)_x = \int_{\partial\Omega} (N(x,w) - N(x,w^*))\, d(\mu_+ - \mu_-)_x = -v_0(w) + v_0(w^*).$$

For the identity $v_0(w^*) = -\int_{\partial\Omega} N(x,w^*)\, d(\mu_+ - \mu_-)_x$ we rely on the boundary limit of the representation formula from Theorem 3.1, which remains valid due to the fact that the measure $\mu_+ - \mu_-$ is given by a C^∞ density near the point w^*. By rearrangement of the above formula,

$$\begin{aligned} v_0(w) &= v_0(w^*) - \int_{\partial\Omega} G(x,w)\, d(\mu_+ - \mu_-)_x \\ &= v_0(w^*) - \sum_{i=1}^{M} \alpha_i^+ G(\mathbf{x}_i^+, w) + \sum_{i=1}^{N} \alpha_i^- G(\mathbf{x}_i^-, w) \\ &\quad - \int_{\partial\Omega} \left(\frac{d_+}{2} e^{v_0} - \frac{d_-}{2} e^{-v_0} \right) G(x,w)\, d\sigma_x. \end{aligned}$$

Introducing $y = \Phi(w)$ and $z = \Phi(x)$ and changing variable of integration (from x to z) we immediately obtain the first representation formula of this lemma. The second formula follows by differentiation. The fact that $F \in L^1(\mathbb{R}, \log(|z|+2)dz)$ follows immediately from the finiteness of the last integral in the above integral identity. ∎

Proof of Theorem 4.1. Let Φ be a conformal equivalence of Ω onto the half-plane $\mathbb{H}$ such that the point $\Phi^{-1}(\infty)$ lies in $\partial\Omega \setminus (\{\mathbf{x}_i^+\}_{i=1}^M \cup \{\mathbf{x}_i^-\}_{i=1}^N)$. Setting $u_{\lambda_n} = v_{\lambda_2} \circ \Phi^{-1}$, we have a family of solutions to

$$\begin{cases} \Delta u_{\lambda_n} = 0 & \text{in } \mathbb{H}, \\ \dfrac{\partial u_{\lambda_n}}{\partial y_2} = -\lambda_n h(y) \sinh(u_{\lambda_n}) & \text{on } \partial\mathbb{H}, \end{cases} \tag{4.4}$$

with $h(y) = |\det(D\Phi(\Phi^{-1}(y)))|^{-1/2}$. The sequence

$$u_{\lambda_n} - s_{\lambda_n}, \quad \text{with } s_{\lambda_n} = \int_{\partial\Omega} v_{\lambda_n}\, d\sigma/|\partial\Omega|,$$

converges to $u_0 = v_0 \circ \Phi^{-1}$ in $H^t(\mathbb{H} \cap \{|y| \le R\})$ for any $t < 1$ and any R; the convergence also takes place in $C^\infty(\overline{\mathbb{H}} \setminus (\{\mathbf{y}_i^+\}_{i=1}^M \cup \{\mathbf{y}_i^-\}_{i=1}^N))$, i.e., in $C^\infty(K)$ for any compact set $K \subset \overline{\mathbb{H}} \setminus (\{\mathbf{y}_i^+\}_{i=1}^M \cup \{\mathbf{y}_i^-\}_{i=1}^N)$. We now introduce functions w_{λ_n} and w_0 by

$$w_{\lambda_n} := \partial_{y_1} u_{\lambda_n} \partial_{y_2} u_{\lambda_n} \quad \text{and} \quad w_0 := \partial_{y_1} u_0 \partial_{y_2} u_0,$$

respectively. Due to the $C^\infty(\overline{\mathbb{H}} \setminus \{\mathbf{y}_i^\pm\})$ convergence of $u_{\lambda_n} - s_{\lambda_n}$ toward u_0, the sequence w_{λ_n} converges in $C^\infty(\overline{\mathbb{H}} \setminus (\{\mathbf{y}_i^+\}_{i=1}^M \cup \{\mathbf{y}_i^-\}_{i=1}^N))$ toward w_0.

Let $\mathbf{y}^* = (y_1^*, 0)$ be one of the points from $\{\mathbf{y}_i^+\}_{i=1}^M \cup \{\mathbf{y}_i^-\}_{i=1}^N$, and given any $\gamma \neq 0$, let $l_{\mathbf{y}^*}$ denote the half-line $l_{\mathbf{y}^*} = \{(y_1, y_2) : y_2 = \gamma y_1 - \gamma y_1^*\} \cap \mathbb{H}$. Suppose there is a point mass contribution $\alpha_*^+ \delta_{\mathbf{x}^*}$ to μ_+ and point mass contribution $\alpha_*^- \delta_{\mathbf{x}^*}$

to μ_- from the point $\mathbf{x}^* = \Phi^{-1}(\mathbf{y}^*)$. This includes the possibility that one of the $\alpha_*^{\pm}$ could be zero, corresponding to $\mathbf{y}^* \notin \{\mathbf{y}_i^+\}_{i=1}^M \cap \{\mathbf{y}_i^-\}_{i=1}^N$. From a combination of Lemmas 4.4 and 4.5 we conclude that

$$\begin{aligned}\frac{\partial u_0}{\partial y_1}(y) &= \frac{1}{\pi}(-\alpha_*^+ + \alpha_*^-)\frac{y_1 - y_1^*}{|y - \mathbf{y}^*|^2} + o(1/y_2)\\ &= \frac{1}{\pi}(-\alpha_*^+ + \alpha_*^-)\frac{\gamma}{1+\gamma^2}\frac{1}{y_2} + o(1/y_2) \quad \text{and}\\ \frac{\partial u_0}{\partial y_2}(y) &= \frac{1}{\pi}(-\alpha_*^+ + \alpha_*^-)\frac{y_2}{|y - \mathbf{y}^*|^2} + o(1/y_2)\\ &= \frac{1}{\pi}(-\alpha_*^+ + \alpha_*^-)\frac{\gamma^2}{1+\gamma^2}\frac{1}{y_2} + o(1/y_2)\end{aligned}$$

as y approaches the point $\mathbf{y}^*$ along the half-line $l_{\mathbf{y}^*}$. As a consequence,

$$w_0(y) = \frac{(\alpha_*^+ - \alpha_*^-)^2}{\pi^2}\frac{\gamma^3}{(1+\gamma^2)^2}\frac{1}{y_2^2} + o(1/y_2^2) \tag{4.5}$$

as y approaches the point $\mathbf{y}^*$ along the half-line $l_{\mathbf{y}^*}$ (and so y_2 approaches 0).

We now proceed to analyze the same asymptotic scenario, using the relationship $w_0(y) = \lim w_{\lambda_n}(y)$, which holds for any point $y \in l_{\mathbf{y}^*}$. Simple calculations yield

$$\Delta w_{\lambda_n} = 0 \quad \text{in } \mathbb{H}$$

and

$$w_{\lambda_n} = -\lambda_n h(y_1,0)\sinh(u_{\lambda_n})\partial_{y_1}u_{\lambda_n} = -\lambda_n h(y_1,0)\partial_{y_1}\left(\cosh(u_{\lambda_n})\right) \quad \text{on } \partial\mathbb{H}, \tag{4.6}$$

with $h(y) = |\det(D\Phi(\Phi^{-1}(y)))|^{-1/2}$. Let $D \subset \mathbb{H}$ be a bounded, smooth domain with $\Gamma_0 = \partial D \cap \partial\mathbb{H} = [-R,R]\times\{0\}$ (for instance, take D to be the half-disk $B_R(0)\cap\mathbb{H}$ with the two corners "smoothed out"). Choose R sufficiently large that all the points $\mathbf{y}_i^{\pm}$ lie strictly inside $\frac{1}{2}\Gamma_0$. Let $G_D(y,z)$ denote the Green's function for the domain D, i.e., the solution to

$$\begin{cases}\Delta G_D(\cdot,z) = \delta_z & \text{in } D,\\ G_D(\cdot,z) = 0 & \text{on } \partial D.\end{cases} \tag{4.7}$$

For any fixed $z \in D$ the harmonic function w_{λ_n} may now be represented as

$$w_{\lambda_n}(z) = \int_{\partial D} w_{\lambda_n}(y)\frac{\partial G_D}{\partial \mathbf{n}_y}(y,z)\, d\sigma_y.$$

We decompose the boundary of D as follows: $\partial D = \Gamma_0 \cup \Gamma_1$, with $\Gamma_0 = \partial D \cap \partial\mathbb{H}$ and $\Gamma_1 = \partial D \cap \mathbb{H}$. In light of (4.6), the above integral representation for w_{λ_n} reads

$$\begin{aligned}w_{\lambda_n}(z) &= \int_{\Gamma_1} w_{\lambda_n}(y)\frac{\partial G_D}{\partial \mathbf{n}_y}(y,z)d\sigma_y\\ &\quad + \int_{\Gamma_0}\lambda_n h(y_1,0)\partial_{y_1}\left(\cosh(u_{\lambda_n}(y_1,0))\right)\frac{\partial G_D}{\partial y_2}((y_1,0),z)dy_1\\ &= I_{1,\lambda_n}(z) + I_{2,\lambda_n}(z).\end{aligned} \tag{4.8}$$

Since the sequence w_{λ_n} converges in $C^\infty(\overline{\mathbb{H}} \setminus (\{\mathbf{y}_i^+\}_{i=1}^M \cup \{\mathbf{y}_i^-\}_{i=1}^N))$ toward w_0

$$\lim_{\lambda_n \to 0} I_{1,\lambda_n}(z) = \int_{\Gamma_1} w_0(y) \frac{\partial G}{\partial \mathbf{n}_y}(y,z) d\sigma_y \quad \forall z \in D. \tag{4.9}$$

Integration by parts yields

$$\begin{aligned} I_{2,\lambda_n}(z) = &- \int_{\Gamma_0} \lambda_n \cosh(u_{\lambda_n}(y_1,0)) h(y_1,0) \partial_{y_1} \left(\frac{\partial G_D}{\partial y_2}((y_1,0),z) \right) dy_1 \\ &- \int_{\Gamma_0} \lambda_n \cosh(u_{\lambda_n}(y_1,0)) \partial_{y_1} h(y_1,0) \frac{\partial G_D}{\partial y_2}((y_1,0),z) dy_1 \\ &+ \left(\lambda_n \cosh(u_{\lambda_n}(y_1,0)) h(y_1,0) \frac{\partial G_D}{\partial y_2}((y_1,0),z) \right) \Bigg|_{y_1=-R}^{y_1=R}. \end{aligned} \tag{4.10}$$

From Lemma 4.3 we know that

$$\lambda_n e^{v_{\lambda_n}} \to 2\mu_+ \quad \text{and} \quad \lambda_n e^{-v_{\lambda_n}} \to 2\mu_-,$$

in the sense of measures on $\partial\Omega$, and so

$$\begin{aligned} \lambda_n \cosh(v_{\lambda_n}) &\to \mu_+ + \mu_- \\ &= \sum_{i=1}^M \alpha_i^+ \delta_{\mathbf{x}_i^+} + \sum_{i=1}^N \alpha_i^- \delta_{\mathbf{x}_i^-} + \frac{d_+}{2} e^{v_0} + \frac{d_-}{2} e^{-v_0}, \end{aligned}$$

in the sense of measures on $\partial\Omega$. In the last case, the left-hand side converges uniformly to the C^∞ function $\frac{d_+}{2} e^{v_0} + \frac{d_-}{2} e^{-v_0}$ away from the points $\mathbf{x}_i^\pm$. When "pushed forward" by the conformal map Φ this last convergence statement translates into

$$\lambda_n h(y_1,0) \cosh(u_{\lambda_n}(y_1,0)) \to \sum_{i=1}^M \alpha_i^+ \delta_{\mathbf{y}_i^+} + \sum_{i=1}^N \alpha_i^- \delta_{\mathbf{y}_i^-} + E(y_1) \tag{4.11}$$

in the sense of measures on $\Gamma_0 = [-R,R] \times \{0\}$, with E given by

$$E(y_1) = \left(\frac{d_+}{2} e^{u_0(y_1,0)} + \frac{d_-}{2} e^{-u_0(y_1,0)} \right) h(y_1,0).$$

As a consequence of (4.10) and (4.11) we immediately obtain the following limit for the integrals I_{2,λ_n}:

$$\begin{aligned} &\lim_{\lambda_n \to 0} I_{2,\lambda_n}(z) \\ &\quad = - \sum_{i=1}^M \alpha_i^+ \frac{\partial^2 G_D}{\partial y_1 \partial y_2}(\mathbf{y}_i^+, z) - \sum_{i=1}^N \alpha_i^- \frac{\partial^2 G_D}{\partial y_1 \partial y_2}(\mathbf{y}_i^-, z) \\ &\qquad - \sum_{i=1}^M \alpha_i^+ \frac{\partial_{y_1} h(\mathbf{y}_i^+)}{h(\mathbf{y}_i^+)} \frac{\partial G_D}{\partial y_2}(\mathbf{y}_i^+, z) - \sum_{i=1}^N \alpha_i^- \frac{\partial_{y_1} h(\mathbf{y}_i^-)}{h(\mathbf{y}_i^-)} \frac{\partial G_D}{\partial y_2}(\mathbf{y}_i^-, z) \\ &\qquad - \int_{\Gamma_0} E(y_1) \left(\frac{\partial^2 G_D}{\partial y_1 \partial y_2}((y_1,0),z) + \frac{\partial_{y_1} h(y_1,0)}{h(y_1,0)} \frac{\partial G_D}{\partial y_2}((y_1,0),z) \right) dy_1 \\ &\qquad + E(y_1) \frac{\partial G_D}{\partial y_2}((y_1,0),z) \Bigg|_{y_1=-R}^{y_1=R} \quad \forall z \in D. \end{aligned} \tag{4.12}$$

Here we also used the fact that $\lambda_n \cosh(u_{\lambda_n}(y_1,0))h(y_1,0)$ converges pointwise to E away from the points $\mathbf{y}_i^{\pm}$, in order to treat the boundary term in (4.10). A combination of (4.9) and (4.12) with (4.8) now yields

$$\begin{aligned} w_0(z) &= \lim_{\lambda_n \to 0} w_{\lambda_n}(z) \\ &= \int_{\Gamma_1} w_0(y) \frac{\partial G_D}{\partial \mathbf{n}_y}(y,z) d\sigma_y \\ &\quad - \sum_{i=1}^{M} \alpha_i^+ \frac{\partial^2 G_D}{\partial y_1 \partial y_2}(\mathbf{y}_i^+, z) - \sum_{i=1}^{N} \alpha_i^- \frac{\partial^2 G_D}{\partial y_1 \partial y_2}(\mathbf{y}_i^-, z) \\ &\quad - \sum_{i=1}^{M} \alpha_i^+ \frac{\partial_{y_1} h(\mathbf{y}_i^+)}{h(\mathbf{y}_i^+)} \frac{\partial G_D}{\partial y_2}(\mathbf{y}_i^+, z) - \sum_{i=1}^{N} \alpha_i^- \frac{\partial_{y_1} h(\mathbf{y}_i^-)}{h(\mathbf{y}_i^-)} \frac{\partial G_D}{\partial y_2}(\mathbf{y}_i^-, z) \\ &\quad - \int_{\Gamma_0} E(y_1) \left(\frac{\partial^2 G_D}{\partial y_1 \partial y_2}((y_1,0),z) + \frac{\partial_{y_1} h(y_1,0)}{h(y_1,0)} \frac{\partial G_D}{\partial y_2}((y_1,0),z) \right) dy_1 \\ &\quad + E(y_1) \frac{\partial G_D}{\partial y_2}((y_1,0),z) \Big|_{y_1=-R}^{y_1=R} \qquad \forall z \in D. \end{aligned} \tag{4.13}$$

Set $\omega = \{z \in D : |z_1| < R/2 \text{ and } 0 < z_2 < \epsilon\}$ for ϵ fixed, but sufficiently small. The Green's function $G_D(y,z)$, $(y,z) \in \overline{D \times \omega} \setminus \{y = z\}$, may now be written

$$G_D(y,z) = \frac{1}{2\pi} \log|y - z| - \frac{1}{2\pi} \log|y - \overline{z}| + g(y,z),$$

where $\overline{z} = \overline{(z_1, z_2)} = (z_1, -z_2)$ and where the function g is in $C^\infty(\overline{D \times \omega})$. We thus compute

$$\begin{aligned} \frac{\partial G_D}{\partial y_2}(y,z) &= -\frac{1}{\pi} \frac{z_2}{|y-z|^2} + \partial_{y_2} g(y,z) \quad \text{and} \\ \frac{\partial^2 G_D}{\partial y_1 \partial y_2}(y,z) &= \frac{2}{\pi} \frac{(y-z)_1 z_2}{|y-z|^4} + \partial_{y_1} \partial_{y_2} g(y,z) \end{aligned} \tag{4.14}$$

for $(y,z) \in \Gamma_0 \times \overline{\omega} \setminus \{y = z\}$. Substituting (4.14) into (4.13) we arrive at

$$\begin{aligned} w_0(z) &= \frac{2}{\pi} \sum_{i=1}^{M} \alpha_i^+ \frac{(z - \mathbf{y}_i^+)_1 z_2}{|z - \mathbf{y}_i^+|^4} + \frac{2}{\pi} \sum_{i=1}^{N} \alpha_i^- \frac{(z - \mathbf{y}_i^-)_1 z_2}{|z - \mathbf{y}_i^-|^4} \\ &\quad + \frac{1}{\pi} \sum_{i=1}^{M} \alpha_i^+ \frac{\partial_{y_1} h(\mathbf{y}_i^+)}{h(\mathbf{y}_i^+)} \frac{z_2}{|z - \mathbf{y}_i^+|^2} + \frac{1}{\pi} \sum_{i=1}^{N} \alpha_i^- \frac{\partial_{y_1} h(\mathbf{y}_i^-)}{h(\mathbf{y}_i^-)} \frac{z_2}{|z - \mathbf{y}_i^-|^2} \\ &\quad + \frac{1}{\pi} \int_{\Gamma_0} E(y_1) \left(2 \frac{(z_1 - y_1) z_2}{|z - (y_1,0)|^4} + \frac{\partial_{y_1} h(y_1,0)}{h(y_1,0)} \frac{z_2}{|z - (y_1,0)|^2} \right) dy_1 \\ &\quad + R(z), \end{aligned} \tag{4.15}$$

where R is in $C^\infty(\overline{\omega})$. Let $\mathbf{y}^* = (y_1^*, 0)$ be one of the points from $\{\mathbf{y}_i^+\}_{i=1}^M \cup \{\mathbf{y}_i^-\}_{i=1}^N$ (the same as before) and let $l_{\mathbf{y}^*}$ denote the half-line $l_{\mathbf{y}^*} = \{(z_1, z_2) : z_2 = \gamma z_1 - \gamma y_1^*\} \cap \mathbb{H}$. Suppose (as before) that there is a point mass contribution $\alpha_*^+ \delta_{\mathbf{x}^*}$ to μ_+ and point mass contribution $\alpha_*^- \delta_{\mathbf{x}^*}$ to μ_- from the point $\mathbf{x}^* = \Phi^{-1}(\mathbf{y}^*)$ (so there are terms with

coefficients α_*^+ and α_*^- corresponding to $\mathbf{y}^*$ in the respective sums above). From a combination of Lemma 4.4 with (4.15) we conclude that

$$\begin{aligned} w_0(z) &= \frac{2}{\pi}(\alpha_*^+ + \alpha_*^-)\frac{(z_1 - y_1^*)z_2}{|z - y^*|^4} + o(1/z_2^2) \\ &= \frac{2}{\pi}(\alpha_*^+ + \alpha_*^-)\frac{\gamma^3}{(1+\gamma^2)^2}\frac{1}{z_2^2} + o(1/z_2^2) \end{aligned} \tag{4.16}$$

as z approaches $\mathbf{y}^*$ along the half-line $l_{\mathbf{y}^*}$ (and therefore z_2 approaches 0). By comparison of the two alternate asymptotic representations, (4.5) and (4.16), for w_0, we infer that

$$\frac{(\alpha_*^+ - \alpha_*^-)^2}{\pi^2} = \frac{2}{\pi}(\alpha_*^+ + \alpha_*^-), \tag{4.17}$$

which immediately yields

$$\alpha_*^+ + \alpha_*^- \geq |\alpha_*^+ - \alpha_*^-| = 2\pi\frac{\alpha_*^+ + \alpha_*^-}{|\alpha_*^+ - \alpha_*^-|} \geq 2\pi$$

or

$$(\mu_+ + \mu_-)(\{\mathbf{x}^*\}) \geq |(\mu_+ - \mu_-)(\{\mathbf{x}^*\})| \geq 2\pi, \tag{4.18}$$

as stated in the formulation of this theorem. If $\mathbf{x}^* \in S \setminus (\{\mathbf{x}_i^+\}_{i=1}^M \cap \{\mathbf{x}_i^-\}_{i=1}^N)$, so that either α_*^+ or α_*^- is zero, then it follows from (4.17) that

$$(\mu_+ + \mu_-)(\{\mathbf{x}^*\}) = |(\mu_+ - \mu_-)(\{\mathbf{x}^*\})| = 2\pi.$$

The statement about the coincidence of the point mass locations of the measures $\mu_+ + \mu_-$ and $\mu_+ - \mu_-$ is a direct consequence of the inequalities (4.18), valid for all $\mathbf{x}^* \in S$.

Let us go back to the representation formula for u_0 from Lemma 4.5,

$$\begin{aligned} u_0(y) = &-\frac{1}{\pi}\sum_{i=1}^M \alpha_i^+ \log|y - \mathbf{y}_i^+| + \frac{1}{\pi}\sum_{i=1}^N \alpha_i^- \log|y - \mathbf{y}_i^-| \\ &-\frac{1}{\pi}\int_{\mathbb{R}} F(z_1)\log|y - (z_1,0)|\, dz_1 + v_0(\Phi^{-1}(\infty)), \quad y \in \mathbb{H}, \end{aligned}$$

with $F(z_1) = (\frac{d_+}{2}e^{u_0(z_1,0)} - \frac{d_-}{2}e^{-u_0(z_1,0)})h(z_1,0)$. By taking the limit as y approaches points $(y_1,0) \in \partial\mathbb{H} \setminus (\{\mathbf{y}_i^+\}_{i=1}^M \cup \{\mathbf{y}_i^-\}_{i=1}^N)$ we obtain

$$\begin{aligned} u_0(y_1,0) = &-\frac{1}{\pi}\sum_{i=1}^M \alpha_i^+ \log|(y_1,0) - \mathbf{y}_i^+| + \frac{1}{\pi}\sum_{i=1}^N \alpha_i^- \log|(y_1,0) - \mathbf{y}_i^-| \\ &-\frac{1}{\pi}\int_{\mathbb{R}} F(z_1)\log|y_1 - z_1|\, dz_1 + v_0(\Phi^{-1}(\infty)). \end{aligned}$$

Let $\mathbf{y}^* = (y_1^*, 0)$ be one of the points of $S = \{\mathbf{y}_i^+\}_{i=1}^M \cup \{\mathbf{y}_i^-\}_{i=1}^N$, and define $d_0 = \min\{\mathrm{dist}(\mathbf{y}^*, S \setminus \mathbf{y}^*), 1\}$. The above representation formula now yields

$$\begin{aligned} u_0(y_1,0) = &-\frac{1}{\pi}(\alpha_*^+ - \alpha_*^-)\log|y_1 - y_1^*| \\ &-\frac{1}{\pi}\int_{|z_1 - y_1^*| < d_0/2} F(z_1)\log|y_1 - z_1|\, dz_1 + O(1) \end{aligned} \tag{4.19}$$

for $0 < |y_1 - y_1^*| < d_0/4$. Recall that $F \in L^1(\mathbb{R}, \log(|x|+2)dx)$.

Now let us consider the case $d_+ > 0$. In that case $d_- = 0$, and so

$$F(z_1) = \frac{d_+}{2} e^{u_0(z_1,0)} |\det(D\Phi(\Phi^{-1}(z_1,0)))|^{-1/2} > 0.$$

Since $\log|y_1 - z_1| < 0$ whenever $|z_1 - y_1^*| < d_0/2$ and $|y_1 - y_1^*| < d_0/4$, we then obtain

$$-\frac{1}{\pi}\int_{|z_1 - y_1^*| < d_0/2} F(z_1) \log|y_1 - z_1| \, dz_1 > 0 \quad \text{for } 0 < |y_1 - y_1^*| < d_0/4.$$

By insertion into (4.19) this yields

$$u_0(y_1,0) \geq \frac{1}{\pi}(\alpha_*^+ - \alpha_*^-) \log|y_1 - y_1^*|^{-1} - C$$

for $0 < |y_1 - y_1^*| < d_0/4$. Now suppose $(\mu_+ - \mu_-)(\{\mathbf{x}^*\}) = \alpha_*^+ - \alpha_*^- \geq 2\pi$. Then it follows immediately that

$$u_0(y_1,0) \geq \log|y_1 - y_1^*|^{-2} - C \quad \text{for } 0 < |y_1 - y_1^*| < d_0/4.$$

As a consequence,

$$e^{u_0(y_1,0)} \geq c|y_1 - y_1^*|^{-2} \quad \text{for } 0 < |y_1 - y_1^*| < d_0/4.$$

However, this contradicts the fact that $e^{u_0(\cdot,0)} \in L^1_{loc}(\mathbb{R})$ ($e^{v_0} \in L^1(\partial\Omega)$). Since we already know that $|\alpha_*^+ - \alpha_*^-| \geq 2\pi$, we may thus conclude that

$$(\mu_+ - \mu_-)(\{\mathbf{x}^*\}) = \alpha_*^+ - \alpha_*^- \leq -2\pi$$

if d_+ is positive. The argument to show that

$$(\mu_+ - \mu_-)(\{\mathbf{x}^*\}) = \alpha_*^+ - \alpha_*^- \geq 2\pi$$

if d_- is positive proceeds similarly. □

As our last result in this paper, we establish a theorem which provides more precise information about the location of the point masses in the case when the limiting measures are "pure" sums of such point masses, i.e., when $d_+ = d_- = 0$. An extension of the proof used to verify this result may be used to derive more precise information about the point mass locations also in the case when either d_+ or d_- is nonzero (see [9]). Since the results obtained are most complete if there is assumed to be no overlap between the points $\{\mathbf{x}_i^+\}$ and $\{\mathbf{x}_i^-\}$, we formulate only the theorem under this assumption. In Remark 4.10 we describe what the corresponding results are without this assumption.

THEOREM 4.6. *Suppose the domain $\Omega \subset \mathbb{R}^2$ is smooth, bounded, and simply connected. Let v_{λ_n}, $\lambda_n \to 0^+$, be the subsequence of solutions to the nonlinear elliptic Neumann problem* (2.1) *extracted in Theorem* 3.1, *for which*

$$\lambda_n \sinh(v_{\lambda_n}^+) = \lambda_n \sinh(v_{\lambda_n})^+ \to \mu_+, \qquad \lambda_n \sinh(v_{\lambda_n}^-) = \lambda_n \sinh(v_{\lambda_n})^- \to \mu_-$$

in the sense of measures on $\partial\Omega$. Suppose that $d_+ = d_- = 0$, i.e., suppose

$$\mu_+ = \sum_{i=1}^M \alpha_i^+ \delta_{\mathbf{x}_i^+}, \qquad \mu_- = \sum_{i=1}^N \alpha_i^- \delta_{\mathbf{x}_i^-}, \tag{4.20}$$

and furthermore, suppose $\{\mathbf{x}_i^+\} \cap \{\mathbf{x}_i^-\} = \emptyset$. *Then*

$$M = N(\geq 1), \quad \alpha_i^+ = \alpha_i^- = 2\pi, \quad 1 \leq i \leq M, \quad \textit{and}$$
$$\lim_{\lambda_n \to 0} \lambda_n \int_{\partial\Omega} |\sinh(v_{\lambda_n})| \, d\sigma = 4M\pi.$$

Set $v^0_{\lambda_n} = v_{\lambda_n} - \frac{1}{|\partial\Omega|}\int_{\partial\Omega} v_{\lambda_n} \, d\sigma$ *and let* v_0 *denote the limit* $v_0 = \lim_{\lambda_n \to 0} v^0_{\lambda_n}$, *whose existence is guaranteed by Theorem* 3.1. *The function* v_0 *satisfies*

$$\Delta v_0 = 0 \quad \textit{in } \Omega, \qquad \frac{\partial v_0}{\partial \mathbf{n}} = 2\pi \sum_{i=1}^M \delta_{\mathbf{x}_i^+} - 2\pi \sum_{i=1}^M \delta_{\mathbf{x}_i^-} \quad \textit{on } \partial\Omega, \quad \int_{\partial\Omega} v_0 \, d\sigma = 0$$

in the sense that

$$v_0(x) = H(x) - 2\sum_{i=1}^M \log|x - \mathbf{x}_i^+| + 2\sum_{i=1}^M \log|x - \mathbf{x}_i^-|,$$

where $H \in C^\infty(\overline{\Omega})$ *is the classical solution to*

$$\Delta H = 0 \quad \textit{in } \Omega, \qquad \frac{\partial H}{\partial \mathbf{n}} = 2\sum_{i=1}^M \frac{(x - \mathbf{x}_i^+)\cdot \mathbf{n}}{|x - \mathbf{x}_i^+|^2} - 2\sum_{i=1}^M \frac{(x - \mathbf{x}_i^-)\cdot \mathbf{n}}{|x - \mathbf{x}_i^-|^2} \quad \textit{on } \partial\Omega,$$
$$\int_{\partial\Omega} H \, d\sigma = 2\sum_{i=1}^M \int_{\partial\Omega} \log|x - \mathbf{x_i}^+| \, d\sigma_x - 2\sum_{i=1}^M \int_{\partial\Omega} \log|x - \mathbf{x}_i^-| \, d\sigma_x.$$

The $2M$ *points* $\{\mathbf{x}_i^+\}_{i=1}^M \cup \{\mathbf{x}_i^-\}_{i=1}^M \subset \partial\Omega$ *satisfy the equations*

$$\begin{aligned} &\frac{\partial}{\partial \tau_x}(v_0(x) + 2\log|x - \mathbf{x}_i^+|)|_{x=\mathbf{x}_i^+} = 0, \quad i = 1, \dots, M, \quad \textit{and} \\ &\frac{\partial}{\partial \tau_x}(v_0(x) - 2\log|x - \mathbf{x}_i^-|)|_{x=\mathbf{x}_i^-} = 0, \quad i = 1, \dots, M, \end{aligned} \tag{4.21}$$

where $\frac{\partial}{\partial \tau_x}$ *denotes a tangential derivative to* $\partial\Omega$.

Remark 4.7. Since $\frac{\partial v_0}{\partial \mathbf{n}} = 2\pi\sum_{i=1}^M \delta_{\mathbf{x}_i^+} - 2\pi\sum_{i=1}^M \delta_{\mathbf{x}_i^-}$ on $\partial\Omega$ we easily calculate that

$$\frac{\partial}{\partial \mathbf{n}_x}(v_0 \pm 2\log|x - \mathbf{x}_i^\pm|)_{x=\mathbf{x}_i^\pm} = \lim_{\substack{x \in \partial\Omega \\ x \to \mathbf{x}_i^\pm}} \pm \frac{2(x - \mathbf{x}_i^\pm)\cdot \mathbf{n}(x)}{|x - \mathbf{x}_i^\pm|^2} = \pm\kappa(\mathbf{x}_i^\pm),$$

where $\kappa(x)$ is the (signed) curvature of $\partial\Omega$ at the point x. If we supplement the identities (4.21) with these identities we obtain

$$\nabla(v_0(x) \pm 2\log|x - \mathbf{x}_i^\pm|)|_{x=\mathbf{x}_i^\pm} = \pm\kappa(\mathbf{x}_i^\pm)\mathbf{n}(\mathbf{x}_i^\pm). \tag{4.22}$$

This is the complete analog of the identities derived in [11] for the singularity locations for the somewhat related problem $\Delta v_\lambda = -\lambda e^{v_\lambda}$.

Remark 4.8. If we apply Theorem 4.6 to the case when Ω is a disk, then $H(x) = 0$, and so

$$v_0(x) = -2\sum_{i=1}^M \log|x - \mathbf{x}_i^+| + 2\sum_{i=1}^M \log|x - \mathbf{x}_i^-|.$$

The equations (4.21) become

$$-\sum_{j=1,\,j\neq i}^{M}\frac{(\mathbf{x}_i^+-\mathbf{x}_j^+)\cdot\tau(\mathbf{x}_i^+)}{|\mathbf{x}_i^+-\mathbf{x}_j^+|^2}+\sum_{j=1}^{M}\frac{(\mathbf{x}_i^+-\mathbf{x}_j^-)\cdot\tau(\mathbf{x}_i^+)}{|\mathbf{x}_i^+-\mathbf{x}_j^-|^2}=0,\quad i=1,\dots,M,$$

and

$$-\sum_{j=1}^{M}\frac{(\mathbf{x}_i^--\mathbf{x}_j^+)\cdot\tau(\mathbf{x}_i^-)}{|\mathbf{x}_i^--\mathbf{x}_j^+|^2}+\sum_{j=1,\,j\neq i}^{M}\frac{(\mathbf{x}_i^--\mathbf{x}_j^-)\cdot\tau(\mathbf{x}_i^-)}{|\mathbf{x}_i^--\mathbf{x}_j^-|^2}=0,\quad i=1,\dots,M,$$

or in terms of the angles θ_j^+ and θ_j^- associated with the points $\mathbf{x}_j^+$ and $\mathbf{x}_j^-$,

$$-\sum_{j=1,\,j\neq i}^{M}\frac{\sin(\theta_i^+-\theta_j^+)}{1-\cos(\theta_i^+-\theta_j^+)}+\sum_{j=1}^{M}\frac{\sin(\theta_i^+-\theta_j^-)}{1-\cos(\theta_i^+-\theta_j^-)}=0,\quad i=1,\dots,M,$$

and

$$-\sum_{j=1}^{M}\frac{\sin(\theta_i^--\theta_j^+)}{1-\cos(\theta_i^--\theta_j^+)}+\sum_{j=1,\,j\neq i}^{M}\frac{\sin(\theta_i^--\theta_j^-)}{1-\cos(\theta_i^--\theta_j^-)}=0,\quad i=1,\dots,M.$$

It is clear that $\theta_j^+=\frac{2j\pi}{M}$, $\theta_j^-=\frac{(2j-1)\pi}{M}$, $j=1,\dots,M$, (and any fixed rotation of this set of angles) is a solution to these equations. This is consistent with the fact that all the explicit solution families constructed in [4] blow up, with alternating signs, at an even number of equidistant points.

The following lemma will be used for the proof of Theorem 4.6.

LEMMA 4.9. *Let Φ be a conformal equivalence on $\overline{\Omega}$, and let $\mathbf{x}^*$ be an arbitrary point on $\partial\Omega\setminus\{\Phi^{-1}(\infty)\}$; then*

$$\frac{\partial}{\partial\tau_x}\left(\log|\Phi(x)-\Phi(\mathbf{x}^*)|-\log|x-\mathbf{x}^*|\right)|_{x=\mathbf{x}^*}=\frac{\frac{\partial}{\partial\tau_x}|\det(D\Phi(x))|_{x=\mathbf{x}^*}}{4|\det(D\Phi(\mathbf{x}^*))|}.$$

Proof. In the following proof we use the Einstein summation convention: Repeated indices indicate summation. We immediately calculate

$$\begin{aligned}
&\frac{\partial}{\partial\tau_x}\left(\log|\Phi(x)-\Phi(\mathbf{x}^*)|-\log|x-\mathbf{x}^*|\right)|_{x=\mathbf{x}^*}\\
&\quad=\lim_{\substack{x\in\partial\Omega\\ x\to\mathbf{x}^*}}\left[\frac{(\Phi(x)-\Phi(\mathbf{x}^*))_k\partial_{x_j}\Phi_k(x)\tau_j(x)}{|\Phi(x)-\Phi(\mathbf{x}^*)|^2}-\frac{(x-\mathbf{x}^*)_j\tau_j(x)}{|x-\mathbf{x}^*|^2}\right]\\
&\quad=\lim_{\substack{x\in\partial\Omega\\ x\to\mathbf{x}^*}}\frac{(\Phi(x)-\Phi(\mathbf{x}^*))_k\partial_{x_j}\Phi_k(x)\tau_j(x)|x-\mathbf{x}^*|^2-(x-\mathbf{x}^*)_j\tau_j(x)|\Phi(x)-\Phi(\mathbf{x}^*)|^2}{|\Phi(x)-\Phi(\mathbf{x}^*)|^2|x-\mathbf{x}^*|^2}.
\end{aligned}$$

The numerator in this last expression may be expanded as follows:

$$\begin{aligned}
&(\Phi(x)-\Phi(\mathbf{x}^*))_k\partial_{x_j}\Phi_k(x)\tau_j(x)|x-\mathbf{x}^*|^2-(x-\mathbf{x}^*)_j\tau_j(x)|\Phi(x)-\Phi(\mathbf{x}^*)|^2\\
&\quad=\frac{1}{2}\partial_{x_l}\partial_{x_m}\Phi_k(\mathbf{x}^*)(x-\mathbf{x}^*)_l(x-\mathbf{x}^*)_m\partial_{x_j}\Phi_k(\mathbf{x}^*)\tau_j(\mathbf{x}^*)|x-\mathbf{x}^*|^2\\
&\qquad+\partial_{x_m}\Phi_k(\mathbf{x}^*)(x-\mathbf{x}^*)_m\partial_{x_l}\partial_{x_j}\Phi_k(\mathbf{x}^*)(x-\mathbf{x}^*)_l\tau_j(\mathbf{x}^*)|x-\mathbf{x}^*|^2\\
&\qquad-(x-\mathbf{x}^*)_j\tau_j(\mathbf{x}^*)\partial_{x_n}\Phi_k(\mathbf{x}^*)(x-\mathbf{x}^*)_n\partial_{x_l}\partial_{x_m}\Phi_k(\mathbf{x}^*)(x-\mathbf{x}^*)_l(x-\mathbf{x}^*)_m\\
&\qquad+O(|x-\mathbf{x}^*|^5)\\
&\quad=|x-\mathbf{x}^*|^4\left(\frac{1}{2}\partial_{x_l}\partial_{x_m}\Phi_k(\mathbf{x}^*)\tau_l(\mathbf{x}^*)\tau_m(\mathbf{x}^*)\partial_{x_j}\Phi_k(\mathbf{x}^*)\tau_j(\mathbf{x}^*)+O(|x-\mathbf{x}^*|)\right),
\end{aligned}\tag{4.23}$$

and the denominator may be expanded as follows:

$$|\Phi(x)-\Phi(\mathbf{x}^*)|^2|x-\mathbf{x}^*|^2 = |x-\mathbf{x}^*|^4\left[|\det(D\Phi(\mathbf{x}^*))| + O(|x-\mathbf{x}^*|)\right]. \tag{4.24}$$

Here we have on several occasions used that Φ is conformal on $\overline{\Omega}\backslash\{\Phi^{-1}(\infty)\}$. Insertion of these new expressions, (4.23) and (4.24), immediately yields

$$\begin{aligned}&\frac{\partial}{\partial\tau_x}\left(\log|\Phi(x)-\Phi(\mathbf{x}^*)| - \log|x-\mathbf{x}^*|\right)|_{x=\mathbf{x}^*}\\ &\quad= \frac{\partial_{x_l}\partial_{x_m}\Phi_k(\mathbf{x}^*)\tau_l(\mathbf{x}^*)\tau_m(\mathbf{x}^*)\partial_{x_j}\Phi_k(\mathbf{x}^*)\tau_j(\mathbf{x}^*)}{2|\det(D\Phi(\mathbf{x}^*))|}.\end{aligned} \tag{4.25}$$

Now

$$\begin{aligned}\frac{\partial}{\partial\tau_x}|\det(D\Phi(x))| &= \frac{\partial}{\partial\tau_x}\sum_{k=1}^{2}(\partial_{x_j}\Phi_k(x)\tau_j(x))^2\\ &= 2\partial_{x_j}\Phi_k(x)\tau_j(x)\partial_{x_m}\left(\partial_{x_l}\Phi_k(x)\tau_l(x)\right)\tau_m(x)\\ &= 2\partial_{x_j}\Phi_k(x)\tau_j(x)\partial_{x_m}\partial_{x_l}\Phi_k(x)\tau_l(x)\tau_m(x)\\ &\quad + 2\partial_{x_j}\Phi_k(x)\tau_j(x)\partial_{x_l}\Phi_k(x)\partial_{x_m}\tau_l(x)\tau_m(x)\\ &= 2\partial_{x_j}\Phi_k(x)\tau_j(x)\partial_{x_m}\partial_{x_l}\Phi_k(x)\tau_l(x)\tau_m(x).\end{aligned} \tag{4.26}$$

For the last identity we used that

$$\left\langle D\Phi(x)\tau(x), D\Phi(x)\frac{\partial}{\partial\tau_x}\tau(x)\right\rangle = -\kappa(x)\langle D\Phi(x)\tau(x), D\Phi(x)\mathbf{n}(x)\rangle = 0.$$

Insertion of (4.26) into (4.25) finally gives

$$\frac{\partial}{\partial\tau_x}\left(\log|\Phi(x)-\Phi(\mathbf{x}^*)| - \log|x-\mathbf{x}^*|\right)|_{x=\mathbf{x}^*} = \frac{\frac{\partial}{\partial\tau_x}|\det(D\Phi(x))|_{x=\mathbf{x}^*}}{4|\det(D\Phi(\mathbf{x}^*))|},$$

exactly as desired. ∎

Proof of Theorem 4.6. The statement about the convergence of $v^0_{\lambda_n} = v_{\lambda_n} - \int_{\partial\Omega} v_{\lambda_n}\, d\sigma/|\partial\Omega|$ follows directly from Theorem 3.1. The limit v_0 has the form

$$\begin{aligned}v_0(x) &= -\int_{\partial\Omega} N(z,x)\; d(\mu_+-\mu_-)_z = -\sum_{i=1}^{M}\alpha_i^+ N(\mathbf{x}_i^+,x) + \sum_{i=1}^{N}\alpha_i^- N(\mathbf{x}_i^-,x)\\ &= H^\alpha(x) - \sum_{i=1}^{M}\frac{\alpha_i^+}{\pi}\log|x-\mathbf{x}_i^+| + \sum_{i=1}^{N}\frac{\alpha_i^-}{\pi}\log|x-\mathbf{x}_i^-|,\end{aligned} \tag{4.27}$$

where H^α is the $C^\infty(\overline{\Omega})$ solution to

$$\Delta H^\alpha = 0 \quad \text{in } \Omega, \qquad \frac{\partial H^\alpha}{\partial\mathbf{n}} = \sum_{i=1}^{M}\frac{\alpha_i^+}{\pi}\frac{(x-\mathbf{x}_i^+)\cdot\mathbf{n}}{|x-\mathbf{x}_i^+|^2} - \sum_{i=1}^{N}\frac{\alpha_i^-}{\pi}\frac{(x-\mathbf{x}_i^-)\cdot\mathbf{n}}{|x-\mathbf{x}_i^-|^2} \quad \text{on } \partial\Omega,$$

$$\int_{\partial\Omega} H^\alpha\, d\sigma = \sum_{i=1}^{M}\frac{\alpha_i^+}{\pi}\int_{\partial\Omega}\log|x-\mathbf{x}_i^+|\, d\sigma_x - \sum_{i=1}^{N}\frac{\alpha_i^-}{\pi}\int_{\partial\Omega}\log|x-\mathbf{x}_i^-|\, d\sigma_x.$$

The form of H^α and the last identity in (4.27) are consequences of (3.1) and (3.2). We introduce a convenient renaming of the points $\mathbf{x}_i^\pm$ and the weights $\alpha_i^\pm$. Define points $\mathbf{x}_i$, $1 \le i \le M+N$, by

$$\mathbf{x}_i = \mathbf{x}_i^+, \quad 1 \le i \le M, \qquad \mathbf{x}_{i+M} = \mathbf{x}_i^-, \quad 1 \le i \le N,$$

and weights $\beta_i^\pm$, $1 \le i \le M+N$, by

$$\begin{aligned} &\beta_i^+ = \alpha_i^+ > 0, \quad 1 \le i \le M, \qquad \beta_{i+M}^+ = 0, \quad 1 \le i \le N, \\ &\beta_i^- = 0, \quad 1 \le i \le M, \qquad \beta_{i+M}^- = \alpha_i^- > 0, \quad 1 \le i \le N. \end{aligned}$$

The points $\mathbf{x}_i$ are distinct, and the product of β_i^+ and β_i^- is zero, since there is, by assumption, no overlap between $\{\mathbf{x}_i^+\}_{i=1}^M$ and $\{\mathbf{x}_i^-\}_{i=1}^N$. With this notation,

$$\mu_\pm = \sum_{i=1}^{M+N} \beta_i^\pm \delta_{\mathbf{x}_i}. \tag{4.28}$$

A part of the following argument is similar to that given in the proof of Theorem 4.1. At this point we need a refined version in order to treat asymptotically smaller terms, and for ease of comprehension we have decided to give the argument in its entirety. Let $x \to y = \Phi(x)$ be a conformal equivalence of Ω onto the upper half-plane $\mathbb{H} = \{(y_1, y_2) : y_2 > 0\}$ with the property that none of the points $\mathbf{y}_i = \Phi(\mathbf{x}_i)$, $1 \le i \le M+N$ are mapped to infinity. Setting $u_{\lambda_n} = v_{\lambda_n} \circ \Phi^{-1}$, we now have a family of solutions to

$$\begin{cases} \Delta u_{\lambda_n} = 0 & \text{in } \mathbb{H}, \\ \dfrac{\partial u_{\lambda_n}}{\partial y_2} = -\lambda_n h(y) \sinh(u_{\lambda_n}) & \text{on } \partial\mathbb{H}, \end{cases} \tag{4.29}$$

with $h(y) = |\det(D\Phi(\Phi^{-1}(y)))|^{-1/2}$. The sequence

$$u_{\lambda_n} - s_{\lambda_n}, \quad \text{with } s_{\lambda_n} = \int_{\partial\Omega} v_{\lambda_n}\, d\sigma/|\partial\Omega|,$$

converges to $u_0 = v_0 \circ \Phi^{-1}$ in $H^t(\mathbb{H} \cap \{|y| \le R\})$ for any $t < 1$ and any R; the convergence also takes place in $C^\infty(\overline{\mathbb{H}} \setminus \{\mathbf{y}_i\}_{i=1}^{M+N})$, i.e., in $C^\infty(K)$ for any compact set $K \subset \overline{\mathbb{H}} \setminus \{\mathbf{y}_i\}_{i=1}^{M+N}$. The function u_0 satisfies

$$\begin{cases} \Delta u_0 = 0 & \text{in } \mathbb{H}, \\ \dfrac{\partial u_0}{\partial y_2} = -\displaystyle\sum_{i=1}^{M+N} (\beta_i^+ - \beta_i^-)\delta_{\mathbf{y}_i} & \text{on } \partial\mathbb{H}. \end{cases} \tag{4.30}$$

We now introduce functions w_{λ_n} and w_0 by

$$w_{\lambda_n} := \partial_{y_1} u_{\lambda_n} \partial_{y_2} u_{\lambda_n} \quad \text{and} \quad w_0 := \partial_{y_1} u_0 \partial_{y_2} u_0,$$

respectively. Due to the $C^\infty(\overline{\mathbb{H}} \setminus \{\mathbf{y}_i\}_{i=1}^{M+N})$ convergence of $u_{\lambda_n} - s_{\lambda_n}$ toward u_0, the sequence w_{λ_n} converges in $C^\infty(\overline{\mathbb{H}} \setminus \{\mathbf{y}_i\}_{i=1}^{M+N})$ toward w_0. Simple calculations yield

$$\Delta w_{\lambda_n} = 0 \quad \text{in } \mathbb{H}$$

and

$$w_{\lambda_n} = -\lambda_n h(y_1,0)\sinh(u_{\lambda_n})\partial_{y_1}u_{\lambda_n} = -\lambda_n h(y_1,0)\partial_{y_1}\left(\cosh(u_{\lambda_n})\right) \quad \text{on } \partial\mathbb{H}. \tag{4.31}$$

From (4.30) and the fact that $u_0 = v_0 \circ \Phi^{-1}$, it follows that u_0 has the form

$$u_0(y) = c_0 - \sum_{i=1}^{M+N} \frac{(\beta_i^+ - \beta_i^-)}{\pi}\log|y - \mathbf{y}_i|, \tag{4.32}$$

where the constant c_0 is given by $c_0 = v_0(\Phi^{-1}(\infty))$. This is a simplified version of Lemma 4.5, corresponding to $F = 0$. We may thus calculate

$$\begin{aligned} w_0(y) &= \partial_{y_1}u_0\partial_{y_2}u_0(y) \\ &= \sum_{i,j}\frac{(\beta_j^+ - \beta_j^-)(\beta_i^+ - \beta_i^-)}{\pi^2}\frac{(y-\mathbf{y}_j)_1 y_2}{|y-\mathbf{y}_j|^2|y-\mathbf{y}_i|^2} \end{aligned}$$

or by a slight regrouping

$$\begin{aligned} w_0(y) = &\sum_i \frac{(\beta_i^+ - \beta_i^-)^2}{\pi^2}\frac{(y-\mathbf{y}_i)_1 y_2}{|y-\mathbf{y}_i|^4} \\ &+ \sum_{i,j,\, i\neq j}\frac{(\beta_j^+ - \beta_j^-)(\beta_i^+ - \beta_i^-)}{\pi^2}\frac{(y-\mathbf{y}_j)_1 y_2}{|y-\mathbf{y}_j|^2|y-\mathbf{y}_i|^2}. \end{aligned} \tag{4.33}$$

We now derive an alternate representation for the function w_0 by use of the relationship $w_0 = \lim_{\lambda_n\to 0} w_{\lambda_n}$. Let $D \subset \mathbb{H}$ be a bounded, smooth domain with $\Gamma_0 = \partial D \cap \partial\mathbb{H} = [-R,R]\times\{0\}$ (for instance, take D to be the half-disk $B_R(0)\cap\mathbb{H}$ with the two corners "smoothed out"). Choose R sufficiently large that all the points $\mathbf{y}_i^{\pm}$ lie inside $\frac{1}{2}\Gamma_0$. Let $G_D(y,z)$ denote the Green's function for the domain D, i.e., the solution to

$$\begin{cases} \Delta G_D(\cdot,z) = \delta_z & \text{in } D, \\ G_D(\cdot,z) = 0 & \text{on } \partial D. \end{cases} \tag{4.34}$$

For any fixed $z \in D$ the harmonic function w_{λ_n} may now be represented as

$$w_{\lambda_n}(z) = \int_{\partial D} w_{\lambda_n}(y)\frac{\partial G_D}{\partial \mathbf{n}_y}(y,z)\, d\sigma_y.$$

We decompose the boundary of D as follows: $\partial D = \Gamma_0 \cup \Gamma_1$, with $\Gamma_0 = \partial D \cap \partial\mathbb{H}$ and $\Gamma_1 = \partial D \cap \mathbb{H}$. In light of (4.31), the above integral representation for w_{λ_n} reads

$$\begin{aligned} w_{\lambda_n}(z) = &\int_{\Gamma_1} w_{\lambda_n}(y)\frac{\partial G_D}{\partial \mathbf{n}_y}(y,z)d\sigma_y \\ &+ \int_{\Gamma_0}\lambda_n h(y_1,0)\partial_{y_1}(\cosh(u_{\lambda_n}(y_1,0)))\frac{\partial G_D}{\partial y_2}((y_1,0),z)dy_1 \\ &= I_{1,\lambda_n}(z) + I_{2,\lambda_n}(z). \end{aligned} \tag{4.35}$$

As a consequence of the $C^\infty(\overline{\mathbb{H}}\setminus\{\mathbf{y}_i\}_{i=1}^{M+N})$ convergence of w_{λ_n} toward w_0

$$\lim_{\lambda_n\to 0} I_{1,\lambda_n}(z) = \int_{\Gamma_1} w_0(y)\frac{\partial G_D}{\partial \mathbf{n}_y}(y,z)d\sigma_y \quad \forall z \in D. \tag{4.36}$$

Integration by parts yields

$$
\begin{aligned}
I_{2,\lambda_n}(z) = &- \int_{\Gamma_0} \lambda_n \cosh(u_{\lambda_n}(y_1,0)) h(y_1,0) \partial_{y_1} \left(\frac{\partial G_D}{\partial y_2}((y_1,0),z) \right) dy_1 \\
&- \int_{\Gamma_0} \lambda_n \cosh(u_{\lambda_n}(y_1,0)) \partial_{y_1} h(y_1,0) \frac{\partial G_D}{\partial y_2}((y_1,0),z) dy_1 \\
&+ \left(\lambda_n \cosh(u_{\lambda_n}(y_1,0)) h(y_1,0) \frac{\partial G_D}{\partial y_2}((y_1,0),z) \right) \Bigg|_{y_1=-R}^{y_1=R}.
\end{aligned} \tag{4.37}
$$

From Lemma 4.3 we know that

$$
\lambda_n e^{v_{\lambda_n}} \to 2\mu_+ = 2 \sum_{i=1}^{M+N} \beta_i^+ \delta_{\mathbf{x}_i} \quad \text{and} \quad \lambda_n e^{-v_{\lambda_n}} \to 2\mu_- = 2 \sum_{i=1}^{M+N} \beta_i^- \delta_{\mathbf{x}_i}
$$

in the sense of measures on $\partial\Omega$, and so

$$
\lambda_n \cosh(v_{\lambda_n}) \to \sum_{i=1}^{M+N} (\beta_i^+ + \beta_i^-) \delta_{\mathbf{x}_i}
$$

in the sense of measures on $\partial\Omega$. The left-hand sides also converge uniformly to zero away from the points $\mathbf{x}_i$. When "pushed forward" by the conformal map Φ, the convergence implies that

$$
\lambda_n h(y_1,0) \cosh(u_{\lambda_n}(y_1,0)) \to \sum_{i=1}^{M+N} (\beta_i^+ + \beta_i^-) \delta_{\mathbf{y}_i} \tag{4.38}
$$

in the sense of measures on $\Gamma_0 = [-R,R] \times \{0\}$. As a consequence of (4.37) and (4.38) we immediately obtain the following limit for the integrals I_{2,λ_n}:

$$
\begin{aligned}
\lim_{\lambda_n \to 0} I_{2,\lambda_n}(z) = &- \sum_{i=1}^{M+N} (\beta_i^+ + \beta_i^-) \frac{\partial^2 G_D}{\partial y_1 \partial y_2}(\mathbf{y}_i, z) \\
&- \sum_{i=1}^{M+N} (\beta_i^+ + \beta_i^-) \frac{\partial_{y_1} h(\mathbf{y}_i)}{h(\mathbf{y}_i)} \frac{\partial G_D}{\partial y_2}(\mathbf{y}_i, z) \quad \forall z \in D.
\end{aligned} \tag{4.39}
$$

Here we also used the fact that $\lambda_n \cosh(u_{\lambda_n}(y_1,0)) h(y_1,0)$ converges pointwise to zero away from the points $\mathbf{y}_i$ to eliminate the boundary term in (4.37). A combination of (4.36) and (4.39) with (4.35) now yields

$$
\begin{aligned}
w_0(z) = \lim_{\lambda_n \to 0} w_{\lambda_n}(z) = &\int_{\Gamma_1} w_0(y) \frac{\partial G_D}{\partial \mathbf{n}_y}(y,z) d\sigma_y \\
&- \sum_{i=1}^{M+N} (\beta_i^+ + \beta_i^-) \frac{\partial^2 G_D}{\partial y_1 \partial y_2}(\mathbf{y}_i, z) \\
&- \sum_{i=1}^{M+N} (\beta_i^+ + \beta_i^-) \frac{\partial_{y_1} h(\mathbf{y}_i)}{h(\mathbf{y}_i)} \frac{\partial G_D}{\partial y_2}(\mathbf{y}_i, z) \quad \forall z \in D.
\end{aligned} \tag{4.40}
$$

Set $\omega = \{z \in D : |z_1| < R/2 \text{ and } 0 < z_2 < \epsilon\}$ for ϵ fixed but sufficiently small. The Green's function $G(y,z)$, $(y,z) \in \overline{D \times \omega} \setminus \{y = z\}$, may now be written

$$
G_D(y,z) = \frac{1}{2\pi} \log|y - z| - \frac{1}{2\pi} \log|y - \overline{z}| + g(y,z),
$$

where $\overline{z} = \overline{(z_1, z_2)} = (z_1, -z_2)$ and where the function g is in $C^\infty(\overline{D \times \omega})$. Since $\mathbf{y}_i \in \Gamma_0 = \partial D \cap \partial \mathbb{H}$, $1 \le i \le M+N$, we thus compute

$$\begin{aligned} \frac{\partial G_D}{\partial y_2}(\mathbf{y}_i, z) &= -\frac{1}{\pi}\frac{z_2}{|\mathbf{y}_i - z|^2} + \partial_{y_2} g(\mathbf{y}_i, z) \quad z \in \overline{\omega} \setminus \{\mathbf{y}_i\} \quad \text{and} \\ \frac{\partial^2 G_D}{\partial y_1 \partial y_2}(\mathbf{y}_i, z) &= \frac{2}{\pi}\frac{(\mathbf{y}_i - z)_1 z_2}{|\mathbf{y}_i - z|^4} + \partial_{y_1}\partial_{y_2} g(\mathbf{y}_i, z) \quad z \in \overline{\omega} \setminus \{\mathbf{y}_i\}. \end{aligned} \tag{4.41}$$

Substituting (4.41) into (4.40) we arrive at

$$\begin{aligned} w_0(z) = \frac{2}{\pi}\sum_{i=1}^{M+N} (\beta_i^+ + \beta_i^-)\frac{(z - \mathbf{y}_i)_1 z_2}{|z - \mathbf{y}_i|^4} \\ + \frac{1}{\pi}\sum_{i=1}^{M+N} (\beta_i^+ + \beta_i^-)\frac{\partial_{y_1} h(\mathbf{y}_i)}{h(\mathbf{y}_i)}\frac{z_2}{|z - \mathbf{y}_i|^2} + R(z), \end{aligned} \tag{4.42}$$

where R is in $C^\infty(\overline{\omega})$. By comparing the two representations (4.33) and (4.42) for w_0, and using the fact that the singular terms of same type (near $\mathbf{y}_i$) must coincide, we now obtain equations for the weights $\{\beta_i^\pm\}_{i=1}^{M+N}$ and the points $\{\mathbf{y}_i\}_{i=1}^{M+N}$. From the terms of type $(\cdot - \mathbf{y}_i)_1(\cdot)_2/|\cdot - \mathbf{y}_i|^4$,

$$\frac{(\beta_i^+ - \beta_i^-)^2}{\pi^2} = \frac{2}{\pi}(\beta_i^+ + \beta_i^-), \quad 1 \le i \le N + M. \tag{4.43}$$

From the terms of type $(\cdot)_2/|\cdot - \mathbf{y}_i|^2$,

$$\begin{aligned} \frac{(\beta_i^+ - \beta_i^-)}{\pi^2}\sum_{j=1,\, j\neq i}^{M+N} (\beta_j^+ - \beta_j^-)\frac{(\mathbf{y}_i - \mathbf{y}_j)_1}{|\mathbf{y}_i - \mathbf{y}_j|^2} \\ = \frac{1}{\pi}(\beta_i^+ + \beta_i^-)\frac{\partial_{y_1} h(\mathbf{y}_i)}{h(\mathbf{y}_i)}, \quad 1 \le i \le N + M. \end{aligned} \tag{4.44}$$

Recall that, by assumption, the two set of points $\{\mathbf{x}_i^+\}_{i=1}^M$ and $\{\mathbf{x}_i^-\}_{i=1}^N$ are disjoint, and $0 < \beta_i^+ = \alpha_i^+$, $1 \le i \le M$, $0 < \beta_{i+M}^- = \alpha_i^-$, $1 \le i \le N$, with $\beta_i^\pm = 0$ otherwise. In terms of the points $\mathbf{y}_i^\pm = \Phi(\mathbf{x}_i^\pm)$ and the weights $\alpha_i^\pm$, the identities (4.43) and (4.44) now reduce to

$$\alpha_i^+ = 2\pi, \quad 1 \le i \le M, \qquad \alpha_i^- = 2\pi, \quad 1 \le i \le N,$$

$$2\sum_{j=1,\, j\neq i}^{M} \frac{(\mathbf{y}_i^+ - \mathbf{y}_j^+)_1}{|\mathbf{y}_i^+ - \mathbf{y}_j^+|^2} - 2\sum_{j=1}^{N} \frac{(\mathbf{y}_i^+ - \mathbf{y}_j^-)_1}{|\mathbf{y}_i^+ - \mathbf{y}_j^-|^2} = \frac{\partial_{y_1} h(\mathbf{y}_i^+)}{h(\mathbf{y}_i^+)}, \quad 1 \le i \le M, \quad \text{and} \tag{4.45}$$

$$-2\sum_{j=1}^{M} \frac{(\mathbf{y}_i^- - \mathbf{y}_j^+)_1}{|\mathbf{y}_i^- - \mathbf{y}_j^+|^2} + 2\sum_{j=1,\, j\neq i}^{N} \frac{(\mathbf{y}_i^- - \mathbf{y}_j^-)_1}{|\mathbf{y}_i^- - \mathbf{y}_j^-|^2} = \frac{\partial_{y_1} h(\mathbf{y}_i^-)}{h(\mathbf{y}_i^-)}, \quad 1 \le i \le N. \tag{4.46}$$

Since all the $\alpha_i^\pm$ have the same value (2π) and since $\sum_{i=1}^M \alpha_i^+ = \mu_+(\partial\Omega) = \mu_-(\partial\Omega) = \sum_{i=1}^N \alpha_i^-$, it follows that $M = N (\ge 1)$. From the definition of h we calculate

$$\frac{\partial h}{\partial y_1}(\mathbf{y}_i^\pm) = -\frac{1}{2}|\det(D\Phi(\mathbf{x}_i^\pm)|^{-2}\frac{\partial}{\partial \tau_x}|\det(D\Phi(x))|_{x=\mathbf{x}_i^\pm},$$

so that

$$\frac{\partial_{y_1} h(\mathbf{y}_i^{\pm})}{h(\mathbf{y}_i^{\pm})^2} = -\frac{1}{2}\frac{\frac{\partial}{\partial\tau_x}\left|\det(D\Phi(x))\right|_{x=\mathbf{x}_i^{\pm}}}{|\det(D\Phi(\mathbf{x}_i^{\pm})|}. \tag{4.47}$$

We have that

$$\begin{aligned} v_0(x) + 2\log|x-\mathbf{x}_i^+| &= v_0(x) + 2\log|\Phi(x)-\Phi(\mathbf{x}_i^+)| \\ &\quad + 2(\log|x-\mathbf{x}_i^+| - \log|\Phi(x)-\Phi(\mathbf{x}_i^+)|) \\ &= [u_0(y) + 2\log|y-\mathbf{y}_i^+|]_{y=\Phi(x)} \\ &\quad + 2(\log|x-\mathbf{x}_i^+| - \log|\Phi(x)-\Phi(\mathbf{x}_i^+)|) \\ &= \left[c_0 - 2\sum_{j=1,\, j\neq i}^{M}\log|y-\mathbf{y}_j^+| + 2\sum_{j=1}^{M}\log|y-\mathbf{y}_j^-|\right]_{y=\Phi(x)} \\ &\quad + 2(\log|x-\mathbf{x}_i^+| - \log|\Phi(x)-\Phi(\mathbf{x}_i^+)|). \end{aligned}$$

For the last identity we used the representation formula (4.32) for u_0 and the facts that $M = N$ and $\beta_i^+ - \beta_i^- = 2\pi$, $1 \le i \le M$, $\beta_i^+ - \beta_i^- = -2\pi$, $M+1 \le i \le 2M$. By differentiation and use of Lemma 4.9 we now get that

$$\begin{aligned} &\frac{\partial}{\partial\tau_x}(v_0(x) + 2\log|x-\mathbf{x}_i^+|)_{x=\mathbf{x}_i^+} \\ &\quad = \frac{1}{h(\mathbf{y}_i^+)}\frac{\partial}{\partial y_1}\left[c_0 - 2\sum_{j=1,\, j\neq i}^{M}\log|y-\mathbf{y}_j^+| + 2\sum_{j=1}^{M}\log|y-\mathbf{y}_j^-|\right]_{y=\mathbf{y}_i^+} \\ &\qquad -\frac{1}{2}\frac{\frac{\partial}{\partial\tau_x}|\det(D\Phi(x))|_{x=\mathbf{x}_i^+}}{|\det(D\Phi(\mathbf{x}_i^+))|}. \end{aligned}$$

In combination with (4.45) and (4.47) this immediately yields

$$\frac{\partial}{\partial\tau_x}(v_0(x) + 2\log|x-\mathbf{x}_i^+|)_{x=\mathbf{x}_i^+} = 0, \quad 1 \le i \le M.$$

A similar approach, based on (4.46), leads to the identities

$$\frac{\partial}{\partial\tau_x}(v_0(x) - 2\log|x-\mathbf{x}_i^-|)_{x=\mathbf{x}_i^-} = 0, \quad 1 \le i \le M.$$

Finally

$$\lim_{\lambda_n\to 0}\lambda_n\int_{\partial\Omega}|\sinh(v_{\lambda_n})|\,d\sigma = (\mu_+ + \mu_-)(\partial\Omega) = \sum_{i=1}^{M}\alpha_i^+ + \sum_{i=1}^{M}\alpha_i^- = 4M\pi.$$

This completes the proof of Theorem 4.6. □

Remark 4.10. If we drop the assumption that $\{\mathbf{x}_i^+\} \cap \{\mathbf{x}_i^-\} = \emptyset$, then the same techniques that were used for the proof of Theorem 4.6 still lead to some interesting conclusions. Let $\mathbf{x}^*$ be any point in $\{\mathbf{x}_i^+\} \cup \{\mathbf{x}_i^-\} = S \subset \partial\Omega$ and suppose μ_+ and μ_- have contributions $\alpha_*^+\delta_{\mathbf{x}^*}$ and $\alpha_*^-\delta_{\mathbf{x}^*}$, respectively (included here is the possibility

that α_*^+ or α_*^- is zero). From (4.43) in the proof of Theorem 4.6 we immediately obtain the identity

$$\frac{(\alpha_*^+ - \alpha_*^-)^2}{\pi^2} = \frac{2}{\pi}(\alpha_*^+ + \alpha_*^-).$$

For points that are in only one of the sets $\{\mathbf{x}_i^+\}$ or $\{\mathbf{x}_i^-\}$ (but not in $\{\mathbf{x}_i^+\} \cap \{\mathbf{x}_i^-\}$) this still asserts that the nonzero weight (α_*^+ or α_*^-) is 2π, but for potential $\mathbf{x}_*$ in $\{\mathbf{x}_i^+\} \cap \{\mathbf{x}_i^-\}$ this identity implies only that

$$\alpha_*^+ + \alpha_*^- > |\alpha_*^+ - \alpha_*^-| > 2\pi,$$

so that the contribution to $\frac{\partial v_0}{\partial \mathbf{n}} = \mu_+ - \mu_-$ and to $\mu_+ + \mu_-$ is of greater magnitude than $2\pi\delta_{\mathbf{x}^*}$. This result was already derived in the proof of Theorem 4.1 in the most general setting, without any assumptions about $d_\pm$.

The equation analogous to (4.21) becomes

$$\frac{\partial}{\partial \tau_x}\left(v_0(x) + \frac{(\alpha_*^+ - \alpha_*^-)}{\pi}\log|x - \mathbf{x}^*|\right)\Big|_{x=\mathbf{x}^*} = 0,$$

or in terms of the full gradient

$$\begin{aligned}\nabla_x\left(v_0(x) + \frac{(\alpha_*^+ - \alpha_*^-)}{\pi}\log|x - \mathbf{x}^*|\right)\Big|_{x=\mathbf{x}^*} &= \frac{(\alpha_*^+ - \alpha_*^-)}{2\pi}\kappa(\mathbf{x}^*)\mathbf{n}(\mathbf{x}^*) \\ &= \frac{(\alpha_*^+ + \alpha_*^-)}{(\alpha_*^+ - \alpha_*^-)}\kappa(\mathbf{x}^*)\mathbf{n}(\mathbf{x}^*).\end{aligned}$$

For points that are in only one of the sets $\{\mathbf{x}_i^+\}$ or $\{\mathbf{x}_i^-\}$ (but not in $\{\mathbf{x}_i^+\} \cap \{\mathbf{x}_i^-\}$) these are exactly the same equations as in Theorem 4.6. We are not at the moment able to derive equations analogous to (4.21) (or (4.22)) for the case when either d_+ or d_- is nonzero.

5. Computational examples. In this section we present some results of numerical calculations of solutions to the boundary value problem (1.1) on two different simply connected domains. As $\lambda \to 0^+$ the solutions we calculate all appear to have boundary fluxes that are bounded in L^1, i.e., they appear to satisfy $\|\lambda \sinh(v_\lambda)\|_{L^1(\partial\Omega)} \leq C$, and as predicted by Theorems 3.1 and 4.1 the limiting fluxes thus all contain at least one nonzero point mass. In section 2 we showed that such solutions also necessarily satisfy

$$|s_\lambda| = \left|\frac{1}{|\partial\Omega|}\int_{\partial\Omega} v_\lambda \, d\sigma\right| \leq \log\frac{1}{\lambda} + D.$$

The second of our numerical experiments gives strong evidence that asymptotic equality is indeed attained for certain solutions. This again implies that one of the expressions,

$$\lambda e^{\pm s_\lambda} = \lambda \exp\left(\pm\frac{1}{|\partial\Omega|}\int_{\partial\Omega} v_\lambda \, d\sigma\right),$$

has a nonzero limit point as $\lambda \to 0^+$. As seen in Theorem 3.1 (and Theorem 4.1), such a nonzero limit point gives rise to a subsequence of solutions with a limiting

boundary flux that contains point masses as well as a regular part. It is therefore not surprising that some of the limiting boundary fluxes arising in the second experiment show strong evidence of the presence of a nonzero regular part. Due to the emergence of nonzero regular parts in some of the limiting boundary fluxes (and the corresponding elimination of some potential point masses) our second experiment also provides potential evidence that the identities (4.21) (or (4.22)), derived in Theorem 4.6 for the locations of point masses, are necessary but not sufficient.

For the purpose of our computations we reformulate the problem (1.1) as a nonlinear boundary integral equation, which we then approximate using a Nyström collocation method and finally solve by means of a Newton iteration scheme [1], [8]. In order to achieve high numerical accuracy in the presence of the points of "blow-up," we first calculate the potential "blow-up" locations (by the approximate solution of the equations from Theorem 4.6) and then use a mesh for the Nyström collocation that is refined appropriately near these points. For more details about these and other Matlab computations, see [9]. Nyström's method is known to achieve exceptional (exponential) accuracy for "uniform" meshes and smooth solutions. Since our solutions develop very strong singularities as $\lambda \to 0^+$, the accuracy using a "uniform" mesh is completely unsatisfactory for small λ. We believe our mesh-refinement strategy provides much higher (though, far from exponential) accuracy. However, we should point out that, due to the fact that even the L^∞ norm of the solution blows up, we do not have any good a priori estimate of the accuracy associated with a particular mesh, valid uniformly in λ. Part of the challenge (short of rigorously developing a posteriori error estimators) is to find decent ways to gauge the accuracy of the computations.

5.1. Pure sums of point masses. The domains we consider are conformal images of the unit disk by the (complex-valued) exponential map, $\Phi(z) = e^{\gamma z}$ with $0 < \gamma < \pi$. For small values of γ, the image of this map is very close to a disk, so we expect the solutions to behave like the solutions on a disk. Indeed, this is the case when we choose $\gamma = 0.5$. We compute the two "first" nontrivial solution families, i.e., the solution families branching off at the first two nonzero Steklov eigenvalues of the linearized problem $\Delta w = 0$ in Ω, $\frac{\partial w}{\partial n} = \lambda w$ on $\partial\Omega$. Since these two eigenvalues represent the "split" of the first nonzero (double) eigenvalue for a disk, we expect to see two distinct families of solutions whose boundary fluxes each develop two point masses of opposite strength. Graphical output from our computations is shown in Figure 1. The first two frames show the normal fluxes plotted against the arclength for the two families of solutions. Arclength = 0 corresponds to the rightmost intersection of the domain with the x-axis (the point $(e^{0.5}, 0)$). In each frame we plot the boundary fluxes $\frac{\partial u_\lambda}{\partial n}$, for a collection of λ ranging between 0.24 and 10^{-3}. We clearly see two point masses develop as λ decreases. There are no nonzero regular parts in the limiting fluxes. The two families have different locations for the point masses. In the third frame we show the mesh grading function used for the computation of the first family of solutions. The horizontal axis represents the node number (from 1 to 512) and the vertical axis represents (the arclength to) the corresponding node location on $\partial\Omega$. There are "exponential-type" refinements near the two potential point mass locations, the kind of refinements that appear well suited to the expected logarithmic near-singularities of the solutions. For more details, see [9]. The fourth frame shows the two potential pairs of point mass locations as computed from the necessary conditions of Theorem 4.6 (with $M = 1$). Each potential pair is in total agreement with the approximate point mass locations observed in exactly one of the first two frames. The locations marked by $\circ$ correspond to the first family of solutions,

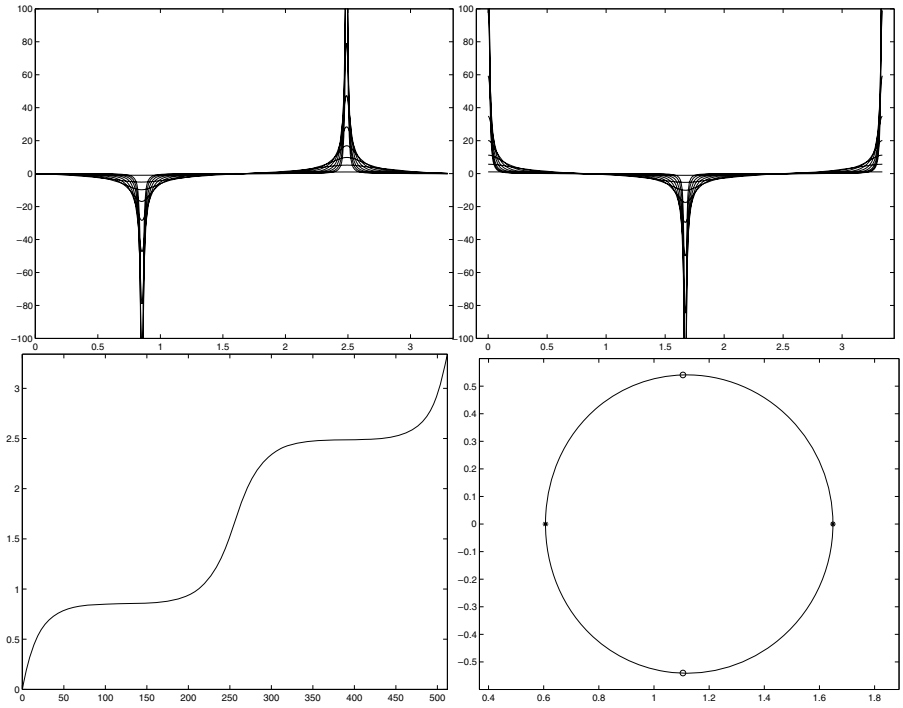

FIG. 1. *The case $\gamma = 0.5$. The two top frames show boundary fluxes for the two first solution families as a function of arclength. The third frame shows the mesh grading function used for the computation of the first solution family. The fourth frame shows the domain with marked point mass locations: $\circ$'s correspond to the first family, *'s to the second.*

the locations marked by * to the second family.

5.2. A regular part. At the other end of the spectrum, for clarity, we look at the case of $\gamma = 2.0$. Figure 2 shows a collection of boundary fluxes of solutions, for λ as small as 10^{-3}, using a mesh that is "exponentially" refined near the two potential "blow-up" points. On this domain (the shape of which is shown in the insert of Figure 2) we see very strong evidence of a positive regular part and a single negative point mass in the limiting boundary flux.

For some of the first uniform (and coarser) meshes we used in our computations, we initially saw behavior like the one showed in Figure 2, but at a certain point in λ, depending on the mesh, the positive part of the boundary flux would accelerate its growth as λ approached 0. In all the meshes we tried it has always been clear that the negative part of the boundary flux converges to a single point mass. A vast difference in scale made it plausible that the negative part was the only point mass to develop in the limit and that the positive part would converge to some smooth function. However, the accelerated growth in the positive part of the flux could indicate some additional point masses. In order to effectively rule this out we need some simple a posteriori tests to indicate, for a certain mesh, what values of λ are simply too small to allow accurate computational results.

To illustrate this point we choose $\lambda = 10^{-2}$ and $\lambda = 10^{-3}$ and compute the approximate boundary fluxes for the same three meshes. The first frame in Figure 3

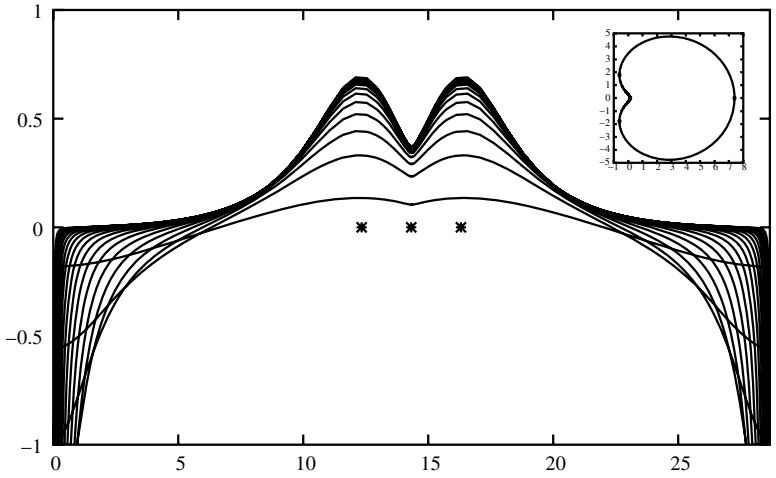

FIG. 2. *An example of a limiting boundary flux with a nonzero regular part. The domain* Ω *(insert) corresponds to* $\gamma = 2$*. The three locations indicated by *'s (below the graphs) correspond to the three marked points on the leftmost part of the boundary of* Ω*.*

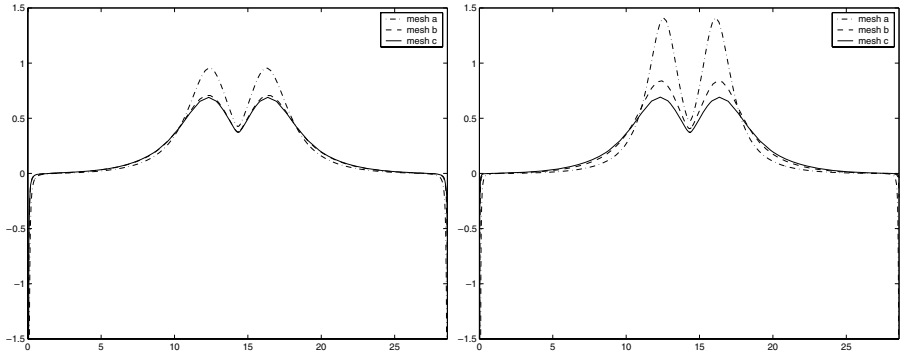

FIG. 3. *Left frame: Boundary fluxes for* $\lambda = 10^{-2}$ *computed using three different meshes. Right frame: Boundary fluxes for* $\lambda = 10^{-3}$ *computed using the same three meshes. The domain in both cases corresponds to* $\gamma = 2$*.*

displays the computed fluxes for $\lambda = 10^{-2}$; the second frame, those for $\lambda = 10^{-3}$. The three meshes we use are "exponentially" refined at the "blow-up" points. They differ by the size of the smallest mesh width near "blow-up" points. Mesh **a** is the coarsest near these points, and mesh **c** is the finest. Using the exact same three meshes, we have computed the integrals of the positive part of the approximate boundary flux and the boundary averages of the approximate solution. These results are displayed in Figure 4. Since it is quite clear (from all three meshes) that a single, isolated negative point mass develops, Theorem 4.1 asserts that this must have a mass of -2π. Correspondingly, the integral of $(\frac{\partial v_\lambda}{\partial n})^+$ should approach 2π. The left frame in Figure 4 clearly indicates that meshes **a** and **b** lack sufficient accuracy for $\lambda = 10^{-3}$, whereas mesh **c** achieves an integral very close to 2π even for $\lambda = 10^{-3}$. According to the same test mesh **b** seems adequate for $\lambda = 10^{-2}$, but mesh **a** already lacks sufficient

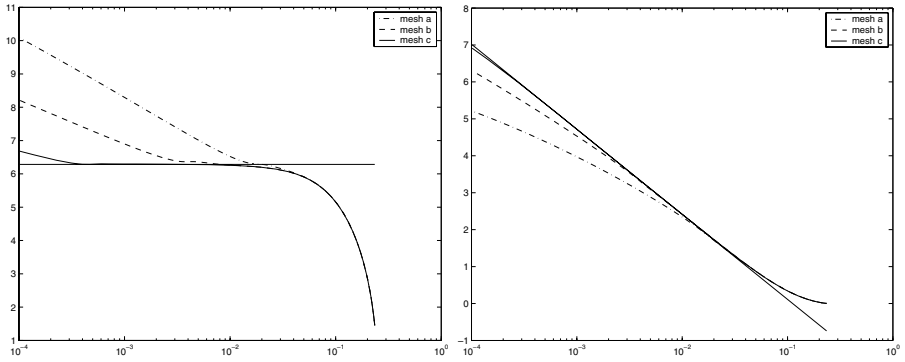

FIG. 4. *Two gauges of accuracy, as explained in text. Left frame: The integrals of the positive part of the approximate boundary flux. Right frame: The boundary averages of the approximate solution. The three meshes are the same as in Figure* 3.

accuracy for this value of λ. Comparing with Figure 3 we are thus inclined to believe the results of meshes **b** and **c** for $\lambda = 10^{-2}$ and those of mesh **c** for $\lambda = 10^{-3}$. These results predict the presence of a nonzero regular part in the limiting boundary flux. The growing "modes" we see in the other results can all be attributed to numerical inaccuracy. The right frame in Figure 4 gives a positive confirmation of the accuracy associated with mesh **c** all the way down to $\lambda = 10^{-3}$ (and the accuracy associated with mesh **b** down to 10^{-2}). In this frame we plot the boundary averages of the approximate solution versus λ. For comparison we also plot the line corresponding to the function $\log(\frac{1}{\lambda}) + D$. D is chosen by fitting the results from the "finest" mesh (mesh **c**) to this logarithmic line. From Theorem 3.1 we know that a nonzero regular part of the limiting boundary flux exists if and only if the boundary averages of the solution behave like $\pm\log(\frac{1}{\lambda}) + D$. The computations corresponding to mesh **c** shown in Figure 3 and in the right frame of Figure 4 are completely consistent with this equivalence, as are the computations corresponding to mesh **b** shown in the left frame of Figure 3 and in the right frame of Figure 4.

Based on these additional computations we believe that the limiting boundary flux for this second family of solutions, in case $\gamma = 2$, does indeed exhibit a nonzero regular part as shown in Figure 2. The equations derived in Theorem 4.6 for the potential locations of two point masses ($M = 1$) continue to have as a solution the pair of points lying at the intersection of Ω and the horizontal coordinate axis (as these equations did in the case of $\gamma = 0.5$). The fact that none of the first two computed solution families (or for that matter none of the first four solution families, as seen in the next figure) develops a pair of singularities at these two points gives some indication that our "location conditions" are necessary but not sufficient.

So far we have considered only the first two nontrivial solution families, but we have in many cases computed solution families branching off at much "higher" Steklov eigenvalues. Figure 5 shows the results of such a computation on the domain corresponding to $\gamma = 2$. The first frame shows the bifurcation diagram for the first six nontrivial solution families. We plot $\|\nabla v_\lambda\|_{L^2(\Omega)}$ versus λ. In the following four frames we show a sequence of eight boundary fluxes corresponding to each of the first four solution branches. The number m is a counter for the Steklov eigenvalues; $m = 1$ corresponds to the eigenvalue 0, $m = 2$ is the first nonzero eigenvalue, etc. The

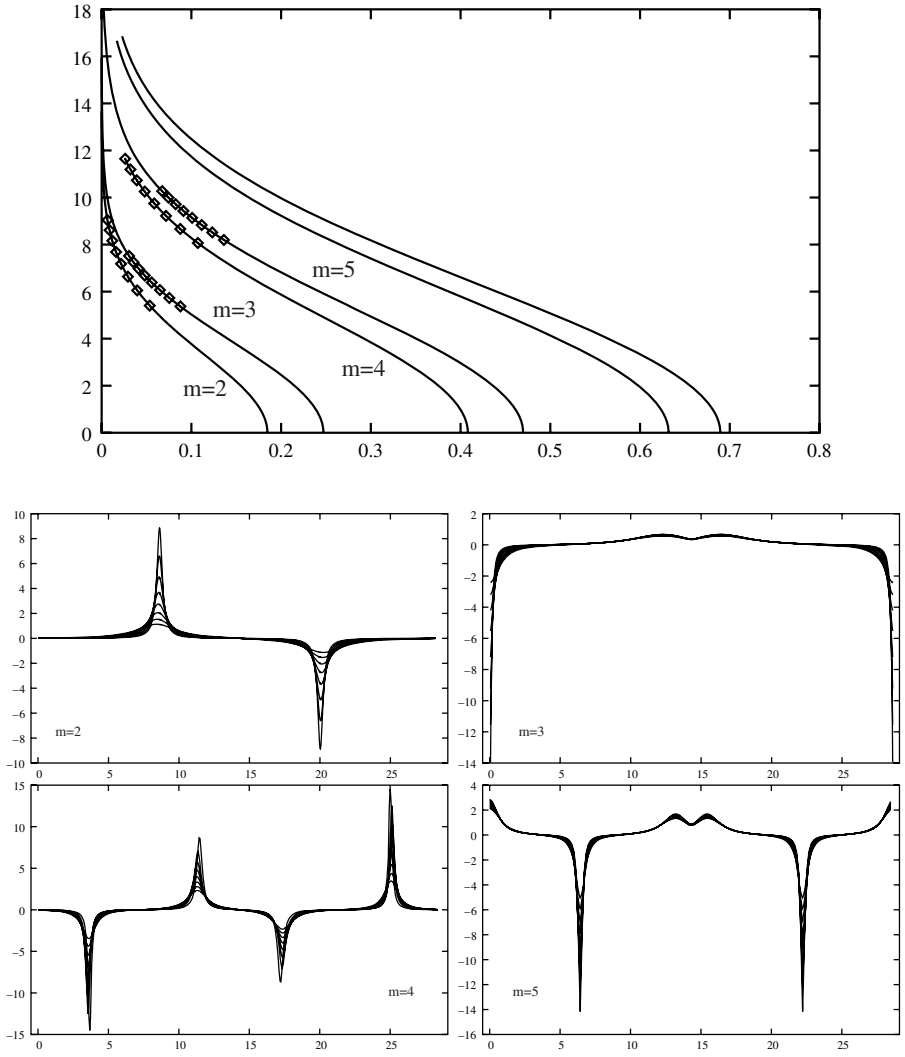

FIG. 5. *A bifurcation diagram showing* $\|\nabla v_\lambda\|_{L^2(\Omega)}$ *as a function of* λ*, and the boundary fluxes at selected points on the first four solution branches. The domain* Ω *corresponds to* $\gamma = 2$.

points in the bifurcation diagram that correspond to the shown boundary fluxes have been marked with squares. As seen in Figure 5, the boundary fluxes corresponding to $m = 2$ and $m = 4$ appear to converge to pure sums of alternating point masses (the computations also confirm that each point mass is of strength $\pm 2\pi$), whereas the limiting boundary fluxes for $m = 3$ and $m = 5$ appear to contain a nonzero regular part as well. The situation $m = 3$ has already been discussed. For $m = 5$ the limiting boundary flux appears to have two negative point masses (each of mass -2π) balanced by a positive regular part, in complete agreement with Theorem 4.1.

Acknowledgments. Part of this work was carried out during a stay at Université de Versailles. The authors thank O. Kavian and J.-P. Puel for their hospitality and

for many interesting discussions.

REFERENCES

[1] K.E. Atkinson, *The Numerical Solution of Integral Equations of the Second Kind*, Cambridge University Press, Cambridge, UK, 1997.

[2] J. Bebernes and D. Eberly, *Mathematical Problems from Combustion Theory*, Appl. Math. Sci. 83, Springer-Verlag, Berlin, 1989.

[3] H. Brezis and F. Merle, *Uniform estimates and blow-up behavior for solutions of* $-\Delta u = V(x)e^u$ *in two dimensions*, Comm. Partial Differential Equations, 16 (1991), pp. 1223–1253.

[4] K. Bryan and M. Vogelius, *Singular solutions to a nonlinear elliptic boundary value problem originating from corrosion modeling*, Quart. Appl. Math., 60 (2002), pp. 675–694.

[5] J. Deconinck, *Current Distributions and Electrode Shape Changes in Electrochemical Systems*, Lecture Notes in Engineering 75, Springer-Verlag, Berlin, 1992.

[6] O. Kavian and M. Vogelius, *On the existence and "blow-up" of solutions to a two-dimensional nonlinear boundary-value problem arising in corrosion modelling*, Proc. Roy. Soc. Edinburgh Sect. A, 133 (2003), pp. 119–149.

[7] O. Kavian and M. Vogelius, *Corrigendum: On the existence and "blow-up" of solutions to a two-dimensional nonlinear boundary-value problem arising in corrosion modelling*, Proc. Roy. Soc. Edinburgh Sect. A, 133 (2003), pp. 729–730.

[8] R. Kress, *Linear Integral Equations*, Applied Mathematical Sciences 82, Springer-Verlag, Berlin, 1989.

[9] K. Medville, *Existence and Blow-up Behavior of Planar Harmonic Functions Satisfying Certain Nonlinear Neumann Boundary Conditions*, Ph.D. thesis, Rutgers University, New Brunswick, NJ, 2004.

[10] J.L. Moseley, *Asymptotic solutions for a Dirichlet problem with an exponential nonlinearity*, SIAM J. Math. Anal., 14 (1983), pp. 719–735.

[11] K. Nagasaki and T. Suzuki, *Asymptotic analysis for two-dimensional elliptic eigenvalue problems with exponentially dominated nonlinearities*, Asymptot. Anal., 3 (1990), pp. 173–188.

[12] M. Vogelius and J.-M. Xu, *A nonlinear elliptic boundary value problem related to corrosion modeling*, Quart. Appl. Math., 56 (1998), pp. 479–505.

[13] V.H. Weston, *On the asymptotic solution of a partial differential equation with an exponential nonlinearity*, SIAM J. Math. Anal., 9 (1978), pp. 1030–1053.

SIAM J. MATH. ANAL.
Vol. 36, No. 6, pp. 1807–1835

FROM KINETIC EQUATIONS TO MULTIDIMENSIONAL ISENTROPIC GAS DYNAMICS BEFORE SHOCKS*

F. BERTHELIN† AND A. VASSEUR†

Abstract. This article is devoted to the proof of the hydrodynamical limit from kinetic equations (including BGK-like equations) to multidimensional isentropic gas dynamics. It is based on a relative entropy method; hence the derivation is valid only before shocks appear on the limit system solution. However, no a priori knowledge on high velocity distributions for kinetic functions is needed. The case of the Saint–Venant system with topography (where a source term is added) is included.

Key words. hydrodynamic limit, entropy method, BGK equation, isentropic gas dynamics, Saint–Venant system

AMS subject classifications. 35L65, 82C40, 76N, 35F20

DOI. 10.1137/S0036141003431554

1. Introduction.

1.1. Context and results. This article is devoted to the study of the hydrodynamical limit of kinetic equations to the multidimensional system of isentropic gas dynamics:

$$\begin{cases} \partial_t \rho + \operatorname{div}_x(\rho u) = 0, \quad t \in \mathbb{R}^+,\ x \in \mathbb{R}^n, \\ \partial_t(\rho u) + \operatorname{div}_x(\rho u \otimes u + I\rho^\gamma) = \rho F, \quad t \in \mathbb{R}^+,\ x \in \mathbb{R}^n, \end{cases} \tag{1.1}$$

for $1 \le \gamma \le \frac{n+2}{n}$ and a given external force field F.

This is a simplified situation of the long-term problem concerning the compressible limit of the Boltzmann equation. In this case, the hydrodynamical limit has been performed by Caflisch [9] only for smooth data during a small time. The asymptotic limit of the Boltzmann equation in low Mach number towards incompressible Euler (or Navier–Stokes) systems has been achieved recently by Saint-Raymond [26] and Lions and Masmoudi [22] following the pioneering work of Bardos, Golse, and Levermore [1]. As for our work, they are still local time results, since it is valid in the lapse of time in which the limit solution remains smooth. However, at the kinetic level, no strong smoothness property is needed. Notice that in our case we deal with compressible gases, and even the existence of a solution to (1.1) after shocks appear is not known in the multidimensional situation.

At the kinetic level, we consider a Fokker–Planck equation for the isothermal case ($\gamma = 1$) and a BGK-like equation for the other values of γ. Originally, BGK equations have been introduced by Bathnagar, Gross, and Krook as a simplification of the Boltzmann equation. This model has been extended in order to construct kinetic equations associated with different hydrodynamical systems (see the book of Perthame [24] for a survey of this field). In our particular case, the BGK model we use has been introduced for the full range of γ by Bouchut [5]. Our main result is the following.

*Received by the editors July 15, 2003; accepted for publication (in revised form) May 21, 2004; published electronically June 14, 2005.
http://www.siam.org/journals/sima/36-6/43155.html
†Laboratoire J. A. Dieudonné, UMR 6621 CNRS, Université de Nice, Parc Valrose, 06108 Nice Cedex 2, France (Florent.Berthelin@unice.fr, Alexis.Vasseur@unice.fr).

THEOREM 1.1. *Let F be in $C^2(\mathbb{R}^n) \cap L^\infty(\mathbb{R}^n)$. Let $(\rho^0, \rho^0 u^0) \in L^1(\mathbb{R}^n)$ be the initial data of a solution $U = (\rho, \rho u) \in C^1([0,T[\times\mathbb{R}^n) \cap L^1([0,T[\times\mathbb{R}^n)$ to (1.1) such that $\rho > 0$; ρ, u, $\partial_x \rho$, $\partial_x u$ are bounded with respect to (t,x); and ρu^2, $h(\rho)$ are integrable with respect to (t,x), where $h(\rho) = \rho^\gamma/(\gamma-1)$ for $\gamma > 1$ and $h(\rho) = \rho \ln \rho$ for $\gamma = 1$. Consider a family of kinetic initial values f_ε^0 verifying $f_\varepsilon^0 \in L^1(\mathbb{R}^{2n})$, $H(f_\varepsilon^0, v) \in L^1(\mathbb{R}^{2n})$ (where H is the kinetic entropy associated with the system (1.1); see section 4). We assume that it verifies*

$$\int_{\mathbb{R}^n} (f_\varepsilon^0, v f_\varepsilon^0, H(f_\varepsilon^0, v))\, dv \underset{\varepsilon\to 0}{\longrightarrow} (\rho^0, \rho^0 u^0, \rho^0 (u^0)^2/2 + h(\rho^0)) \quad \text{in } L^1(\mathbb{R}^n).$$

Let f_ε be the solution to the BGK equation (4.1) for $1 < \gamma \le n/(n+2)$ or the solution to the Fokker–Planck equation (4.5) for the isotherm case ($\gamma = 1$). We denote

$$(\rho_\varepsilon, \rho_\varepsilon u_\varepsilon) = \left(\int_{\mathbb{R}^n} f_\varepsilon\, dv, \int_{\mathbb{R}^n} v f_\varepsilon\, dv \right).$$

Then ρ_ε converges strongly in $C^0(0,T; L^p_{\mathrm{loc}}(\mathbb{R}^n))$ to ρ for every $1 \le p < \gamma$ and $\rho_\varepsilon u_\varepsilon$ converges strongly to ρu in $C^0(0,T; L^q_{\mathrm{loc}}(\mathbb{R}^n))$ for every $1 \le q < 2\gamma/(\gamma+1)$.

In the monodimensional case ($n = 1$) a stronger result has been achieved by Berthelin and Bouchut [3] in the similar situation where we have only one entropy. This result is valid even when shocks appear. The simpler case dealing with the complete family of entropies has been performed by Berthelin and Bouchut [2] (see also Serre [27] for regular systems). However, notice that in our case, no a priori assumption on the support of f_ε in v is needed. Everything is controlled by the energy bound. (We also refer the reader to [28], [17] for the convergence of discrete kinetic models to the Lagrangian version of the p-system in the one-dimensional case but even after the appearance of shocks.)

The main tool is a relative entropy method. It relies on the "weak-strong" uniqueness principle, established by Dafermos for multidimensional systems of hyperbolic conservation laws admitting a convex entropy functional [10]. It is close to the concept of dissipative solutions for the Euler equations of Lions [21]. It has been frequently used for systems of particles and rarefied gas dynamics; see Yau [29] and Golse, Levermore, and Saint-Raymond [14] (see also Goudon, Jabin, and Vasseur [18]). For different asymptotic problems it is called the "modulated energy" method (Brenier [7], Masmoudi [23], and Brenier [8]).

1.2. Numerical motivation. The kinetic structure of hyperbolic conservation laws have been used for a long time to construct entropic numerical schemes (see Kaniel [20], Giga and Miyakawa [13], the "transport-collapse" method of Brenier [6], etc.). This method has been intensively developed by the group of Perthame (see [24] for a review). In this framework, study of hydrodynamical limits of BGK-like equations can give a first step for the proof of the convergence of those schemes.

Recently an intense activity has been produced to solve numerically hyperbolic conservation laws with source terms. As a test, the Saint–Venant system with bottom topography is often proposed:

$$\begin{cases} \partial_t h + \mathrm{div}_x(hu) = 0, \quad t > 0,\ x \in \mathbb{R}^2, \\ \partial_t (hu) + \mathrm{div}_x(hu \otimes u) + \nabla_x h^2 + Z'(x) h = 0, \quad t > 0,\ x \in \mathbb{R}^2, \\ (h, hu)|_{t=0} = (h^0, h^0 u^0), \quad x \in \mathbb{R}^2, \end{cases} \tag{1.2}$$

where Z is the given bottom topography, h is the unknown depth of the water, and u is the unknown velocity of the water. This system models the evolution of a river.

Different numerical methods have been proposed to solve such problems (see Gosse [16], [15], Jin [19], Gallouët, Hérard, and Seguin [12], etc.). Botchorishvili, Perthame, and Vasseur [4] developed a kinetic procedure to construct numerical schemes and showed the convergence in the scalar case. This method has been successfully implemented by Perthame and Simeoni [25] for the Saint–Venant system. Notice that the Saint–Venant system corresponds exactly to the sytem (1.1) with $\gamma = 2$ and $Z' = F$. Our result can be seen as a first attempt to show the convergence of kinetic schemes in this framework.

1.3. Idea of the proof. As mentioned above, the proof relies on a relative entropy method. We consider the following abstract conservation law:

$$\partial_t U + \operatorname{div}_x A(U) = Q(U, x), \tag{1.3}$$

with $U(t,x) \in \mathcal{U} \subset \mathbb{R}^p$ for $t \in \mathbb{R}^+$, $x \in \mathbb{R}^n$, $A : \mathcal{U} \to \mathbb{R}^p$, and $Q : \mathcal{U} \times \mathbb{R}^n \to \mathbb{R}^p$. We assume that there exists an entropy, entropy flux couple (η, G) with $\eta \in C^2(\mathcal{U}, \mathbb{R})$ convex such that

$$\partial_i G_k(W) = \sum_j \partial_j \eta(W)\, \partial_i A_{jk}(W) \quad \forall k, i, \ \forall W. \tag{1.4}$$

For smooth solutions of this system, we have the entropy equality

$$\partial_t \eta(U) + \partial_x G(U) = \eta'(U) Q(U, x). \tag{1.5}$$

Following the notations of Dafermos, for every function $\Phi \in C^1(\mathbb{R}^p)$ of U we introduce the associated related quantity $\Phi(\cdot|\cdot) \in C^0(\mathbb{R}^p \times \mathbb{R}^p)$:

$$\Phi(U_1|U_2) = \Phi(U_1) - \Phi(U_2) - \nabla\Phi(U_2)(U_1 - U_2). \tag{1.6}$$

For example, the relative entropy is defined by

$$\eta(U_1|U_2) = \eta(U_1) - \eta(U_2) - \eta'(U_2) \cdot (U_1 - U_2), \tag{1.7}$$

and the entropy flux by

$$A(U_1|U_2) = A(U_1) - A(U_2) - A'(U_2) \cdot (U_1 - U_2). \tag{1.8}$$

We notice that if Φ is convex, then $\Phi(U_1|U_2) \geq 0$. Moreover, if it is strictly convex, then $\Phi(U_1|U_2) = 0$ if and only if $U_1 = U_2$.

We consider in the same way an abstract kinetic equation

$$\partial_t f_\varepsilon + v \cdot \nabla_x f_\varepsilon + q(f_\varepsilon) = \frac{\mathcal{Q}(f_\varepsilon, v)}{\varepsilon}, \tag{1.9}$$

where $f_\varepsilon = f_\varepsilon(t, x, v) \in \mathbb{R}$ with $t \in \mathbb{R}^+$, $x, v \in \mathbb{R}^n$, and q a linear operator, $\mathcal{Q} : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ and with $a : \mathbb{R}^n \to \mathbb{R}^p$, where the collision term $\mathcal{Q}$ satisfies

$$\int_{\mathbb{R}^n} a(v) \mathcal{Q}(f, v)\, dv = 0 \quad \text{for any } f \in \mathbb{R}. \tag{1.10}$$

We assume the existence of a kinetic entropy $H : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ which is well related to the entropy η of the hyperbolic system we want to relax. In more precise words, we need that the following nonincrease is checked for the solution of the kinetic equation,

$$\frac{d}{dt} \iint_{\mathbb{R}^{2n}} H(f_\varepsilon, v)\, dv\, dx \leq \int_{\mathbb{R}^n} \eta'(U_\varepsilon) Q(U_\varepsilon)\, dx, \tag{1.11}$$

where

$$U_\varepsilon(t,x) = \int_{\mathbb{R}^n} a(v) f_\varepsilon(t,x,v)\, dv, \tag{1.12}$$

and that the following compatibility between the entropy η and the kinetic entropy H is satisfied:

$$\eta(U_\varepsilon) \le \int_{\mathbb{R}^n} H(f_\varepsilon, v)\, dv. \tag{1.13}$$

In addition we assume that

$$\begin{aligned} &\left| \int_{\mathbb{R}^n} (|a(v)| + |a(v)\otimes v|) f_\varepsilon\, dv + \int_{\mathbb{R}^n} a(v) q(f_\varepsilon)\, dv \right| \\ &\qquad \le \int_{\mathbb{R}^n} (f_\varepsilon + H(f_\varepsilon, v))\, dv. \end{aligned} \tag{1.14}$$

Then we have the following abstract theorem.

THEOREM 1.2. *Let $U \in [C^1([0,T]\times\mathbb{R}^n)]^p$ be a strong solution on $[0,T]$ of the multidimensional hyperbolic system* (1.3)*, a system with a convex, C^2 entropy, for an initial data U^0. We assume in addition that U, $\eta'(U)$, and $\partial_x \eta'(U)$ are bounded and that U and $\eta(U)$ are integrable with respect to x. Let f_ε be a solution to the kinetic equation* (1.9)*, satisfying* (1.10)–(1.14) *and $f_\varepsilon + H(f_\varepsilon, v)$ integrable with respect to x and v for every t. We set*

$$U_\varepsilon(t,x) = \int_{\mathbb{R}^n} a(v) f_\varepsilon(t,x,v)\, dv.$$

We assume the convergence of initial data

$$\int_{\mathbb{R}^n} \eta(U_\varepsilon^0 | U^0)\, dx \le C_0 \sqrt{\varepsilon}, \tag{1.15}$$

and the following compatibility for the initial data

$$\left| \int_{\mathbb{R}^n} H(f_\varepsilon^0, v)\, dv - \eta(U_\varepsilon^0) \right| \le C_0 \sqrt{\varepsilon}. \tag{1.16}$$

If we have the control of the kinetic quantities

$$\int_0^T \int_{\mathbb{R}^n} \left| A(U_\varepsilon) - \int_{\mathbb{R}^n} v \otimes a(v) f_\varepsilon\, dv \right| dx\, dt \le C_1 \sqrt{\varepsilon}, \tag{1.17}$$

$$\int_0^T \int_{\mathbb{R}^n} \left| Q(U_\varepsilon, x) + \int_{\mathbb{R}^n} a(v) q(f_\varepsilon)\, dv \right| dx\, dt \le C_1 \sqrt{\varepsilon}, \tag{1.18}$$

and the control of the relative flux and the source terms by the relative entropy as

$$|A(U_\varepsilon | U)| \le C_2\ \eta(U_\varepsilon | U), \tag{1.19}$$

$$|Q(U)\eta'(U_\varepsilon | U) + [Q(U_\varepsilon) - Q(U)](\eta'(U_\varepsilon) - \eta'(U))| \le C_2\ \eta(U_\varepsilon | U), \tag{1.20}$$

where C_1 and C_2 are positive constants, then we get, for a constant C,

$$\int_{\mathbb{R}^n} \eta(U_\varepsilon | U)(t,x)\, dx \le C\sqrt{\varepsilon} \quad \text{for any } t \in [0,T]. \tag{1.21}$$

Hypothesis (1.15) and (1.16) are compatibility conditions on the initial data. Hypothesis (1.19) and (1.20) are structure conditions on the system (1.3). This theorem shows that if (1.17) and (1.18) are fulfilled (which will be derived from kinetic dissipation), and the system (1.3) has a good structure, then the total relative entropy of U_ε with respect to U converges to 0. This applies to the convergence of U_ε to U. Notice that this presentation splits nicely the kinetic dissipation effect from the control of the nonlinearities. The kinetic dissipation is needed in order to fulfill the consistency (1.17) and (1.18). The nonlinearity is driven by the relative entropy method which can be applied if (1.3) verifies (1.19) and (1.20). Notice that the method depends only on the structure of the system whatever the kinetic equation is.

Then we show that the isentropic system (1.1) verifies the structure compatibility and that the involved kinetic equations verify the dissipation properties needed. Notice that the full Euler system (with the added energy equation) does not verify (1.19). Hence this method cannot be applied directly to the convergence from the Boltzmann equation to the Euler system, for instance (see the appendix). The problem relies already on the structure itself of the system (the relative flux cannot be controlled by the relative entropy because of the high macroscopic velocities). Of course an additional difficulty lies in the kinetic level to control high velocities to obtain (1.17) and (1.18).

2. Study of the abstract problem. We consider the abstract equation (1.3) and abstract kinetic equation (1.9). This section is devoted to the proof of Theorem 1.2.

2.1. The key estimate. In the following proposition, we describe the evolution of the relative entropy using canonical quantities associated with the system (1.3) and entropy equation (1.5). We do not claim any originality in this result. It can be found in [10], except for the slight generalization concerning the source term Q. However, we give the proof for the sake of completeness.

PROPOSITION 2.1. *For the entropy $\eta \in C^2(\mathbb{R}^p)$ and for any $U, V \in [C^1(\mathbb{R}^n)]^p$, we have*

$$\begin{aligned}
\partial_t \eta(V|U) = {} & [\partial_t \eta(V) + \operatorname{div}_x G(V) - \eta'(V)Q(V)] \\
& - [\partial_t \eta(U) + \operatorname{div}_x G(U) - \eta'(U)Q(U)] \\
& - \eta''(U) \cdot [\partial_t U + \operatorname{div}_x A(U) - Q(U))] \cdot (V - U) \\
& - \eta'(U) \cdot [\partial_t V + \operatorname{div}_x A(V) - Q(V)] \\
& + \eta'(U) \cdot [\partial_t U + \operatorname{div}_x A(U) - Q(U))] \\
& + \operatorname{div}_x [G(U) - G(V)] + \sum_{ik} \partial_{x_k} [\partial_i G_k(U)(V_i - U_i)] \\
& + \sum_{jk} \partial_j \eta(U) \partial_{x_k} [A(V|U)] \\
& + Q(U)\eta'(V|U) + [Q(V) - Q(U)](\eta'(V) - \eta'(U)).
\end{aligned}$$

Remark 2.2. Notice that if U and V are regular solutions to (1.3), the first five lines vanish. The sixth line has a divergence form, hence its integral is vanishing. Finally, the two last terms are quadratic with respect to $V-U$ (at least when $|V-U| \leq R$) as η is. Hence, from this proposition, we can expect to have a good structure to use Gronwall's lemma on $\int \eta(V|U)\, dx$.

Proof. From the definition of relative quantity (1.6), we have

$$\begin{aligned}\partial_t\eta(V|U) &= \partial_t\eta(V) - \partial_t\eta(U) - \partial_t\left[\eta'(U)\right]\cdot(V-U) - \eta'(U)\cdot\partial_t(V-U)\\ &= \left[\partial_t\eta(V) + \operatorname{div}_x G(V) - \eta'(V)Q(V)\right]\\ &\quad - \left[\partial_t\eta(U) + \operatorname{div}_x G(U) - \eta'(U)Q(U)\right]\\ &\quad - \eta''(U)\cdot\left[\partial_t U + \operatorname{div}_x A(U) - Q(U)\right]\cdot(V-U)\\ &\quad - \eta'(U)\cdot\left[\partial_t V + \operatorname{div}_x A(V) - Q(V)\right]\\ &\quad + \eta'(U)\cdot\left[\partial_t U + \operatorname{div}_x A(U) - Q(U)\right] + R_1 + R_2,\end{aligned} \tag{2.1}$$

where

$$\begin{aligned}R_1 &= \eta'(V)Q(V) - \eta'(U)Q(U) - \eta''(U)\cdot Q(U)\cdot(V-U)\\ &\quad - \eta'(U)\cdot Q(V) + \eta'(U)\cdot Q(U)\\ &= Q(U)\eta'(V|U) + \left[\eta'(U) - \eta'(V)\right]\cdot\left[Q(U) - Q(V)\right]\end{aligned} \tag{2.2}$$

and

$$\begin{aligned}R_2 &= \operatorname{div}_x\left[G(U) - G(V)\right]\\ &\quad + \eta''(U)\cdot\operatorname{div}_x A(U)\cdot(V-U)\\ &\quad + \eta'(U)\cdot\operatorname{div}_x\left[A(V) - A(U)\right].\end{aligned}$$

The existence of the associated entropy flux G gives the relation (see (1.4))

$$\partial_i G_k(W) = \sum_j \partial_j\eta(W)\,\partial_i A_{jk}(W) \quad \forall k, i,\ \forall W.$$

A derivation of this relation with respect to W_l gives

$$\sum_j \partial_{lj}\eta(W)\partial_i A_{jk}(W) = \partial_{il}G_k(W) - \sum_j \partial_j\eta(W)\partial_{il}A_{jk}(W).$$

We use this relation with $W = U$ and get

$$\begin{aligned}&\eta''(U)\cdot\operatorname{div}_x A(U)\cdot(V-U)\\ &\quad = \sum \partial_{lj}\eta(U)\partial_{x_k}[A_{jk}(U)](V_l - U_l)\\ &\quad = \sum \partial_{lj}\eta(U)\partial_i A_{jk}(U)\partial_{x_k}U_i(V_l - U_l)\\ &\quad = \sum \partial_{il}G_k(U)\partial_{x_k}U_i(V_l - U_l) - \sum \partial_j\eta(U)\partial_{il}A_{jk}(U)\partial_{x_k}U_i(V_l - U_l),\end{aligned}$$

and now

$$\begin{aligned}&-\partial_j\eta(U)\partial_{il}A_{jk}(U)\partial_{x_k}U_i(V_l - U_l)\\ &\quad = \partial_j\eta(U)\left[-\partial_{x_k}\left[\partial_l A_{jk}(U)\right](V_l - U_l)\right]\\ &\quad = \partial_j\eta(U)\left[-\partial_{x_k}\left[\partial_l A_{jk}(U)(V_l - U_l)\right] + \partial_l A_{jk}(U)\partial_{x_k}(V_l - U_l)\right];\end{aligned}$$

therefore, we obtain

$$\begin{aligned}
R_2 &= \operatorname{div}_x [G(U) - G(V)] + \sum \partial_{il} G_k(U) \partial_{x_k} U_i (V_l - U_l) \\
&\quad + \sum \partial_j \eta(U) \left[-\partial_{x_k} \left[\partial_l A_{jk}(U)(V_l - U_l) \right] + \partial_l A_{jk}(U) \partial_{x_k} (V_l - U_l) \right] \\
&\quad + \eta'(U) \cdot \operatorname{div}_x [a(V) - A(U)] \\
&= \operatorname{div}_x [G(U) - G(V)] + \sum \partial_{x_k} [\partial_l G_k(U)](V_l - U_l) \\
&\quad - \sum \partial_j \eta(U) \partial_{x_k} \left[\partial_l A_{jk}(U)(V_l - U_l) \right] \\
&\quad + \sum \partial_j \eta(U) \partial_l A_{jk}(U) \partial_{x_k} (V_l - U_l) \\
&\quad + \sum \partial_j \eta(U) \partial_{x_k} [A_{jk}(V) - A_{jk}(U)].
\end{aligned}$$

Permuting indexes i and l, we can rewrite (1.4) in the following way:

$$\sum_j \partial_j \eta(U) \partial_l A_{jk}(U) = \partial_l G_k(U).$$

Thus we find

$$\begin{aligned}
R_2 &= \operatorname{div}_x [G(U) - G(V)] + \sum \partial_{x_k} [\partial_i G_k(U)(V_i - U_i)] \\
&\quad + \sum \partial_j \eta(U) \partial_{x_k} [A(V|U)].
\end{aligned} \tag{2.3}$$

Equation (2.1), with (2.2) and (2.3), gives the desired relation. □

Remark 2.3. We notice that in particular, the term R_2 of the proof satisfies

$$\begin{aligned}
\int_{\mathbb{R}^n} R_2 \, dx &= \int_{\mathbb{R}^n} \sum_{jk} \partial_j \eta(U) \partial_{x_k} \left[A(V|U) \right] \, dx \\
&= - \int_{\mathbb{R}^n} \sum_{jk} \partial_{x_k} \left[\partial_j \eta(U) \right] A_{jk}(V|U) \, dx.
\end{aligned}$$

This result is now used to obtain information if one deals with weak, strong, or/and approximated solutions.

2.2. Weak and strong solutions. This subsection is completely imbedded in Dafermos [10] (except for the slight generalization of the source term). Moreover, it is completely independent of the remainder of the paper. We give it since it clarifies the structure conditions (1.19) and (1.20) needed to use the relative entropy method without a priori condition on V in L^∞ for instance. We assume here that U is a strong solution of (1.3) (and as a consequence (1.5) is satisfied), and that V is a weak solution of (1.3) satisfying the entropy inequality

$$\partial_t \eta(V) + \partial_x G(V) \leq \eta'(V) Q(V). \tag{2.4}$$

Thus applying Proposition 2.1 (on a regularization of V and passing to the limit in the regularization), we get with the notations (2.2) and (2.3)

$$\partial_t \eta(V|U) \leq R_1 + R_2, \tag{2.5}$$

and using Remark 2.3, it leads to the following result.

COROLLARY 2.4. *Let $U \in [C^1([0,T]\times\mathbb{R}^n)]^p$ be a strong solution of* (1.3) *such that U, $\eta'(U)$, $\partial_x\eta'(U)$ are bounded and U, $\eta(U)$ are integrable. Let V be an entropy weak solution of* (1.3). *Then we have*

$$\begin{aligned}\frac{d}{dt}\int_{\mathbb{R}^n}\eta(V|U)\,dx \le &-\int_{\mathbb{R}^n}\sum_{jk}\partial_{x_k}\left[\partial_j\eta(U)\right]A_{jk}(V|U)\,dx\\ &+\int_{\mathbb{R}^n}Q(U)\eta'(V|U)\,dx\\ &+\int_{\mathbb{R}^n}[Q(V)-Q(U)](\eta'(V)-\eta'(U))\,dx.\end{aligned}\tag{2.6}$$

This result clarifies necessary information needed on the structure of the system (1.3). If we have for every $V,U\in\mathbb{R}^p$

$$|A(V|U)| \le C\eta(V|U)\tag{2.7}$$

and

$$|Q(U)\eta'(V|U)+[Q(V)-Q(U)](\eta'(V)-\eta'(U))| \le C\eta(V|U),\tag{2.8}$$

then we get

$$\frac{d}{dt}\int_{\mathbb{R}^n}\eta(V|U)\,dx \le (C(U)+1)C\int_{\mathbb{R}^n}\eta(V|U)\,dx,$$

and by a Gronwall's argument, it gives

$$\int_{\mathbb{R}^n}\eta(V|U)(t,x)\,dx \le \int_{\mathbb{R}^n}\eta(V|U)(0,x)\,dx\ e^{(C(U)+1)Ct}.$$

Thus if $U^0=V^0$, then

$$\eta(V|U)(t,x)=0\quad\forall t\in[0,T],\ \text{a.e. } x\in\mathbb{R}^n.$$

It gives $V=U$ if η is strictly convex. We recover here part of the classical results for weak = strong solutions. In fact, estimates as (2.7)–(2.8) are the important point to perform our entropy method. If we do not have a source term, it says that we need a control of the relative flux of the system by the relative entropy. This was already the case in Brenier [8]. We want now to extend the possible applications by studying the link between a strong solution and some approximations of it.

2.3. Strong and approximated solutions. We now assume that U is a strong solution of (1.3), and U_ε is any approximation of a solution, coming for example from a kinetic model. We get the following corollary from Proposition 2.1.

COROLLARY 2.5. *Let $U\in[C^1([0,T]\times\mathbb{R}^n)]^p$ be a strong solution of* (1.3) *such that U, $\eta'(U)$, $\partial_x\eta'(U)$ are bounded and U, $\eta(U)$ are integrable. Then we have, for any function $U_\varepsilon\in[C^1([0,T]\times\mathbb{R}^n)]^p$,*

$$\begin{aligned}\partial_t\eta(U_\varepsilon|U) = &\left[\partial_t\eta(U_\varepsilon)+\operatorname{div}_x G(U_\varepsilon)-\eta'(U_\varepsilon)Q(U_\varepsilon)\right]\\ &-\eta'(U)\cdot\left[\partial_t U_\varepsilon+\operatorname{div}_x A(U_\varepsilon)-Q(U_\varepsilon)\right]\\ &+\operatorname{div}_x\left[G(U)-G(U_\varepsilon)\right]+\sum\partial_{x_k}[\partial_i G_k(U)((U_\varepsilon)_i-U_i)]\\ &+\sum_{jk}\partial_j\eta(U)\partial_{x_k}\left[A(U_\varepsilon|U)\right]\\ &+Q(U)\eta'(U_\varepsilon|U)+[Q(U_\varepsilon)-Q(U)](\eta'(U_\varepsilon)-\eta'(U)).\end{aligned}$$

In this situation we have

$$\begin{aligned}\frac{d}{dt}\int_{\mathbb{R}^n}[\eta(U_\varepsilon|U)-\eta(U_\varepsilon)]\,dx &= -\int_{\mathbb{R}^n}\eta'(U_\varepsilon)Q(U_\varepsilon)\,dx\\ &\quad -\int_{\mathbb{R}^n}\eta'(U)\cdot\left[\partial_t U_\varepsilon+\operatorname{div}_x A(U_\varepsilon)-Q(U_\varepsilon)\right]\,dx\\ &\quad -\int_{\mathbb{R}^n}\sum_{jk}\partial_{x_k}\left[\partial_j\eta(U)\right]A_{jk}(U_\varepsilon|U)\,dx\\ &\quad +\int_{\mathbb{R}^n}Q(U)\eta'(U_\varepsilon|U)\,dx\\ &\quad +\int_{\mathbb{R}^n}[Q(U_\varepsilon)-Q(U)](\eta'(U_\varepsilon)-\eta'(U))\,dx.\end{aligned}$$

We use now this relation in the case where U_ε comes from a kinetic equation.

2.4. Approximation from a kinetic equation. We consider f_ε a solution to the kinetic model (1.9) which satisfies (1.10)–(1.14) with $f_\varepsilon+H(f_\varepsilon,v)$ integrable with respect to x and v for every t. Let U_ε be the moments of f_ε defined by (1.12). We set

$$\Delta_\varepsilon=\eta(U_\varepsilon|U)+\int_{\mathbb{R}^n}H(f_\varepsilon,v)\,dv-\eta(U_\varepsilon). \tag{2.9}$$

From (1.13) and the convexity of η, we have

$$\Delta_\varepsilon\geq 0. \tag{2.10}$$

Using (1.11) and the relation of the previous section, we obtain (again after a regularization)

$$\begin{aligned}\frac{d}{dt}\int_{\mathbb{R}^n}\Delta_\varepsilon\,dx &\leq -\int_{\mathbb{R}^n}\eta'(U)\cdot\left[\partial_t U_\varepsilon+\operatorname{div}_x A(U_\varepsilon)-Q(U_\varepsilon)\right]\,dx\\ &\quad -\int_{\mathbb{R}^n}\sum_{jk}\partial_{x_k}\left[\partial_j\eta(U)\right]A_{jk}(U_\varepsilon|U)\,dx\\ &\quad +\int_{\mathbb{R}^n}Q(U)\eta'(U_\varepsilon|U)\,dx\\ &\quad +\int_{\mathbb{R}^n}[Q(U_\varepsilon)-Q(U)](\eta'(U_\varepsilon)-\eta'(U))\,dx.\end{aligned}$$

Now multiplying the kinetic equation (1.9) by $a(v)$ and then integrating it with respect to v and using (1.10), we have

$$\partial_t U_\varepsilon+\operatorname{div}_x\int_{\mathbb{R}^n}v\otimes a(v)f_\varepsilon\,dv+\int_{\mathbb{R}^n}a(v)q(f_\varepsilon)\,dv=0.$$

It gives

$$\begin{aligned}\partial_t U_\varepsilon+\operatorname{div}_x A(U_\varepsilon)-Q(U_\varepsilon) &= \operatorname{div}_x\left(A(U_\varepsilon)-\int_{\mathbb{R}^n}v\otimes a(v)f_\varepsilon\,dv\right)\\ &\quad -\left[\int_{\mathbb{R}^n}a(v)q(f_\varepsilon)\,dv+Q(U_\varepsilon)\right].\end{aligned}$$

Therefore we get the following result.

PROPOSITION 2.6. *We assume that the system* (1.3) *admits a strictly convex entropy* $\eta \in C^2(\mathbb{R}^p)$. *Let* $U \in [C^1([0,T]\times\mathbb{R}^n)]^p$ *be a strong solution of* (1.3) *such that* U, $\eta'(U)$, $\partial_x\eta'(U)$ *are bounded and* U, $\eta(U)$ *are integrable. Let* f_ε *be a solution of* (1.9) *such that* (1.10)–(1.14) *are satisfied and* $f_\varepsilon + H(f_\varepsilon, v)$ *are integrable with respect to* x *and* v *for every time* t. *We set*

$$U_\varepsilon(t,x) = \int_{\mathbb{R}^n} a(v) f_\varepsilon(t,x,v)\, dv.$$

Then there exists a constant $C(U)$ *such that*

$$\begin{aligned}
&\frac{d}{dt}\int_{\mathbb{R}^n}\left[\eta(U_\varepsilon|U) + \int_{\mathbb{R}^n} H(f_\varepsilon)\, dv - \eta(U_\varepsilon)\right] dx \\
&\le C(U)\left(\int_{\mathbb{R}^n}\left|A(U_\varepsilon) - \int_{\mathbb{R}^n} v\otimes a(v) f_\varepsilon\, dv\right| dx \right. \qquad (2.11)\\
&\qquad + \int_{\mathbb{R}^n}\left|Q(U_\varepsilon) + \int_{\mathbb{R}^n} a(v) q(f_\varepsilon)\, dv\right| dx \qquad (2.12)\\
&\qquad + \int_{\mathbb{R}^n} |\, Q(U)\eta'(U_\varepsilon|U) + [Q(U_\varepsilon) - Q(U)](\eta'(U_\varepsilon) - \eta'(U))\,|\, dx \qquad (2.13)\\
&\qquad \left. + \int_{\mathbb{R}^n} |A(U_\varepsilon|U)|\, dx\right). \qquad (2.14)
\end{aligned}$$

Remark 2.7. This inequality uncouples the various structures which come into play. The term (2.11) is related to the kinetic approximation, the term (2.14) is related to the structure of the system, the term (2.12) is related to the kinetic structure of the source term, and the term (2.13) is related to the structure of the source term with respect to the hyperbolic system.

We use this majoration to get the convergence result from a solution of a kinetic equation to a strong solution of a multidimensional hyperbolic system, that is, Theorem 1.2.

2.5. Proof of Theorem 1.2. We use again the notation Δ_ε given by (2.9). From Proposition 2.6, we get

$$\begin{aligned}
\frac{d}{dt}\int_{\mathbb{R}^n}\Delta_\varepsilon(t,x)\, dx \le C(U)\left(\int_{\mathbb{R}^n}\left|A(U_\varepsilon) - \int_{\mathbb{R}^n} v\otimes a(v) f_\varepsilon\, dv\right| dx\right.\\
+ \int_{\mathbb{R}^n}\left|Q(U_\varepsilon) + \int_{\mathbb{R}^n} a(v) q(f_\varepsilon)\, dv\right| dx\\
\left. + 2C_2\int_{\mathbb{R}^n}\eta(U_\varepsilon|U)\, dx\right),
\end{aligned}$$

and thus, for $t \in [0,T]$,

$$\begin{aligned}
&\int_{\mathbb{R}^n}\Delta_\varepsilon(t,x)\, dx\\
&\quad \le \int_{\mathbb{R}^n}\Delta_\varepsilon(0,x)\, dx + 2C(U)C_1\sqrt{\varepsilon} + 2C(U)C_2\int_0^t\int_{\mathbb{R}^n}\Delta_\varepsilon(s,x)\, dx\, ds.
\end{aligned}$$

By Gronwall's argument, it gives

$$\int_{\mathbb{R}^n}\Delta_\varepsilon(t,x)\, dx \le \left(\int_{\mathbb{R}^n}\Delta_\varepsilon(0,x)\, dx + 2C(U)C_1\sqrt{\varepsilon}\right) e^{2C(U)C_2 t}.$$

From (1.15)–(1.16), we have

$$\left| \int_{\mathbb{R}^n} \Delta_\varepsilon(0,x)\,dx \right| \leq C\sqrt{\varepsilon},$$

and consequently, since $0 \leq \eta(U_\varepsilon|U) \leq \Delta_\varepsilon$, we obtain the result. □

Two independant studies to apply this result for a given example are thus necessary: the study of the system structure and the study of the dissipation of the kinetic model.

3. Study of the system structure. We consider here the case of the multidimensional isentropic gas dynamics system (1.1) avoiding the appearance of the vacuum. The associated entropy is

$$\eta(\rho,\rho u) = \rho\frac{u^2}{2} + h(\rho), \tag{3.1}$$

where $h(\rho) = \frac{1}{\gamma-1}\rho^\gamma$ for $\gamma > 1$ and $h(\rho) = \rho\ln\rho$ for the isotherm case $\gamma = 1$. The existence of strong solution for this problem is related to the classical result for regular solution for hyperbolic systems endowed with a strong entropy (see for instance [11]). In order to apply the convergence result of the previous section, we require that the structure of the system and the source terms be controllable by the relative entropy. For the system (1.1), the relative entropy is given by

$$\eta(U_1|U_2) = \frac{\rho_1}{2}|u_1-u_2|^2 + h(\rho_1|\rho_2). \tag{3.2}$$

The relative flux of the system is

$$A(U_1|U_2) = (0, \rho_1(u_1-u_2)\otimes(u_1-u_2) + h(\rho_1|\rho_2)I). \tag{3.3}$$

We clearly have the existence of a constant C such that

$$|A(U_1|U_2)| \leq C\eta(U_1|U_2) \tag{3.4}$$

for every $U_1, U_2 \in \mathbb{R}^{n+1}$. This fulfills estimate (1.19).

For the system (1.1), the source terms reads

$$Q(\rho,\rho u,x) = (0,\rho F(x)). \tag{3.5}$$

This gives

$$Q(U_2)\eta'(U_1|U_2) = -(u_2-u_1)(\rho_2-\rho_1)F \tag{3.6}$$

and

$$[Q(U_1)-Q(U_2)](\eta'(U_1)-\eta'(U_2)) = (u_2-u_1)(\rho_2-\rho_1)F, \tag{3.7}$$

and finally

$$Q(U_2)\eta'(U_1|U_2) + [Q(U_1)-Q(U_2)](\eta'(U_1)-\eta'(U_2)) = 0. \tag{3.8}$$

Thus the term (1.20) associated with the system (1.1) does not appear in the system structure study.

We turn now to the study of the terms related to the kinetic structure.

4. Study of the kinetic structure. We begin introducing the kinetic model we are dealing with. For $\gamma > 1$ we consider the following BGK kinetic equation:

$$\partial_t f_\varepsilon + v \cdot \nabla_x f_\varepsilon + F(x) \cdot \nabla_v f_\varepsilon = \frac{M f_\varepsilon - f_\varepsilon}{\varepsilon}, \tag{4.1}$$

where the unknown is $f_\varepsilon = f_\varepsilon(t,x,v) \in \mathbb{R}$ with $t \in \mathbb{R}^+$, $x, v \in \mathbb{R}^n$. The force term $F : \mathbb{R}^n \to \mathbb{R}^n$ is given. The equilibrium function $M f_\varepsilon$ is defined in the following way:

$$M f_\varepsilon(t,x,v) = M(\rho_\varepsilon(t,x), \rho_\varepsilon u_\varepsilon(t,x), v) \tag{4.2}$$

with

$$\rho_\varepsilon(t,x) = \int_{\mathbb{R}^n} f_\varepsilon(t,x,v)\, dv,$$
$$\rho_\varepsilon u_\varepsilon(t,x) = \int_{\mathbb{R}^n} v f_\varepsilon(t,x,v)\, dv,$$

where the Maxwellian $M : \mathbb{R}^p \times \mathbb{R}^n \to \mathbb{R}$ is given by

$$M(\rho, \rho u, v) = \mathbf{1}_{|u-v|^n \le c_n \rho} \quad \text{for } \gamma = \frac{n+2}{n}, \tag{4.3}$$

$$M(\rho, \rho u, v) = c \left(\frac{2\gamma}{\gamma - 1} \rho^{\gamma - 1} - |v-u|^2 \right)_+^{d/2} \quad \text{else.} \tag{4.4}$$

The constants are given by

$$c_n = n/|\mathbb{S}_n|,$$
$$d = \frac{2}{\gamma - 1} - n,$$
$$c = \left(\frac{2\gamma}{\gamma - 1} \right)^{-1/(\gamma-1)} \frac{\Gamma(\frac{\gamma}{\gamma-1})}{\pi^{n/2} \Gamma(d/2+1)}.$$

In the isothermal case $\gamma = 1$ we consider the following Fokker–Planck equation:

$$\partial_t f_\varepsilon + v \cdot \nabla_x f_\varepsilon + F(x) \cdot \nabla_v f_\varepsilon = \frac{1}{\varepsilon} \operatorname{div}_v((v - u_\varepsilon) f_\varepsilon + \nabla_v f_\varepsilon), \tag{4.5}$$

where

$$\rho_\varepsilon = \int_{\mathbb{R}^n} f_\varepsilon\, dv, \qquad \rho_\varepsilon u_\varepsilon = \int_{\mathbb{R}^n} v f_\varepsilon\, dv. \tag{4.6}$$

This section is devoted to the proof of the estimates needed to apply Theorem 1.2 for each model. For each case, we will first show that it verifies (1.10)–(1.14), and in the second step that it verifies the more difficult estimates (1.17) and (1.18).

4.1. BGK structure for isentropic gas with $1 < \gamma \le (n+2)/n$. In this section, we start by the study of kinetic BGK equations whose hydrodynamic limit is the isentropic system with an external force field.

We denote

$$U_\varepsilon = (\rho_\varepsilon, \rho_\varepsilon u_\varepsilon),$$
$$U = (\rho, \rho u),$$
$$a(v) = (1, v),$$
$$q(f) = F(x) \cdot \nabla_v f,$$
$$\mathcal{Q}(f, v) = Mf - f.$$

The Maxwellian M satisfies (see Bouchut [5])

$$\int_{\mathbb{R}^n} a(v)M(U,v)\,dv = U \qquad \forall U \in \mathbb{R}^p, \tag{4.7}$$

$$\int_{\mathbb{R}^n} v \otimes a(v)M(U,v)\,dv = A(U) \qquad \forall U \in \mathbb{R}^p, \tag{4.8}$$

$$\int_{\mathbb{R}^n} a(v)q(M(U,v))\,dv = -Q(U,x) \quad \forall U \in \mathbb{R}^p, \quad x \in \mathbb{R}^n. \tag{4.9}$$

This is the classical compatibility conditions required for the kinetic equation to be related to the system. Notice that thanks to (4.7), we have (1.10).

The kinetic entropy is the following:

$$H(f,v) = \frac{|v|^2}{2} f \quad \text{for } \gamma = \frac{n+2}{n},$$
$$H(f,v) = \frac{|v|^2}{2} f + \frac{1}{2c^{2/d}} \frac{f^{1+2/d}}{1+2/d} \quad \text{else.}$$

We have (see Bouchut [5]) that, for any f satisfying $\int_{\mathbb{R}^n} (f + H(f,v))\,dv < \infty$, and denoting $U = \int_{\mathbb{R}^n} a(v)f(v)\,dv$, the following minimization principle holds:

$$\int_{\mathbb{R}^n} H(M(U,v),v)\,dv \le \int_{\mathbb{R}^n} H(f)\,dv, \tag{4.10}$$

and a compatibility between the entropy η and the kinetic entropy H is satisfied as

$$\int_{\mathbb{R}^n} H(M(U,v),v)\,dv = \eta(U) \quad \text{for any } U \in \mathbb{R}^p. \tag{4.11}$$

First notice that (1.14) is verified. As a consequence of (4.10) and (4.11), we get

$$\eta(U) = \int_{\mathbb{R}^n} H(M(U,v),v)\,dv \le \int_{\mathbb{R}^n} H(f,v)\,dv,$$

which in particular gives (1.13). We prove now the decrease (1.11). We give it for $\gamma = (n+2)/n$. The other case is similar. Multiplying (4.1) by $|v|^2/2$ and then integrating in (v,x), we get

$$\begin{aligned} \frac{d}{dt} \iint_{\mathbb{R}^{2n}} \frac{|v|^2}{2} f_\varepsilon\,dv\,dx &= \iint_{\mathbb{R}^{2n}} F(x) \cdot v f_\varepsilon\,dv\,dx \\ &\quad + \frac{1}{\varepsilon} \iint_{\mathbb{R}^{2n}} \frac{|v|^2}{2} (M f_\varepsilon - f_\varepsilon)\,dv\,dx, \end{aligned} \tag{4.12}$$

since

$$\iint_{\mathbb{R}^{2n}} \frac{|v|^2}{2} F(x) \cdot \nabla_v f_\varepsilon\,dv\,dx = -\iint_{\mathbb{R}^{2n}} F(x) \cdot v f_\varepsilon\,dv\,dx.$$

In particular, from (4.10), it gives

$$\begin{aligned} \frac{d}{dt} \iint_{\mathbb{R}^{2n}} H(f_\varepsilon,v)\,dv\,dx &\le \int_{\mathbb{R}^n} F(x) \cdot \left(\int_{\mathbb{R}^n} v f_\varepsilon\,dv \right) dx \\ &\le \int_{\mathbb{R}^n} F(x) \cdot \rho_\varepsilon u_\varepsilon\,dx. \end{aligned} \tag{4.13}$$

Since

$$\int_{\mathbb{R}^n} \eta'(U_\varepsilon) Q(U_\varepsilon)\, dx = \int_{\mathbb{R}^n} F(x) \cdot \rho_\varepsilon u_\varepsilon\, dx,$$

this leads to (1.11).

We want now to prove (1.17)–(1.18). Since

$$A(U_\varepsilon) = \int_{\mathbb{R}^n} v \otimes a(v) M f_\varepsilon\, dv$$

and

$$Q(U_\varepsilon) = \sum_j \int_{\mathbb{R}^n} F_j(x) \partial_{v_j} a(v) M f_\varepsilon\, dv,$$

it gives

$$\begin{aligned} &\int_0^T \int_{\mathbb{R}^n} \left| A(U_\varepsilon) - \int_{\mathbb{R}^n} v \otimes a(v) f_\varepsilon\, dv \right| dx\, dt \\ &\quad = \int_0^T \int_{\mathbb{R}^n} \left| \int_{\mathbb{R}^n} v \otimes a(v)(M f_\varepsilon - f_\varepsilon)\, dv \right| dx\, dt \end{aligned} \tag{4.14}$$

and

$$\begin{aligned} &\int_0^T \int_{\mathbb{R}^n} \left| Q(U_\varepsilon, x) - \int_{\mathbb{R}^n} F(x) \nabla_v a(v) f_\varepsilon\, dv \right| dx\, dt \\ &\quad = \int_0^T \int_{\mathbb{R}^n} \left| \int_{\mathbb{R}^n} F(x) \nabla_v a(v)(f_\varepsilon - M f_\varepsilon)\, dv \right| dx\, dt. \end{aligned} \tag{4.15}$$

We have

$$\int_{\mathbb{R}^n} \partial_{v_i} a_j(v)(f_\varepsilon - M f_\varepsilon)\, dv = \delta_{i+1,j} \int_{\mathbb{R}^n} (f_\varepsilon - M f_\varepsilon)\, dv = 0;$$

thus the kinetic structure of the source term (1.18) vanishes. In order to apply the convergence result for this kinetic model, it only remains to control the entropy dissipation (1.17). It is the technical point of this example.

4.1.1. Control of the entropy dissipation. This subsection is devoted to the proof of the following proposition.

PROPOSITION 4.1. *Let f_ε be a solution to the BGK equation of the previous section with initial value f_ε^0 bounded in $L^1(\mathbb{R}^{2n})$ verifying (finite energy)*

$$\iint_{\mathbb{R}^{2n}} |v|^2 f_\varepsilon^0(x,v)\, dv\, dx \le C^0 < \infty, \tag{4.16}$$

and with $\gamma = (n+2)/n$. Then there exists C_n such that for every $\varepsilon < 1$, we have

$$\int_0^T \int_{\mathbb{R}^n} \left| \int_{\mathbb{R}^n} v \otimes a(v)(M f_\varepsilon - f_\varepsilon)\, dv \right| dx\, dt \le C_n \sqrt{\varepsilon}.$$

We define

$$D_\varepsilon(t,x)=\int_{\mathbb{R}^n}|v|^2(f_\varepsilon(t,x,v)-Mf_\varepsilon(t,x,v))\,dv.$$

From (4.13), we have

$$\begin{aligned}\frac{d}{dt}\iint_{\mathbb{R}^{2n}}H(f_\varepsilon,v)\,dv\,dx&\le\|F\|_{L^\infty}\iint_{\mathbb{R}^{2n}}|v|f_\varepsilon\,dx\,dv\\&\le\|F\|_{L^\infty}\iint_{\mathbb{R}^{2n}}(1+|v|^2)f_\varepsilon\,dx\,dv\\&\le\|F\|_{L^\infty}\left(\|\rho_\varepsilon\|_{L^1}+2\iint_{\mathbb{R}^{2n}}H(f_\varepsilon,v)\,dx\,dv\right);\end{aligned}$$

thus by Gronwall's argument, we obtain

$$\iint_{\mathbb{R}^{2n}}|v|^2f_\varepsilon(t,x,v)\,dv\,dx\le C,\quad 0\le t\le T. \tag{4.17}$$

Integrating now (4.13) with respect to t leads to

$$\begin{aligned}&\int_0^T\int_{\mathbb{R}^n}D_\varepsilon(t,x)\,dx\,dt\\&\quad\le\varepsilon\left(\iint_{\mathbb{R}^{2n}}|v|^2f^0(x,v)\,dv\,dx-2\iiint_{[0,T]\times\mathbb{R}^{2n}}F(x)vf_\varepsilon\,dv\,dx\,dt\right)\\&\quad\le\varepsilon\left(C^0+2\|F\|_{L^\infty}\iiint_{[0,T]\times\mathbb{R}^{2n}}|v|f_\varepsilon\,dv\,dx\,dt\right)\\&\quad\le\varepsilon\left(C^0+2\|F\|_{L^\infty}\iiint_{[0,T]\times\mathbb{R}^{2n}}(1+|v|^2)f_\varepsilon\,dv\,dx\,dt\right)\\&\quad\le\varepsilon(C^0+2\|F\|_{L^\infty}T\|f_\varepsilon^0\|_{L^1}+2\|F\|_{L^\infty}C)\\&\quad\le\varepsilon\,\tilde{C}.\end{aligned}\tag{4.18}$$

This gives a bound in ε for

$$\int_0^T\iint_{\mathbb{R}^{2n}}|v|^2(f_\varepsilon-Mf_\varepsilon)\,dv\,dx\,dt,$$

but we need to control

$$\int_0^T\int_{\mathbb{R}^n}\left|\int_{\mathbb{R}^n}v\otimes a(v)(Mf_\varepsilon-f_\varepsilon)\,dv\right|dx\,dt,$$

which is more delicate.

We set $a_1(v)=1$ and $a_2(v)=v$ such that $a=(a_1,a_2)$. Similarly, we define $A_1(U)=\rho u$ and $A_2(U)=\rho u\otimes u+I\rho^\gamma$.

Since

$$A(U_\varepsilon)=\int_{\mathbb{R}^n}v\otimes a(v)Mf_\varepsilon\,dv,$$

the first component of $\left|\int_{\mathbb{R}^n} v\otimes a(v)(Mf_\varepsilon - f_\varepsilon)\,dv\right|$ is still zero here. Now we have

$$\left|A_2(U_\varepsilon) - \int_{\mathbb{R}^n} v\otimes v f_\varepsilon\,dv\right| = \left|\int_{\mathbb{R}^n} (v-u)\otimes(v-u)(Mf_\varepsilon - f_\varepsilon)\,dv\right| \le \int_{\mathbb{R}^n} |v-u|^2 |Mf_\varepsilon - f_\varepsilon|\,dv.$$

The first equality uses (4.7). Thus to control the second component, we want to show that

$$\int_{\mathbb{R}^n} |v-u|^2 |Mf_\varepsilon - f_\varepsilon|\,dv$$

can be controlled (at least for bounded mass ρ) by the dissipation of entropy

$$\int_{\mathbb{R}^n} |v|^2 (f_\varepsilon - Mf_\varepsilon)\,dv.$$

It is the aim of the following proposition.

PROPOSITION 4.2. *For every $f\in L^1(\mathbb{R}^n)$ verifying $0\le f\le 1$, and every $u\in\mathbb{R}^n$ we denote*

$$\rho = \int_{\mathbb{R}^n} f(v)\,dv,$$
$$F = \int_{\mathbb{R}^n} |v-u|^2 |f(v) - M(\rho,u,v)|\,dv,$$
$$D = \int_{\mathbb{R}^n} |v|^2 (f(v) - M(\rho,u,v))\,dv.$$

Then there exists a constant C_n such that, for every $f\in L^1(\mathbb{R}^n)$ verifying $0\le f\le 1$,

$$F \le C_n(\rho^{\frac{n+2}{2n}}\sqrt{D} + D).$$

To prove this result, we first introduce some notations and prove preliminary results. Notice that, thanks to (4.7),

$$D = \int_{\mathbb{R}^n} |v-u|^2 (f(v) - M(\rho,u,v))\,dv.$$

Then changing v by $v+u$ if necessary, we see that we can restrict ourselves to the case $u=0$. We first reduce the problem to a one-dimensional problem. We introduce the following quantities:

$$\overline{f}(r) = \frac{1}{|\mathbb{S}_n|}\int_{\mathbb{S}_n} f(r\sigma)\,d\sigma,$$
$$\overline{M}(r) = \frac{1}{|\mathbb{S}_n|}\int_{\mathbb{S}_n} M(\rho,0,r\sigma)\,d\sigma = \mathbb{1}_{\{r^n\le c_n\rho\}}(r).$$

Since the integral of f is equal to the integral of $M(\rho,0,\cdot)$, we have

$$\int_0^\infty r^{n-1}\overline{f}(r)\,dr = \int_0^\infty r^{n-1}\overline{M}(r)\,dr. \tag{4.19}$$

We denote $r_1 = (c_n\rho)^{\frac{1}{n}}$, and we have

$$\begin{aligned}
F &= |\mathbb{S}_n| \int_0^\infty r^{n+1} |\overline{f}(r) - \overline{M}(r)|\, dr \\
&= |\mathbb{S}_n| \left(\int_0^{r_1} r^{n+1} (1 - \overline{f}(r))\, dr + \int_{r_1}^\infty r^{n+1} \overline{f}(r)\, dr \right), \\
D &= |\mathbb{S}_n| \int_0^\infty r^{n+1} (\overline{f}(r) - \overline{M}(r))\, dr \\
&= |\mathbb{S}_n| \left(- \int_0^{r_1} r^{n+1} (1 - \overline{f}(r))\, dr + \int_{r_1}^\infty r^{n+1} \overline{f}(r)\, dr \right).
\end{aligned}$$

We define in addition

$$M = \int_0^{r_1} r^{n-1} (1 - \overline{f}(r))\, dr = \int_{r_1}^\infty r^{n-1} \overline{f}(r)\, dr;$$

the last equality comes from (4.19) and $\overline{M}(r) = \mathbb{1}_{\{r \le r_1\}}(r)$. We have to do a different treatment for values close to r_1 and far from this value. For this purpose we consider $r_2 > r_1$ a new number which will be fixed later on. Then we denote

$$\begin{aligned}
M_1 &= \int_{r_1}^{r_2} r^{n-1} \overline{f}(r)\, dr, \\
M_2 &= \int_{r_2}^\infty r^{n-1} \overline{f}(r)\, dr.
\end{aligned}$$

We have $M = M_1 + M_2$. Then we define $0 < r_0 < r_1$ (in a unique way when r_2 is chosen) in the following way:

$$M_1 = \int_{r_0}^{r_1} r^{n-1} (1 - \overline{f}(r))\, dr.$$

Then, from the definition of M and since M is the sum of M_1 and M_2, we have

$$M_2 = \int_0^{r_0} r^{n-1} (1 - \overline{f}(r))\, dr.$$

In the same way we define F_1, F_2, D_1, D_2 in the following way:

$$\begin{aligned}
F_1 &= \int_{r_0}^{r_2} r^{n+1} |\overline{f}(r) - \overline{M}(r)|\, dr \\
&= \int_{r_0}^{r_1} r^{n+1} (1 - \overline{f}(r))\, dr + \int_{r_1}^{r_2} r^{n+1} \overline{f}(r)\, dr, \\
F_2 &= \int_0^{r_0} r^{n+1} |\overline{f}(r) - \overline{M}(r)|\, dr + \int_{r_2}^\infty r^{n+1} |\overline{f}(r) - \overline{M}(r)|\, dr \\
&= \int_0^{r_0} r^{n+1} (1 - \overline{f}(r))\, dr + \int_{r_2}^\infty r^{n+1} \overline{f}(r)\, dr,
\end{aligned}$$

$$
\begin{aligned}
D_1 &= \int_{r_0}^{r_2} r^{n+1}(\overline{f}(r) - \overline{M}(r))\,dr \\
&= -\int_{r_0}^{r_1} r^{n+1}(1-\overline{f}(r))\,dr + \int_{r_1}^{r_2} r^{n+1}\overline{f}(r)\,dr, \\
D_2 &= \int_{0}^{r_0} r^{n+1}(\overline{f}(r) - \overline{M}(r))\,dr + \int_{r_2}^{\infty} r^{n+1}(\overline{f}(r) - \overline{M}(r))\,dr \\
&= -\int_{0}^{r_0} r^{n+1}(1-\overline{f}(r))\,dr + \int_{r_2}^{\infty} r^{n+1}\overline{f}(r)\,dr.
\end{aligned}
$$

Notice that F_1, F_2, M_1, M_2 are nonnegative (as integrals of nonnegative functions) and verify

$$
\begin{aligned}
M &= M_1 + M_2, \\
F &= F_1 + F_2, \\
D &= D_1 + D_2.
\end{aligned}
$$

We can show, in addition, that D_1 and D_2 are nonnegative too.

LEMMA 4.3. *We have*

$$D_1 \geq 0, \qquad D_2 \geq 0.$$

Proof. We show the result for D_1 (the proof is similar for D_2). We have

$$
\begin{aligned}
\int_{r_1}^{r_2} r^{n+1}\overline{f}(r)\,dr &= \int_{r_1}^{r_2} r^2(r^{n-1}\overline{f}(r))\,dr \geq r_1^2 M_1, \\
\int_{r_0}^{r_1} r^{n+1}\overline{f}(r)\,dr &= \int_{r_0}^{r_1} r^2(r^{n-1}\overline{f}(r))\,dr \leq r_1^2 M_1.
\end{aligned}
$$

Since D_1 is the difference of those two terms, we find that D_1 is nonnegative. ∎

We first consider the values far from r_1.

LEMMA 4.4. *We can dominate F_2 by D_2 in the following way:*

$$F_2 \leq D_2 \left(\frac{r_1^2 + r_2^2}{r_2^2 - r_1^2} \right).$$

Proof. We have

$$
\begin{aligned}
\int_{r_2}^{\infty} r^{n+1}\overline{f}(r)\,dr &\geq r_2^2 M_2 \\
&\geq r_2^2 \frac{1}{r_0^2} \int_0^{r_0} r^{n+1}(1-\overline{f}(r))\,dr \\
&\geq \frac{r_2^2}{r_1^2} \int_0^{r_0} r^{n+1}(1-\overline{f}(r))\,dr.
\end{aligned}
$$

Hence we have

$$D_2 \geq \left(\frac{r_2^2}{r_1^2} - 1 \right) \int_0^{r_0} r^{n+1}(1-\overline{f}(r))\,dr.$$

But F_2 can be expressed in the following way:

$$F_2 = D_2 + 2\int_0^{r_0} r^{n+1}(1-\overline{f}(r))\,dr.$$

Those two expressions lead to

$$F_2 \le D_2 \left(\frac{r_1^2 + r_2^2}{r_2^2 - r_1^2} \right). \qquad \square$$

We consider now the values close to r_1.

LEMMA 4.5. *There exist $a > 0$ and a constant C_n depending only on n such that if $|r_2 - r_1| \le a r_1$, then*

$$F_1 \le C_n a^2 \rho^{\frac{n-2}{2n}} \sqrt{D_1}.$$

Proof. We split the proof in several parts.

(i) *Minimization of the entropy dissipation.* We define α and β such that

$$M_1 = \int_{r_1}^{\beta} r^{n-1}\, dr = \int_{\alpha}^{r_1} r^{n-1}\, dr.$$

From the definition of M_1, notice that $\beta \le r_2$. In the same way we have $\alpha \ge r_0$. We want to show that

$$D_1 \ge \int_{r_1}^{\beta} r^{n+1}\, dr - \int_{\alpha}^{r_1} r^{n+1}\, dr.$$

First we calculate

$$\begin{aligned}
&\int_{r_1}^{r_2} r^{n+1} \overline{f}(r)\, dr - \int_{r_1}^{\beta} r^{n+1}\, dr \\
&\quad = \int_{r_1}^{\beta} r^2 [r^{n-1}(\overline{f}(r) - 1)]\, dr + \int_{\beta}^{r_2} r^2 [r^{n-1} \overline{f}(r)]\, dr \\
&\quad = \int_{\beta}^{r_2} r^2 [r^{n-1} \overline{f}(r)]\, dr - \int_{r_1}^{\beta} r^2 [r^{n-1}(1 - \overline{f}(r))]\, dr \\
&\quad \ge \beta^2 \left[\int_{\beta}^{r_2} r^{n-1} \overline{f}(r)\, dr - \int_{r_1}^{\beta} r^{n-1}(1 - \overline{f}(r))\, dr \right] \\
&\quad \ge \beta^2 (M_1 - M_1) = 0.
\end{aligned}$$

In the same way we calculate

$$\begin{aligned}
&\int_{r_0}^{r_1} r^{n+1} (\overline{f}(r) - 1)\, dr + \int_{\alpha}^{r_1} r^{n+1}\, dr \\
&\quad = \int_{r_0}^{\alpha} r^{n+1} (\overline{f}(r) - 1)\, dr + \int_{\alpha}^{r_1} r^{n+1} \overline{f}(r)\, dr \\
&\quad \ge \alpha^2 \left[\int_{r_0}^{\alpha} r^{n-1} (\overline{f}(r) - 1)\, dr + \int_{\alpha}^{r_1} r^{n-1} \overline{f}(r)\, dr \right] \\
&\quad \ge 0.
\end{aligned}$$

Summing those two last inequalities gives the desired result.

(ii) *Taylor expansion of the critical entropy dissipation.* We call critical entropy dissipation the function defined by

$$D^c = \left(\int_{r_1}^{\beta} r^{n+1}\, dr - \int_{\alpha}^{r_1} r^{n+1}\, dr \right),$$

where α and β are defined in (i). Then we have

$$\begin{aligned} nM_1 &= \beta^n - r_1^n, \\ nM_1 &= r_1^n - \alpha^n, \\ (n+2)D^c &= \beta^{n+2} - 2r_1^{n+2} + \alpha^{n+2}, \end{aligned}$$

and therefore

$$\begin{aligned} \frac{D^c}{r_1^{n+2}} = \frac{\alpha+\beta-2r_1}{r_1} + \frac{n+1}{2}\left(\left(\frac{\beta-r_1}{r_1}\right)^2 + \left(\frac{\alpha-r_1}{r_1}\right)^2\right) \\ + O\left(\left(\frac{\beta-r_1}{r_1}\right)^3 + \left(\frac{\alpha-r_1}{r_1}\right)^3\right). \end{aligned}$$

Now

$$\begin{aligned} \frac{M_1}{r_1^n} &= \frac{\beta-r_1}{r_1} + \frac{n-1}{2}\left(\frac{\beta-r_1}{r_1}\right)^2 + O\left(\frac{\beta-r_1}{r_1}\right)^3 \\ &= \frac{r_1-\alpha}{r_1} - \frac{n-1}{2}\left(\frac{r_1-\alpha}{r_1}\right)^2 + O\left(\frac{r_1-\alpha}{r_1}\right)^3, \end{aligned}$$

and hence

$$\begin{aligned} 0 = \frac{\beta+\alpha-2r_1}{r_1} + \frac{n-1}{2}\left[\left(\frac{\beta-r_1}{r_1}\right)^2 + \left(\frac{r_1-\alpha}{r_1}\right)^2\right] \\ + O\left(\left(\frac{\beta-r_1}{r_1}\right)^3 + \left(\frac{r_1-\alpha}{r_1}\right)^3\right). \end{aligned}$$

Finally, we obtain

$$\begin{aligned} \frac{D^c}{r_1^{n+2}} &= \left[\left(\frac{\beta-r_1}{r_1}\right)^2 + \left(\frac{r_1-\alpha}{r_1}\right)^2\right] + O\left(\left(\frac{\beta-r_1}{r_1}\right)^3 + \left(\frac{r_1-\alpha}{r_1}\right)^3\right) \\ &= 2\left(\frac{M_1}{r_1^n}\right)^2 + O\left(\left(\frac{\beta-r_1}{r_1}\right)^3 + \left(\frac{r_1-\alpha}{r_1}\right)^3\right). \end{aligned}$$

Hence, there exist $\eta > 0$ and $\delta > 0$ such that

$$D^c \geq \frac{\delta}{r_1^{n-2}} M_1^2$$

whenever

$$\left|\frac{\beta-r_1}{r_1}\right| + \left|\frac{r_1-\alpha}{r_1}\right| \leq \eta.$$

(iii) *Final estimation.* From the definition of α, there exists $a > 0$ such that $|\frac{r_1-\alpha}{r_1}| \leq \eta$ whenever $|\frac{\beta-r_1}{r_1}| \leq a$. Remember that $r_2 \leq \beta$. Hence if $|r_2 - r_1| \leq ar_1$, then

$$\left|\frac{\beta-r_1}{r_1}\right| + \left|\frac{r_1-\alpha}{r_1}\right| \leq \eta$$

and

$$\begin{aligned} F_1 &\le r_2^2 M_1 \le a^2 \delta \sqrt{D^c} r_1^{\frac{n+2}{2}} \\ &\le C_n a^2 \sqrt{D_1} \rho^{\frac{n+2}{2n}}. \end{aligned}$$

The first inequality uses the definition of F_1, the second one uses the result of (ii), and the third one uses the definition of r_1 and the result of (i). ∎

Now we are able to prove the estimate of Proposition 4.2.

Proof of Proposition 4.2. We fix a and r_2 verifying the properties of Lemma 4.5. Thanks to Lemmas 4.4 and 4.5, we have

$$\begin{aligned} F \le F_1 + F_2 &\le D_2 \left(\frac{1+a}{a} \right) + C_n \rho^{\frac{n+2}{2n}} \sqrt{D_1} \\ &\le C_n'(D + \rho^{\frac{n+2}{2n}} \sqrt{D}). \end{aligned}$$ ∎

We are now able to prove the announced result.

Proof of Proposition 4.1. Thanks to Proposition 4.2, we have

$$\begin{aligned} &\int_0^T \int_{\mathbb{R}^n} \left| \int_{\mathbb{R}^n} v \otimes a(v)(M f_\varepsilon - f_\varepsilon)\, dv \right| dx\, dt \\ &\le C_n \sqrt{\left(\int_0^T \int_{\mathbb{R}^n} \rho_\varepsilon^{\frac{n+2}{n}}(t,x,v)\, dx\, dt \right) \left(\int_0^T \iint_{\mathbb{R}^{2n}} D_\varepsilon(t,x)\, dx\, dt \right)} \\ &\quad + C_n \int_0^T \iint_{\mathbb{R}^{2n}} D_\varepsilon(t,x)\, dx\, dt. \end{aligned}$$

From (4.17) and

$$\begin{aligned} \rho_\varepsilon |u_\varepsilon|^2 + n \rho_\varepsilon^{\frac{n+2}{n}} &= \int_{\mathbb{R}^n} |v|^2 M f_\varepsilon(t,x,v)\, dv \\ &\le \int_{\mathbb{R}^n} |v|^2 f_\varepsilon(t,x,v)\, dv, \end{aligned} \tag{4.20}$$

we have

$$\int_0^T \int_{\mathbb{R}^n} \rho_\varepsilon^{\frac{n+2}{n}}(t,x,v)\, dx\, dt \le \frac{T}{n} C.$$

Using (4.18), those lead to

$$\begin{aligned} &\int_0^T \int_{\mathbb{R}^n} \left| \int_{\mathbb{R}^n} v \otimes a(v)(M f_\varepsilon - f_\varepsilon)\, dv \right| dx\, dt \\ &\le C_n \sqrt{\frac{\varepsilon T}{n}} \tilde{C}^{1/2} + C_n \varepsilon \tilde{C}. \end{aligned}$$ ∎

We can then conclude the convergence result for $\gamma = (n+2)/n$.

Proof of Theorem 1.1. We apply Theorem 1.2 to get

$$\int_{\mathbb{R}^n} \eta(U_\varepsilon | U)(t,x)\, dx \to 0 \quad \text{for } t \in [0,T], \text{ as } \varepsilon \to 0. \tag{4.21}$$

Now

$$\eta(U_\varepsilon|U) = \int_0^1 \eta''(U + \vartheta(U_\varepsilon - U)) \cdot (U_\varepsilon - U)^2 \vartheta \, d\vartheta$$

with

$$\eta''(\rho, \rho u) \cdot (X_0, X_1)^2 = \gamma \rho^{\gamma-2} X_0^2 + \frac{1}{\rho}(X_1 - u X_0)^2,$$

and thus we have

$$\int_{\mathbb{R}^n} \int_0^1 \vartheta(\rho + \vartheta(\rho_\varepsilon - \rho))^{\gamma-2}(\rho_\varepsilon - \rho)^2 \, d\vartheta \, dx \to 0 \quad \text{for } t \in [0,T], \text{ as } \varepsilon \to 0 \tag{4.22}$$

and

$$\int_{\mathbb{R}^n} \int_0^1 \frac{\vartheta \rho^2}{(\rho + \vartheta(\rho_\varepsilon - \rho))^3}(\rho_\varepsilon(u_\varepsilon - u))^2 \, d\vartheta \, dx \to 0 \quad \text{for } t \in [0,T], \text{ as } \varepsilon \to 0. \tag{4.23}$$

For $\gamma \geq 2$, (4.22) gives that, up to a subsequence, $\rho_\varepsilon \to \rho$ a.e. as $\varepsilon \to 0$ since $\rho > 0$. For $\gamma < 2$, it gives this result except at the points where $\rho_\varepsilon \to +\infty$, but in this case, as ρ stays bounded,

$$(\rho + \vartheta(\rho_\varepsilon - \rho))^{\gamma-2}(\rho_\varepsilon - \rho)^2 \underset{\varepsilon \to 0}{\sim} \vartheta^{\gamma-2}\rho_\varepsilon^\gamma \to 0 \text{ a.e.},$$

and thus this case is impossible. Now (4.23) gives that, up to a subsequence, $\rho_\varepsilon(u_\varepsilon - u) \to 0$ a.e. as $\varepsilon \to 0$ and therefore $\rho_\varepsilon u_\varepsilon \to \rho u$ a.e. as $\varepsilon \to 0$. But from (4.17) and (4.20), we have that ρ_ε is bounded in $L^\infty(0,T;L^\gamma(\mathbb{R}^n))$ and that $\sqrt{\rho_\varepsilon}u_\varepsilon$ is bounded in $L^\infty(0,T;L^2(\mathbb{R}^n))$. Hence, the whole family ρ_ε converges strongly in $L^\infty(0,T;L^p(\mathbb{R}^n))$ to ρ for every $1 \leq p < \gamma$ and $\sqrt{\rho_\varepsilon}u_\varepsilon$ converges strongly to $\sqrt{\rho}u$ in $L^\infty(0,T;L^q(\mathbb{R}^n))$ for every $1 \leq q < 2$. In particular, $\rho_\varepsilon u_\varepsilon$ converges strongly to ρu in $L^\infty(0,T;L^q(\mathbb{R}^n))$ for every $1 \leq q < 2\gamma/(\gamma+1)$. □

4.1.2. Extension to every γ. The previous model works for $\gamma = (n+2)/n$. In order to deal with the values of $\gamma \in]1,(n+2)/n[$, we use an other model which was introduced in [5] and is written as follows.

We consider the BGK equation

$$\partial_t f_\varepsilon + \xi \cdot \nabla_x f_\varepsilon + F(x) \cdot \nabla_\xi f_\varepsilon = \frac{M f_\varepsilon - f_\varepsilon}{\varepsilon}, \tag{4.24}$$

where $f_\varepsilon = f_\varepsilon(t,x,v) \in \mathbb{R}$ with $t \in \mathbb{R}^+$, $x \in \mathbb{R}^n$, $v = (\xi, I) \in \mathbb{R}^n \times \mathbb{R}^+$, and $F : \mathbb{R}^n \to \mathbb{R}^n$, with

$$M f_\varepsilon(t,x,v) = M(U_\varepsilon(t,x), v), \qquad U_\varepsilon(t,x) = \int_{\mathbb{R}^{n+1}} a(v) f_\varepsilon(t,x,v) \, dv \tag{4.25}$$

with $a : \mathbb{R}^{n+1} \to \mathbb{R}^p$, $a(v) = (1, \xi)$, $p = n+1$, and

$$dv = b_1 I^{d-1} \, dI \, d\xi, \qquad b_1 = 2\,\pi^{d/2}/\Gamma(d/2),$$

where d is the number of degrees of freedom satisfying

$$n + d = \frac{2}{\gamma - 1}. \tag{4.26}$$

We notice that the function f_ε takes its values in $[0, b_2]$, and the Maxwellian M is defined by

$$M(U,v) = b_2 \mathbb{1}_{|\xi-u|^2+I^2<b_3\rho^{\gamma-1}}, \quad U=(\rho,\rho u), \tag{4.27}$$

where

$$b_2 = \left(\frac{2\pi\gamma}{\gamma-1}\right)^{-1/(\gamma-1)} \Gamma\left(\frac{\gamma}{\gamma-1}\right), \qquad b_3 = \frac{2\gamma}{\gamma-1},$$

and satisfies (4.7)–(4.8). It satisfies also (4.9) with $\partial_{\xi_j} a(v)$ instead of $\partial_{v_j} a(v)$. The kinetic entropy is

$$H(f,v) = \frac{1}{2}|v|^2 f = \frac{1}{2}(|\xi|^2 + I^2) f, \tag{4.28}$$

and satisfies (4.10)–(4.11). We get, as in the previous kinetic model, (1.11) and (1.18).

We recover the BGK model introduced previously integrating (4.24) (and the function f_ε) with respect to I with the measure $b_1 I^{d-1} dI$ [5].

Now for the control of the dissipation, we set

$$\overline{f}(r) = \frac{1}{s_n}\int_{\mathbb{S}^+_{n+1}} f(r\sigma)(\cos\theta)^{d-1}\, d\sigma,$$
$$\overline{M}(r) = \frac{1}{s_n}\int_{\mathbb{S}^+_{n+1}} M(\rho,0,r\sigma)(\cos\theta)^{d-1}\, d\sigma = b_2 \mathbb{1}_{\{r^2\le b_3\rho^{\gamma-1}\}}(r),$$

where $\mathbb{S}^+_{n+1} = \{(\xi, I) \in \mathbb{S}_{n+1};\ I \ge 0\}$, $I = r\cos\theta$ and $s_n = \int_{\mathbb{S}^+_{n+1}} (\cos\theta)^{d-1}\, d\sigma$. Then, we get from the mass conservation

$$\int_0^\infty r^{n+d-1}\overline{f}(r)\, dr = \int_0^\infty r^{n+d-1}\overline{M}(r)\, dr. \tag{4.29}$$

By similar techniques to those in the previous section, we get the following estimate.

PROPOSITION 4.6. *For every $f \in L^1_{dv}(\mathbb{R}^{n+1})$ verifying $0 \le f \le b_2$ and every $u \in \mathbb{R}^n$ we denote*

$$\rho = \int_{\mathbb{R}^{n+1}} f(v)\, dv,$$
$$F = \int_{\mathbb{R}^{n+1}} |v-u|^2 |f(v) - M(\rho,u,v)|\, dv,$$
$$D = \int_{\mathbb{R}^{n+1}} |v|^2 (f(v) - M(\rho,u,v))\, dv.$$

Then there exists a constant C_n such that, for every $f \in L^1_{dv}(\mathbb{R}^{n+1})$ verifying $0 \le f \le b_2$,

$$F \le C_n(\rho^{\frac{n+d+2}{2(n+d)}}\sqrt{D} + D).$$

Now since

$$\frac{n+d+2}{n+d} = \gamma,$$

we get a similar dissipation result to that in Proposition 4.1, and we can conclude the convergence in this case for every γ such that (4.26) is satisfied with $d > 0$; that is to say,

$$1 < \gamma < \frac{n+2}{n}.$$

We then obtain Theorem 1.1 in the same way as in the previous case.

We thank Bouchut for noticing that the proof of the case $\gamma = (n+2)/n$ (of the previous subsection) is also valid for every $1 < \gamma < (n+2)/n$ using this model.

4.2. Fokker–Planck. In this subsection, we study the convergence from the Fokker–Planck kinetic equation to the isothermal system, that is, the case $\gamma = 1$.

4.2.1. The kinetic model. The Fokker–Planck equation on $f_\varepsilon = f_\varepsilon(t,x,v) \in \mathbb{R}$, with $t \in \mathbb{R}^+$ and $x, v \in \mathbb{R}^n$, is given by

$$\partial_t f_\varepsilon + v \cdot \nabla_x f_\varepsilon + F(x) \cdot \nabla_v f_\varepsilon = \frac{1}{\varepsilon} \operatorname{div}_v((v - u_\varepsilon) f_\varepsilon + \nabla_v f_\varepsilon), \tag{4.30}$$

where

$$\rho_\varepsilon = \int_{\mathbb{R}^n} f_\varepsilon \, dv, \qquad \rho_\varepsilon u_\varepsilon = \int_{\mathbb{R}^n} v f_\varepsilon \, dv, \tag{4.31}$$

with the kinetic entropy

$$H(f,v) = \left(\frac{1}{2}|v|^2 + \ln f \right) f. \tag{4.32}$$

Here, we have $q(f) = F(x) \cdot \nabla_v f$, $a(v) = (1, v)$, and

$$\mathcal{Q}(f,v) = \operatorname{div}_v((v-u)f + \nabla_v f), \qquad (\rho, \rho u) = \int_{\mathbb{R}^n} a(v) f \, dv.$$

The property (1.10) is clear, and the property (1.13) comes from the following majorations: for $f \in L^1(\mathbb{R}^n)$ such that $\int_{\mathbb{R}^n} H(f,v)\, dv < \infty$, denoting $(\rho, \rho u) = \int_{\mathbb{R}^n} a(v) f(v)\, dv$, we have

$$\rho u^2 = \frac{(\int_{\mathbb{R}^n} v f \, dv)^2}{\int_{\mathbb{R}^n} f(v)\, dv} \le \int_{\mathbb{R}^n} |v|^2 f(v)\, dv,$$
$$\rho \ln \rho = \left(\int_{\mathbb{R}^n} f(v)\, dv \right) \ln \left(\int_{\mathbb{R}^n} f(v)\, dv \right) \le \int_{\mathbb{R}^n} f(v) \ln f(v)\, dv,$$

by Cauchy–Schwarz and by Jensen's inequality. As in the previous section we can check that

$$Q(\rho_\varepsilon, \rho_\varepsilon u_\varepsilon) = (0, F\rho_\varepsilon) = -\int_{\mathbb{R}^n} a(v) q(f_\varepsilon)\, dv,$$

so (1.18) is verified. It remains to verify (1.11) and (1.17).

4.2.2. Control of the entropy estimate and convergence result. We have the following estimate.

PROPOSITION 4.7. *Let f_ε be a solution to the kinetic equation* (4.5) *with initial value f_ε^0 bounded in $L^1(\mathbb{R}^{2n})$ verifying (finite energy)*

$$\iint_{\mathbb{R}^{2n}} H(f_\varepsilon^0(x,v),v)\,dv\,dx \le C^0 < \infty. \tag{4.33}$$

Then f_ε satisfies (1.11) *and*

$$\int_0^T \int_{\mathbb{R}^n} \left| A(U_\varepsilon) - \int_{\mathbb{R}^n} v \otimes a(v) f_\varepsilon\,dv \right| dx\,dt \le C\sqrt{\varepsilon}, \tag{4.34}$$

where A is the flux of the isothermal system.

Proof. We have

$$\partial_f H(f_\varepsilon, v) = \frac{|v|^2}{2} + 1 + \ln f_\varepsilon.$$

So

$$\begin{aligned}\int_{\mathbb{R}^n} \partial_f H(f_\varepsilon,v) F(x)\cdot\nabla_v f_\varepsilon\,dv &= -\rho_\varepsilon u_\varepsilon F(x)\\ &= -\eta'(\rho_\varepsilon,\rho_\varepsilon u_\varepsilon) Q(\rho_\varepsilon,\rho_\varepsilon u_\varepsilon)\end{aligned}$$

and

$$\begin{aligned}&\int_{\mathbb{R}^n} \partial_f H(f_\varepsilon,v)\,\mathrm{div}_v((v-u_\varepsilon)f_\varepsilon + \nabla_v f_\varepsilon)\,dv\\ &\quad= -\int_{\mathbb{R}^n} (v(v-u_\varepsilon)f_\varepsilon + v\nabla_v f_\varepsilon)\,dv - \int_{\mathbb{R}^n}\left(\frac{\nabla_v f_\varepsilon}{f_\varepsilon}(v-u_\varepsilon)f_\varepsilon + \frac{(\nabla_v f_\varepsilon)^2}{f_\varepsilon}\right) dv\\ &\quad= -\int_{\mathbb{R}^n} \frac{((v-u_\varepsilon)f_\varepsilon + \nabla_v f_\varepsilon)^2}{f_\varepsilon}\,dv - \int_{\mathbb{R}^n} u_\varepsilon(\nabla_v f_\varepsilon + (v-u_\varepsilon)f_\varepsilon)\,dv\\ &\quad= -\int_{\mathbb{R}^n} \frac{((v-u_\varepsilon)f_\varepsilon + \nabla_v f_\varepsilon)^2}{f_\varepsilon}\,dv.\end{aligned}$$

That is to say,

$$\begin{aligned}&\partial_t \int_{\mathbb{R}^n} H(f_\varepsilon,v)\,dv + \int_{\mathbb{R}^n} v\cdot\nabla_x H(f_\varepsilon,v)\,dv\\ &\quad= \int_{\mathbb{R}^n} \partial_f H(f_\varepsilon,v)(\partial_t f_\varepsilon + v\cdot\nabla_x f_\varepsilon)\,dv\\ &\quad= F(x)\rho_\varepsilon u_\varepsilon - \frac{1}{\varepsilon}\int_{\mathbb{R}^n} \frac{((v-u_\varepsilon)f_\varepsilon + \nabla_v f_\varepsilon)^2}{f_\varepsilon}\,dv.\end{aligned}$$

The first consequence of this relation is

$$\frac{d}{dt}\iint_{\mathbb{R}^{2n}} H(f_\varepsilon,v)\,dv\,dx \le \int_{\mathbb{R}^n} \eta'(\rho_\varepsilon,\rho_\varepsilon u_\varepsilon) Q(\rho_\varepsilon,\rho_\varepsilon u_\varepsilon)\,dx.$$

In particular this implies property (1.11). Moreover,

$$\begin{aligned}\frac{d}{dt}\iint_{\mathbb{R}^{2n}} H(f_\varepsilon, v)\,dv\,dx &\le F(x)\iint_{\mathbb{R}^{2n}} v f_\varepsilon\,dv\,dx\\ &\le \|F\|_{L^\infty}\iint_{\mathbb{R}^{2n}} (|v|^2+1) f_\varepsilon\,dv\,dx\\ &\le C_1\left(\|\rho_\varepsilon^0\|_{L^1} + \iint_{\mathbb{R}^{2n}} H(f_\varepsilon, v)\,dv\,dx\right),\end{aligned}$$

since the quantity $\iint |v|^2 f_\varepsilon\,dx\,dv$ is controlled in a classical way by $\iint H(f_\varepsilon, v)\,dx\,dv$. Using Gronwall's lemma, we deduce that there exists a constant C depending on T and f_ε^0 such that, for every $0 \le t \le T$,

$$\iint_{\mathbb{R}^{2n}} H(f_\varepsilon(t,x,v), v)\,dv\,dx \le C. \tag{4.35}$$

The second consequence is

$$\begin{aligned}&\left|\iiint_{[0,T]\times\mathbb{R}^{2n}} \frac{((v-u_\varepsilon)f_\varepsilon + \nabla_v f_\varepsilon)^2}{f_\varepsilon}\,dv\,dx\,dt\right|\\ &\quad\le \varepsilon\left(\iint_{\mathbb{R}^{2n}} H(f_\varepsilon^0)\,dv\,dx + \iint_{[0,T]\times\mathbb{R}^n} F(x)\rho_\varepsilon u_\varepsilon\,dx\,dt\right)\\ &\quad\le \varepsilon\left(C^0 + \|F\|_{L^\infty}\iiint_{[0,T]\times\mathbb{R}^{2n}} |v| f_\varepsilon\,dv\,dx\,dt\right)\\ &\quad\le \varepsilon\left(C^0 + \|F\|_{L^\infty}\iiint_{[0,T]\times\mathbb{R}^{2n}} (|v|^2+1) f_\varepsilon\,dv\,dx\,dt\right)\\ &\quad\le \varepsilon\left(C^0 + \|F\|_{L^\infty}\left(C' + T\iint_{[0,T]\times\mathbb{R}^n} f_\varepsilon^0\,dv\,dx\right)\right)\\ &\quad\le \varepsilon\, C_2.\end{aligned}$$

We have to estimate

$$\int_{\mathbb{R}^n}\left|A(U_\varepsilon) - \int_{\mathbb{R}^n} v\otimes a(v) f_\varepsilon\,dv\right|\,dx.$$

The first componant is zero. The second one is

$$E_2 = \int_{\mathbb{R}^n}\left|\rho_\varepsilon u_\varepsilon\otimes u_\varepsilon + \rho_\varepsilon I - \int_{\mathbb{R}^n} v\otimes v f_\varepsilon\,dv\right|\,dx,$$

which can be rewritten as

$$E_2 = \iint_{\mathbb{R}^{2n}} |v\otimes[(u_\varepsilon - v)f_\varepsilon - \nabla_v f_\varepsilon]|\,dv\,dx,$$

since

$$\int_{\mathbb{R}^n} v\otimes\nabla_v f_\varepsilon\,dv = -\int_{\mathbb{R}^n} f_\varepsilon\,dv I.$$

Thus we get

$$\int_0^T\!\!\int_{\mathbb{R}^{2n}} \left| A_2(U_\varepsilon) - \int_{\mathbb{R}^n} v \otimes a_2(v) f_\varepsilon \, dv \right| dx\, dt$$
$$\leq \left(\int_0^T\!\!\iint_{\mathbb{R}^{2n}} |v|^2 f_\varepsilon \, dv\, dx\, dt \right)^{1/2} \left(\int_0^T\!\!\iint_{\mathbb{R}^{2n}} \frac{((v-u_\varepsilon) f_\varepsilon + \nabla_v f_\varepsilon)^2}{f_\varepsilon} \, dv\, dx\, dt \right)^{1/2},$$

which concludes the proof. □

We can then apply the convergence result (Theorem 1.2). As in the previous section, this leads to Theorem 1.1 for the isothermal case.

Appendix. Euler. In this appendix, we calculate the various quantities which appear in our study in the case of the Euler system in order to see what prevents us from applying the method. For the full gas dynamics of Euler, the conservative variables are

$$U = (\rho, q, E) = \left(\rho, \rho u, \rho \frac{|u|^2}{2} + \frac{n}{2} \rho T \right),$$

and the flux is

$$A(U) = \left(\rho u, \rho u \otimes u + \rho T I, \rho u \frac{|u|^2}{2} + \frac{n+2}{2} \rho T u \right).$$

The entropy is

$$\eta(U) = \rho \ln \left(\frac{\rho}{(2\pi T)^{n/2}} \right) - \frac{n}{2} \rho,$$

and the associated flux is $G(U) = \eta(U) u$. The expression of the flux A in conservative variables is

$$A(U) = \left(q, \frac{1}{\rho} q \otimes q + \frac{2E}{n} I - \frac{1}{n\rho} |q|^2 I, \frac{n+2}{n} \frac{q}{\rho} E - \frac{1}{n} \frac{q}{\rho^2} |q|^2 \right).$$

Then we get

$$\begin{aligned}
\partial_\rho A_q(U) &= -u \otimes u + \frac{1}{n} |u|^2 I, \\
\partial_{q_i} (A_q)_{jk}(U) &= \delta_{ij} u_k + \delta_{ik} u_j - \delta_{jk} \frac{2u_i}{n}, \\
\partial_E A_q(U) &= \frac{2}{n} I, \\
\partial_\rho A_E(U) &= -\frac{n-2}{2} u \frac{|u|^2}{2} - \frac{n+2}{2} u T, \\
\partial_{q_i} (A_E)_j(U) &= \delta_{ij} \left(\frac{|u|^2}{2} + \frac{n+2}{2} T \right) - \frac{2}{n} u_i u_j, \\
\partial_E A_E(U) &= \frac{n+2}{n} u,
\end{aligned}$$

and the relative flux is

$$A_\rho(U_1|U_2) = 0, \tag{A.1}$$

$$A_q(U_1|U_2) = \rho_1(u_1-u_2)\otimes(u_1-u_2) - \frac{1}{n}\rho_1|u_1-u_2|^2 I, \tag{A.2}$$

$$A_E(U_1|U_2) = \frac{1}{2}\rho_1(|u_1|^2-|u_2|^2)(u_1-u_2) + \frac{n+2}{2}\rho_1(u_1-u_2)(T_1-T_2) - \frac{1}{n}\rho_1 u_2|u_1-u_2|^2. \tag{A.3}$$

We compute now the relative entropy. Since the linear part in a function disappeared in any relative quantity, we have to compute the flux of

$$\tilde\eta(U) = \left(1+\frac{n}{2}\right)\rho\ln\rho - \frac{n}{2}\rho\ln\left(\frac{2E}{n}-\frac{|q|^2}{n\rho}\right),$$

which satisfies

$$\partial_\rho\tilde\eta(U) = 1+\ln\rho+\frac{n}{2}-\frac{n}{2}\ln T - \frac{|u|^2}{2T}, \qquad \partial_q\tilde\eta = \frac{u}{T}, \qquad \partial_E\tilde\eta = -\frac{1}{T},$$

and thus we get

$$\eta(U_1|U_2) = h(\rho_1|\rho_2) + \frac{n\rho_1}{2T_2}h(T_2|T_1) + \frac{\rho_1}{2T_2}|u_1-u_2|^2, \tag{A.4}$$

where $h(x) = x\ln x$.

We see that we cannot apply our method in this case because of the cubic power in velocity in $A_E(U_1|U_2)$ since such a term does not appear in $\eta(U_1|U_2)$.

REFERENCES

[1] C. Bardos, F. Golse, and C. D. Levermore, *The acoustic limit for the Boltzmann equation*, Arch. Ration. Mech. Anal., 153 (2000), pp. 177–204.

[2] F. Berthelin and F. Bouchut, *Kinetic invariant domains and relaxation limit from a BGK model to isentropic gas dynamics*, Asymptot. Anal., 31 (2002), pp. 153–176.

[3] F. Berthelin and F. Bouchut, *Relaxation to isentropic gas dynamics for a BGK system with single kinetic entropy*, Methods Appl. Anal., 9 (2002), pp. 313–327.

[4] R. Botchorishvili, B. Perthame, and A. Vasseur, *Equilibrium schemes for scalar conservation laws with stiff sources*, Math. Comp., 72 (2003), pp. 131–157.

[5] F. Bouchut, *Construction of BGK models with a family of kinetic entropies for a given system of conservation laws*, J. Statist. Phys., 95 (1999), pp. 113–170.

[6] Y. Brenier, *Résolution d'équations d'évolution quasilinéaires en dimension N d'espace à l'aide d'équations linéaires en dimension N+1*, J. Differential Equations, 50 (1983), pp. 375–390.

[7] Y. Brenier, *Convergence of the Vlasov-Poisson system to the incompressible Euler equations*, Comm. Partial Differential Equations, 25 (2000), pp. 737–754.

[8] Y. Brenier, *Hydrodynamic structure of the augmented Born-Infeld equations*, Arch. Ration. Mech. Anal., 172 (2004), pp. 65–91.

[9] R. E. Caflisch, *The fluid dynamic limit of the nonlinear Boltzmann equation*, Comm. Pure Appl. Math., 33 (1980), pp. 651–666.

[10] C. M. Dafermos, *The second law of thermodynamics and stability*, Arch. Ration. Mech. Anal., 70 (1979), pp. 167–179.

[11] C. M. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Grundlehren Math. Wiss. 325, Springer-Verlag, Berlin, 2000.

[12] T. Gallouët, J.-M. Hérard, and N. Seguin, *Some approximate Godunov schemes to compute shallow-water equations with topography*, Comput. & Fluids, 32 (2003), pp. 479–513.

[13] Y. Giga and T. Miyakawa, *A kinetic construction of global solutions of first order quasilinear equations*, Duke Math. J., 50 (1983), pp. 505–515.

[14] F. Golse, C. D. Levermore, and L. Saint-Raymond, *La méthode de l'entropie relative pour les limites hydrodynamiques de modèles cinétiques*, in Séminaire: Équations aux Dérivées Partielles, 1999–2000, Sémin. Équ. Dériv. Partielles, Exp. No. XIX, 23, École Polytech., Palaiseau, 2000.

[15] L. Gosse, *A well-balanced scheme using non-conservative products designed for hyperbolic systems of conservation laws with source terms*, Math. Models Methods Appl. Sci., 11 (2001), pp. 339–365.

[16] L. Gosse, *Localization effects and measure source terms in numerical schemes for balance laws*, Math. Comp., 71 (2002), pp. 553–582.

[17] L. Gosse and A. E. Tzavaras, *Convergence of relaxation schemes to the equations of elastodynamics*, Math. Comp., 70 (2001), pp. 555–577.

[18] T. Goudon, P.-E. Jabin, and A. Vasseur, *Hydrodynamic limits for the Vlasov-Navier-Stokes equations. Part* II: *Fine particles regime*, Indiana Univ. Math. J., 53 (2004), pp. 1517–1536.

[19] S. Jin, *A steady-state capturing method for hyperbolic systems with geometrical source terms*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 631–645.

[20] S. Kaniel, *Approximation of the hydrodynamic equations by a transport process*, in Approximation Methods for Navier-Stokes Problems (Proc. Sympos., Univ. Paderborn, Paderborn, 1979), Lecture Notes in Math. 771, Springer-Verlag, Berlin, 1980, pp. 272–286.

[21] P.-L. Lions, *Mathematical Topics in Fluid Mechanics. Vol.* 1, *Incompressible Models*, Oxford Lecture Ser. Math. Appl. 3, The Clarendon Press, Oxford University Press, New York, 1996.

[22] P.-L. Lions and N. Masmoudi, *From the Boltzmann equations to the equations of incompressible fluid mechanics.* I, II, Arch. Ration. Mech. Anal., 158 (2001), pp. 173–193, 195–211.

[23] N. Masmoudi, *From Vlasov-Poisson system to the incompressible Euler system*, Comm. Partial Differential Equations, 26 (2001), pp. 1913–1928.

[24] B. Perthame, *Kinetic Formulation of Conservation Laws*, Oxford Lecture Ser. Math. Appl. 21, Oxford University Press, New York, 2002.

[25] B. Perthame and C. Simeoni, *A kinetic scheme for the Saint-Venant system with a source term*, Calcolo, 38 (2001), pp. 201–231.

[26] L. Saint-Raymond, *Convergence of solutions to the Boltzmann equation in the incompressible Euler limit*, Arch. Ration. Mech. Anal., 166 (2003), pp. 47–80.

[27] D. Serre, *Relaxation semi-linéaire et cinétique des systèmes de lois de conservation*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 17 (2000), pp. 169–192.

[28] A. E. Tzavaras, *Materials with internal variables and relaxation to conservation laws*, Arch. Ration. Mech. Anal., 146 (1999), pp. 129–155.

[29] H.-T. Yau, *Relative entropy and hydrodynamics of Ginzburg-Landau models*, Lett. Math. Phys., 22 (1991), pp. 63–80.

SIAM J. MATH. ANAL.
Vol. 36, No. 6, pp. 1836–1861

EXPANDING LORENZ ATTRACTORS THROUGH RESONANT DOUBLE HOMOCLINIC LOOPS*

C. A. MORALES†, M. J. PACIFICO†, AND B. SAN MARTIN‡

Abstract. In this paper we study the existence of Lorenz attractors in the unfolding of resonant double homoclinic loops in dimension three. Our results generalize the ones obtained in [C. Robinson, *SIAM J. Math. Anal.*, 32 (2000), pp. 119–141] in two ways. First, we obtain *attractors* instead of *weak attractors* obtained there. Second, we enlarge considerably the region in the parameter space corresponding to flows presenting expanding Lorenz attractors. The proof is based on rescaling techniques [J. Palis and F. Takens, *Hyperbolicity and Sensitive Choatic Dynamics at Homoclinic Bifurcations. Fractal Dimensions and Infinitely Many Attractors*, Cambridge University Press, Cambridge, UK, 1993] to obtain convergence to noncontinuous maps.

Key words. Lorenz attractors, resonant double homoclinic loops

AMS subject classifications. 37G10, 37G15, 37G25

DOI. 10.1137/S0036141002415785

1. Introduction. In the series of papers [Rob1, Rob2, Rob3] Robinson studied the existence of transitive attractors of Lorenz type in generic unfoldings of resonant double homoclinic loops in dimension three. For instance, [Rob1, Theorem 3.1, p. 130] says that under certain conditions such unfoldings produce transitive weak attractors containing the singularity. In this paper we improve this result in two ways. First, we obtain expanding Lorenz attractors instead of weak attractors as defined in [Rob1, p. 120]. Second, we enlarge considerably the region in the parameter space which corresponds to flows presenting expanding Lorenz attractors. By *attractor* we mean a transitive set which is maximal invariant in a positively invariant open set. A set is *transitive* if it is the omega-limit set of one of its orbits. An attractor of a three-dimensional vector field is *expanding Lorenz* if it contains a unique singularity whose eigenvalues $\lambda_1, \lambda_2, \lambda_3$ are real and satisfy $\lambda_2 < \lambda_3 < 0 < -\lambda_3 < \lambda_1$. The classical example of an expanding Lorenz attractor is the geometric Lorenz attractor in [GW, ABS]. We shall consider parametrized families of vector fields unfolding a resonant double homoclinic loop at $\eta = \eta_0$ as in [Rob1]. We assume the same hypotheses (A1)–(A7) of [Rob1], except for (A5) that we replace with

$$B = \frac{C^+_{\eta_0} + C^-_{\eta_0}}{C^+_{\eta_0} C^-_{\eta_0}} > 1,$$

where $C^\pm_{\eta_0}$ are defined as in that paper: the constants $C^\pm_{\eta_0}$ measure the change in area within a certain bundle over Γ, the resonant double homoclinic loop. Note that (A5) in [Rob1] implies $B > 1$ but not conversely. The proof is based on rescaling techniques [PT] to obtain convergence to noncontinuous maps. In a forthcoming paper we use

*Received by the editors October 10, 2002; accepted for publication (in revised form) October 3, 2003; published electronically June 14, 2005. This work was partially supported by CNPq, FAPERJ, PRONEX on Dynamical Systems, FONDECYT grants 1981241 and 100047, and FUNDACION ANDES.

http://www.siam.org/journals/sima/36-6/41578.html

†Instituto de Matemática, Universidade Federal do Rio de Janeiro, C. P. 68.530, CEP 21.945-970, Rio de Janeiro, R. J., Brazil (morales@im.ufrj.br, pacifico@im.ufrj.br).

‡Departamento de Mathemáticas, Universidad Católica del Norte, Casilla 1280, Antofagasta, Chile (sanmarti@socompa.ucn.cl).

the rescaling techniques developed here to study the existence of contracting Lorenz attractors (Rovella attractors) in the unfolding of resonant double homoclinic loops. This will answer a question posed in [Rob1, Remark 5.1, p. 138].

Before we announce in a precise way our results, let us comment on some other results concerning Lorenz attractors and its bifurcations.

1.1. Related results and comments.

1.1.1. Lorenz equations and the geometrical model. The main motivation for all these results is the Lorenz attractor [Lo], given by the solutions of the polynomial vector field in $\mathbb{R}^3$:

$$X(x,y,z) = \begin{cases} \dot{x} = -\alpha x + \alpha y, \\ \dot{y} = \beta x - y - xz, \\ \dot{z} = -\gamma z + xy, \end{cases} \tag{1.1}$$

where α, β, γ are real parameters. Numerical experiments performed by Lorenz (for $\alpha = 10, \beta = 28,$ and $\gamma = 8/3$) suggested the existence of a strange attractor toward which tends a full neighborhood of positive trajectories of the above system. Moreover, the strange attractor seemed to be robust: it cannot be destroyed by any perturbation of the parameters. On the other hand, this attractor contains an equilibrium point $(0,0,0)$, and periodic points accumulating on it, and hence cannot be hyperbolic. The book [Sp] contains an extensive presentation of analytical and numerical facts about the original Lorenz equations. We point out that it was proved [Tu1, Tu2] that the solutions of (1.1) satisfy such a property for values α, β, γ near the ones considered by Lorenz. See the feature review [Tu2, V] for a survey and a discussion of Tucker's proof.

However, already in the mid-seventies, the existence of robust nonhyperbolic attractors was proved for flows introduced in [ABS, GW], which we now call geometrical models for Lorenz attractors. At the same time the theory of uniformly hyperbolic systems was being developed, and it was increasingly clear that such systems are not dense. One of those examples was precisely the geometrical Lorenz attractor.

Although geometrical Lorenz attractors fail to be uniformly hyperbolic, they do exhibit a certain amount of hyperbolicity (called singular hyperbolicity in [MPP2, MPP3]), which has been exploited to understand their structure. The bifurcations of these attractors and their relations to homoclinic bifurcations of codimensions 2 and 3 have also been intensively studied by several authors, both theoretically and with a view toward applications. This included great interest in the so-called Lorenz maps, one-dimensional models to which the dynamics of a geometrical Lorenz attractor can be reduced. In addition a number of bifurcation mechanisms have been found which yield attractors with similar properties. In what follows we give an overview of the literature, without trying to be exhaustive.

1.1.2. Attractors that resemble a Lorenz attractor. First, let us list some papers about the existence of chaotic attractors that resemble geometrical Lorenz models. We start with [MPP2, MPP3] where it is proved that any robust attractor of a flow in three manifolds containing equilibria looks like a geometric Lorenz attractor. In [BPV] the authors construct a multidimensional Lorenz-like attractor that is C^r-robust (r large) and contains a singularity with at least two positive eigenvalues. They also investigate the Sinai–Bowen–Ruelle (SBR) measures of these attractors. Their construction works in dimension greater than or equal to 5. In [ST] the authors present an example of a four-dimensional quasi-attractor and study its perturbations. The

quasi-attractor is pseudohyperbolic, contains a singularity with a complex eigenvalue, and cannot be destroyed by small perturbations of the system. In [Lo84] the author reports a careful numerical study of what seems to be a strange (chaotic) attractor in four dimensions for a system of 2-degree polynomial equations. In [Rov] the author proves existence and persistence of contracting Lorenz attractors, that is, with the contracting eigenvalue condition $-\lambda_3 > \lambda_1$.

In [PRV] the authors prove that certain parametrized families of one-dimensional maps with infinitely many critical points exhibit global chaotic behavior in a persistent way. An application of the methods developed there yields a proof of existence and even persistence of global spiral attractors for smooth flows in three dimensions, to be given in [CPRV].

In [P, S] the authors propose abstract models for attractors with singularities, called generalized hyperbolic attractors, and study their properties.

1.1.3. Topological dynamics. Some aspects of the topological dynamics of the geometric model was studied in [Ko1, Ko2], where it was proved that most geometrical Lorenz attractors do not have the shadowing property, and their expansive properties are investigated. In [Kl] the author finds a topological invariant for the Lorenz attractor allowing him to exhibit an uncountable number of nonhomeomorphic Lorenz attractors in the unfolding of a certain homoclinic loop. In [Ya] the author shows that the geometrical Lorenz attractor can be approximated by horseshoes with entropy close to that of the Lorenz attractor. In [BW] the knot type of the geometric model is analyzed, and in [GH] the Lorenz attractor is used to investigate the existence of flows realizing all links and knots as periodic orbits in 3-manifolds and an explicit ODE with such properties is exhibited. One can also see the survey [PS].

1.1.4. Dimension theory, ergodic and statistical properties. Concerning fractal dimensions of Lorenz attractors we mention the results in [Le1, Le2] and [BL]. The first contains an explicit formula for the Liapunov dimension of the Lorenz attractor and in the second a simple upper bound on the Hausdorff dimension of Lorenz attractors is given in terms of parameters σ, α, β in (1.1).

Statistical and ergodic properties of the geometrical model were investigated in [Bu]. In [Me1, Me2] the existence of SBR measures and the stochastic stability for the contracting model are proved.

1.1.5. Lorenz maps of the interval. Lorenz-like maps of the interval and their bifurcations have been studied in [LM1, LM2]. See also the references therein. In [AL] the authors describe the use of kneading theory to study the dynamics of Lorenz maps. In [HS] the Lorenz maps are classified, up to topological conjugacy, by their kneading invariants. In [MdeM] the notion of monotone Lorenz families is introduced, and the authors prove that all possible topological dynamics behavior of Lorenz maps is realized within every monotone Lorenz family. In [KS] the authors study the case where the Lorenz map has negative Schwarzian derivative and the derivative vanishes at both sides of the discontinuity point. The paper has both a survey and a research flavor. Among other results they characterize the global attractor associated in terms of renormalization properties of the Lorenz map. A modified Lorenz interval map is studied in [LV], combining a finite number of critical points and a finite number of discontinuities with infinite derivative. The authors prove that the associated attractor is nonuniformly hyperbolic. In [DY] the authors study the asymptotic periodicity of a Lorenz interval map. They prove that a Lorenz-like map is asymptotically periodic if the derivative set of the pre-image set of the discontinuity point is countable. See

also the book [AH] and the references therein for more about this topic.

1.1.6. Bifurcations. An extensive series of works have been dedicated to the bifurcations of systems leading to the appearance of Lorenz attractors, contracting or not, besides the papers by Robinson quoted before. We start with the Shimizu–Morioka model, a three-dimensional system of first-order ODEs which depend on two parameters. It has been derived to study the Lorenz system for large Rayleigh number. In [SM] the authors proved that there are two types of Lorenz-like attractors in this model. In [Sh], bifurcations near two codimension-two points in the parameter plane are analyzed in great detail. These points lie on the boundary of the region of existence of the Lorenz-like attractors. In [MPP1] a bifurcation of hyperbolic vector fields on three-dimensional manifolds leading to robust strange attractors with singularities is studied. In [DKO] the authors prove that bifurcations of a certain double homoclinic loop associated with a degenerated singularity can produce geometrical Lorenz attractors, and they also exhibit an explicit ODE presenting such attractors. [Ry] gives sufficient conditions for a butterfly inclination-flip loop to generate geometrical Lorenz attractors. In [ACL] the authors investigate the existence of Lorenz attractors in the unfolding of a certain singular cycle involving a saddle-node periodic orbit and a Lorenz-like singularity. In [MPu] the authors show the appearance of Lorenz attractors in the unfolding of a cycle formed by a Lorenz-like singularity and a saddle-node periodic orbit, and in [Mo] the appearance of Lorenz attractors through a saddle-node bifurcation is shown. See also [KKO] and the references therein for more about bifurcations generating geometrical Lorenz attractors.

1.2. Our hypotheses. Let us state our hypotheses in a precise way. In what follows X_η is a family of C^r, $r \geq 1$, vector fields on $\mathbb{R}^3$ satisfying the following conditions.

(A1) *For every η, X_η has a hyperbolic singularity Q_η such that the eigenvalues of $DX_\eta(Q_\eta)$ are real with $\lambda_{ss}(\eta) < \lambda_s(\eta) < 0 < \lambda_u(\eta)$, and with eigenvectors v^{ss}, v^s, and v^u, respectively.*

With this assumption, there are several invariant manifolds for the singularity Q_η. We denote the one-dimensional unstable manifold tangent to v^u by $W^u(Q_\eta, \eta)$, and the two-dimensional stable manifold tangent to v^{ss} and v^s by $W^s(Q_\eta, \eta)$. Next, there is a one-dimensional strong stable manifold $W^{ss}(Q_\eta, \eta)$. This latter manifold is made of points which converge to Q_η at an asymptotic rate determined by the eigenvalue λ_{ss}. All these manifolds are C^r if the vector field is C^r. Finally, there is a two-dimensional extended stable manifold tangent to v^s and v^u, which we denote by $W^{cu}(Q_\eta, \eta)$. The later manifold is at least C^1. With this notation we can make the second assumption about the existence of a double homoclinic connection.

(A2) *For the bifurcation value η_0, there is a double homoclinic connection with the unstable manifold of Q_{η_0} contained in the stable manifold but outside the strong stable manifold, that is,*

$$\Gamma = W^u(Q_{\eta_0}, \eta_0) \subset W^s(Q_{\eta_0}, \eta_0) \setminus W^{ss}(Q_{\eta_0}, \eta_0).$$

In fact, we assume that the two branches $\Gamma^\pm$ of $\Gamma \setminus \{Q_{\eta_0}\}$ are contained in the same component of $W^s(Q_{\eta_0}, \eta_0) \setminus W^{ss}(Q_{\eta_0}, \eta_0)$. Note that $\Gamma = \{Q_{\eta_0}\} \cup \Gamma^+ \cup \Gamma^-$.

(A3) *For η_0, the central manifold $W^{cu}(Q_{\eta_0}, \eta_0)$ is transverse to the stable manifold $W^s(Q_{\eta_0}, \eta_0)$ along Γ.*

Let

$$P(q) = T_q W^{cu}(Q_{\eta_0}, \eta_0) \quad \text{for } q \in \Gamma.$$

The transversality condition in (A3) with the condition

$$W^u(Q_{\eta_0}, \eta_0) \cap W^{ss}(Q_{\eta_0}, \eta_0) = Q_{\eta_0}$$

in assumption (A2) implies that $P(q)$ converges to $P(Q_{\eta_0})$ as q converges to Q_{η_0} along Γ (by the inclination lemma [dMP]). Therefore, $\{P(q) : q \in \Gamma\}$ is a continuous bundle over Γ. Considering one half of the homoclinic connection $\Gamma^+ \cup Q_{\eta_0}$, let $\nu^+ = 1$ if the bundle $\{P(q) : q \in \Gamma^+ \cup Q_{\eta_0}\}$ is orientable and $\nu^+ = -1$ if the bundle is nonorientable. In the same way, considering the other half of the homoclinic connection $\Gamma^- \cup Q_{\eta_0}$, let $\nu^- = \pm 1$ whenever the bundle $\{P(q) : q \in \Gamma^- \cup Q_{\eta_0}\}$ is orientable or nonorientable, respectively.

(A4) *We assume that*

$$\lambda_{ss}(\eta_0) - \lambda_s(\eta_0) + \lambda_u(\eta_0) < 0 \quad \textit{and} \quad \lambda_{ss}(\eta_0) < 2\lambda_s(\eta_0).$$

We shall use the notation $\alpha(\eta) = -\frac{\lambda_s(\eta)}{\lambda_u(\eta)}$ and $\beta(\eta) = -\frac{\lambda_{ss}(\eta)}{\lambda_u(\eta)}$.

These are open conditions and so do not add a codimension to the bifurcation. The second inequality in (A4) assures that $W^{cu}(Q_{\eta_0}, \eta_0)$ is C^2.

Let $q^{\pm}(t)$ be a parametrization of the solution along $\Gamma^{\pm}$ and $\operatorname{div}_2(q^{\pm}(t))$ the Jacobian of X_{η_0} at t restricted to $T_{\Gamma^{\pm}} W^{cu}$. Define $C^{\pm}_{\eta_0}$ by

$$C^{\pm}_{\eta_0} = \exp\left(\int_{-\infty}^{\infty} \operatorname{div}_2(q^{\pm}(t)) dt \right).$$

The quantity $C^{\pm}_{\eta_0}$ is the change in area within the planes $P(q)$ along the whole length of $\Gamma^{\pm}$.

(A5) $B > 1$, *where*

$$B = \frac{C^+_{\eta_0} + C^-_{\eta_0}}{C^+_{\eta_0} C^-_{\eta_0}}.$$

We observe that condition (A5) in [Rob1] requires either $0 < C^{\pm}_{\eta_0} < 1$ or $0 < C^{\pm}_{\eta_0} < 2$ and $\frac{C^+_{\eta_0}}{C^-_{\eta_0}} \in [(1+\sqrt{2})^{-1}, 1+\sqrt{2}]$. Robinson also has a polynomial vector field realizing such conditions.

We, on the contrary, do not make such assumptions on $C^{\pm}_{\eta_0}$. See Figure 1.1. On the other hand, we are still working on the problem of finding a polynomial vector field realizing the unfolding described in this paper.

Note that (A5) in [Rob1] implies $B > 1$ but not conversely. Assumption (A5) is open.

(A6) *There is a one-to-one resonance between the unstable and weak stable eigenvalue for η_0:*

$$\lambda_u(\eta_0) + \lambda_s(\eta_0) = 0.$$

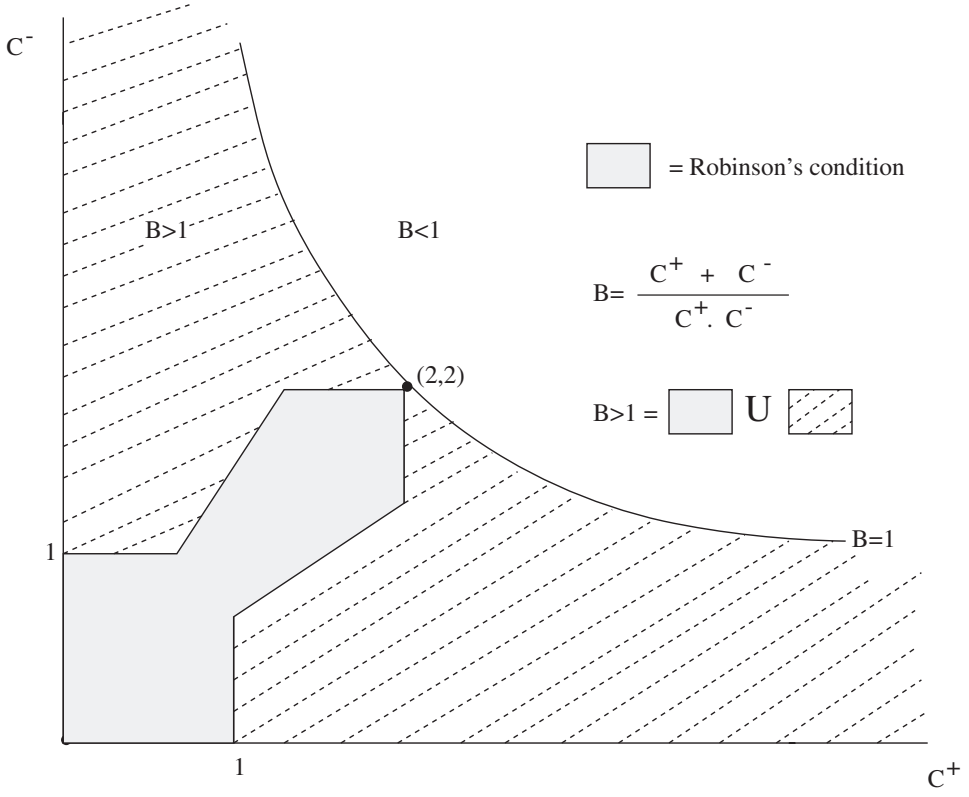

FIG. 1.1.

Observe that condition (A6) means $\alpha(\eta) = 1$. This condition is needed to have (A5) satisfied; see [Rob1]. This resonance condition is a codimension-one condition; in total, the conditions on η_0 are codimension three. (Two conditions are from the double homoclinic connection and resonance gives the third and final codimension.)

The final assumption is related to the unfolding of the bifurcation. We assume that the parameter space is large enough in order to break the double homoclinic loop in a correct way.

(A7) *Let $\mathcal{N} \subset \mathcal{X}^1(\mathbb{R}^3)$ be the 3-submanifold defined by conditions* (A1)–(A6). *We assume that the family $\{X_\eta\}$ is transverse to $\mathcal{N}$ at η_0.*

1.3. The main result. It is now possible to announce our main result. Given an A, a subset in the parameter space, we set $Cl(A)$ for the closure of A.

THEOREM 1.1. *Let $\{X_\eta\}$ be a C^k-parametrized family of C^r-vector fields (where $r, k \geq 3$) satisfying* (A1)–(A7). *Then, there is an open set $\mathcal{O}$ in the parameter space with $\eta_0 \in Cl(\mathcal{O})$ such that X_η has an expanding Lorenz attractor for all $\eta \in \mathcal{O}$.*

The tools used in the proof are reduction of the dynamics to a one-dimensional Poincaré map and the existence of a suitable rescaling for such maps with a well-defined limit dynamics. Rescaling techniques in the unfolding of homoclinic loops were used in [N] to obtain convergence to the Henon map. Here we use such techniques to obtain convergence to noncontinuous maps. Convergence to noncontinuous maps via rescaling was already considered in [MPu] for the study of certain heteroclinic

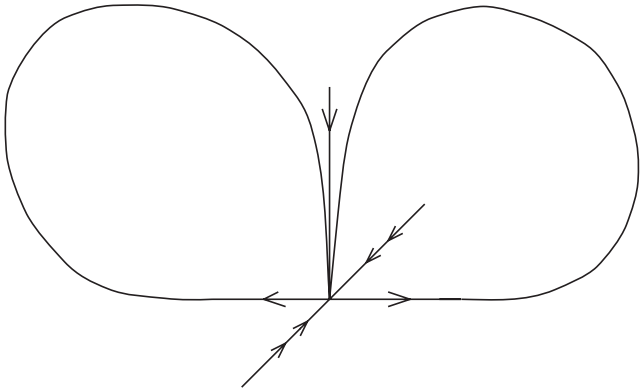

FIG. 1.2. *Vector field* X_{η_0}.

connections involving saddle-node periodic orbits. See [MSV] for more about rescaling techniques for singular cycles. Figure 1.2 displays a double homoclinic loop for the vector field X_{η_0}.

1.4. Sketch of the proof. Let us present the idea of the proofs. As in [Rob1], we observe that (A1)–(A3) imply the existence of a strong stable invariant foliation close to the loop. By (A4), the C^r-section theorem [S] implies that such a foliation is C^1 and varies C^1 with the parameters. As usual we consider the Poincaré map along the homoclinic loop. Using the strong stable foliation we reduce the dynamics of the return map to a one-dimensional map $f_\eta(\tau)$. Following [Rob2] we denote by α the order of $f_\eta(\tau)$. Clearly α depends on η. At this point the proof in [Rob1] requires solving certain inequalities allowing trapping regions for suitable parameter values. Recall that a compact interval J is a trapping region for f_η if $f_\eta(J) \subset \text{int}(J)$, where $\text{int}(J)$ means the interior of J. Instead, we use a different approach in which we consider α as a parameter. In Lemma 2.2 we fix α and prove the existence of *good parameters values*, i.e., parameters for which the critical values $f_\eta(0^\pm)$ of f_η are either fixed or pre-fixed or periodic (with period 2) expanding points. Such parameters are solutions of certain equations that can be solved only for $\alpha < 1$ because of (A5). Using the critical values we construct, for those good parameters, an f_η-invariant closed interval $[p, q]$ containing $\tau = 0$. We also get $|f'_\eta(\tau)| > 1$ uniformly for $\tau \in [p, q]$. Afterward we use rescaling techniques [PT]: we take a suitable parameter-depending change of coordinates in a neighborhood of $[p, q]$ and, at the same time, we normalize the parameter space in a small neighborhood of those good parameters. This yields a new family $g_\alpha(\mu, \nu, \cdot)$ and new good parameters $(\mu(\alpha), \nu(\alpha))$. In Lemma 3.3 we show the existence of bounds for the derivative of $g_\alpha(\nu, \mu, \cdot)$. This is used to prove that $g_\alpha(\mu, \nu, \cdot)$ converges (in a C^1-sense to be defined below) to a map $g(\mu, \nu, \cdot)$ as $\alpha \to 1^-$. The limit map $g(\mu, \nu, \cdot)$ is piecewise linear expanding that looks like the one in Figure 3.1 in section 3. In the same lemma we show that $\lim_{\alpha \to 1^\pm}(\mu(\alpha), \nu(\alpha)) = (\mu(1), \nu(1))$ exists.

The maps $g(\mu, \nu, \cdot)$ above do not have trapping regions, but they can be approximated in our family by ones having them. Indeed, by Theorem 4.1, we have that $g_\alpha(\mu, \nu, \cdot)$ has trapping regions for $\alpha < 1$ close to 1 and for (μ, ν) close to $(\mu(\alpha), \nu(\alpha))$. That is, for each $\alpha < 1$ close to 1 there is a compact interval J_α such that $g_\alpha(\mu, \nu, J_\alpha) \subset \text{int}(J_\alpha)$. Then for all $\alpha < 1$ and close to 1 the flow associated

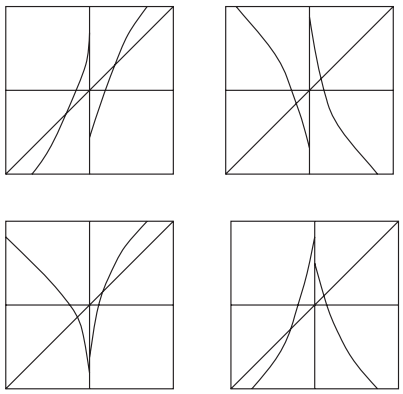

FIG. 2.1. *Possible* f_η.

with $g_\alpha(\mu,\nu,\cdot)$ has a trapping region. Thus a trapping region for the flow does exist in an open set $\mathcal{O}$ of parameters accumulating $\eta = 0$. To prove Theorem 1.1 we use the closeness of $g_\alpha(\mu,\nu,\cdot)$ to $g(\mu(1),\nu(1),\cdot)$. Such a closeness and Theorem 5.1 imply both transitiveness and expansiveness of $g_\alpha(\mu,\nu,\cdot)$ for the parameters with trapping regions; see also [Rob4, Theorem A].

2. One-dimensional reduction and good parameters. To start with, consider a cross section Σ of X_{η_0} close to Q_{η_0} transversal $W^s(Q_{\eta_0})$ intersecting both branches of $W^u(Q_{\eta_0})$. There is a neighborhood V of $\Sigma \cap W^s(Q_{\eta_0})$ in Σ such that the positive orbit of every point at $V \setminus W^s(Q_{\eta_0})$ intersects Σ for every parameter η near enough to η_0, defining in this way a Poincaré map $F_\eta : V \setminus W^s(Q_{\eta_0}) \subset \Sigma \to \Sigma$.

As X_{η_0} satisfies conditions (A1)–(A4), the standard stable manifold theory applies to show the existence of a C^1 stable foliation in a small neighborhood (that for convenience we assume equals to V) of $W^s(Q_\eta)$ varying C^1 with the parameter. As in [Rob2] the existence of a C^r stable foliation ($r \geq 1$) depends on the relation

$$C_3 e^{T(\lambda_{ss}(\eta_0)-\lambda_s(\eta_0))}(e^{T\lambda_u(\eta_0)})^r < 1\,.$$

By the first eigenvalue inequality in (A4) we have the above relation for $r = 1$. Then, we can use the C^r-section theorem (see [S, Theorem 5.18]) in the same way as in [Rob2]. Thus, via projection along the leaves of the strong stable foliation, the problem can be reduced to a one-dimensional Poincaré map $f_\eta : V' \setminus \{c_\eta\} \subset [-1,1] \to [-1,1]$. Here c_η is the projection of $W^s(Q_{\eta_0}) \cap V$ onto V'. We assume $c_\eta = 0$ for every η. Denote $a_\eta^\pm = \lim_{\tau\to\pm 0} f_\eta(\tau)$, $\tau \in [-1,1]$. Recall the description of the coefficients $\nu^\pm$ in assumption (A3).

LEMMA 2.1. *There is an interval J, $0 \in J$, such that for every η sufficiently near to η_0, the map $f_\eta : J \subset [-1,1] \to [-1,1]$ has the following form:*

$$f_\eta(\tau) = \begin{cases} a_\eta^+ + \nu^+ C_\eta^+ |\tau|^{\alpha_\eta} + O_{\eta,1}(|\tau|^{\alpha_\eta}) & \text{if } \tau > 0, \\ a_\eta^- - \nu^- C_\eta^- |\tau|^{\alpha_\eta} + O_{\eta,2}(|\tau|^{\alpha_\eta}) & \text{if } \tau < 0, \end{cases}$$

where $O_{\eta,i}$ are C^1, varying C^1 with respect to η, and $\lim_{x\to 0} \frac{O_{\eta,i}(x)}{x} = 0$ uniformly on η. Moreover, $C_\eta^\pm$ depends C^1 on η.

This result was proved in [Rob2]. See Figure 2.1.

From now on we assume that X_η is a three-parameter family for which there is an open set $U \subset \mathbb{R}^3$ such that for $\eta \in U$, X_η satisfies conditions (A1)–(A7).

It follows from (A7) that the map $\eta \mapsto (\alpha_\eta, a^+_\eta, a^-_\eta)$ is a diffeomorphism from a neighborhood of η_0 onto a neighborhood of $(1,0,0)$. So, we can reparametrize η by $(\alpha, a^+, a^-) \mapsto \eta(\alpha, a^+, a^-)$ in such a way that $\alpha_{\eta(\alpha,a^+,a^-)} = \alpha$, $a^+_{\eta(\alpha,a^+,a^-)} = a^+$ and $a^-_{\eta(\alpha,a^+,a^-)} = a^-$.

LEMMA 2.2. *There are $\Lambda > 0$, an open dense set $O \subset (1-\Lambda, 1)$, and C^1 maps $a^+(\cdot), a^-(\cdot), p(\cdot), q(\cdot) : O \to \mathbb{R}$ such that $p(\alpha) < 0 < q(\alpha)$ and for $\eta = \eta(\alpha, a^+(\alpha), a^-(\alpha))$ the following hold:*

(a) *if $\nu^+ = \nu^- = 1$, then $f_\eta(p(\alpha)) = p(\alpha)$, $f_\eta(q(\alpha)) = q(\alpha)$, $f_\eta(0^+) = p(\alpha)$, and $f_\eta(0^-) = q(\alpha)$;*

(b) *if $\nu^+ = \nu^- = -1$, then $f_\eta(p(\alpha)) = q(\alpha)$, $f_\eta(q(\alpha)) = p(\alpha)$, $f_\eta(0^+) = q(\alpha)$, and $f_\eta(0^-) = p(\alpha)$;*

(c) *if $\nu^+ = -\nu^- = 1$, then $f_\eta(p(\alpha)) = q(\alpha)$, $f_\eta(q(\alpha)) = q(\alpha)$, and $f_\eta(0^+) = f_\eta(0^-) = p(\alpha)$;*

(d) *if $\nu^+ = -\nu^- = -1$, then $f_\eta(p(\alpha)) = p(\alpha)$, $f_\eta(q(\alpha)) = p(\alpha)$, and $f_\eta(0^+) = f_\eta(0^-) = q(\alpha)$.*

In any case, $\lim_{\alpha\to 1^-} |p(\alpha)|/q(\alpha) = C^+_{\eta_0}/C^-_{\eta_0}$, $\lim_{\alpha\to 1^-} q(\alpha) = 0 = \lim_{\alpha\to 1} p(\alpha)$, and $\lim_{\alpha\to 1^-} q(\alpha)^{\alpha-1} = \lim_{\alpha\to 1^-} |p(\alpha)|^{\alpha-1} = B$, where B is the number given in (A5).

Proof. We prove (a). Define, for $\alpha < 1$, $p < 0$, and $q > 0$, the functions h, $\tilde{h}$, ℓ, and $\tilde{\ell}$ by

$$h(\alpha,p,q) = q^\alpha \left(\frac{1 + \left(\frac{C^+_\eta q^\alpha + O_{\eta,1}(|q|^\alpha)}{C^-_\eta |p|^\alpha - O_{\eta,2}(|p|^\alpha)} \frac{|p|^\alpha}{|q|^\alpha} \right)^{\frac{1}{\alpha}}}{C^+_\eta q^\alpha + O_{\eta,1}(|q|^\alpha)} \right),$$

$$\tilde{h}(\alpha,p,q) = |p|^\alpha \left(\frac{1 + \left(\frac{C^-_\eta |p|^\alpha - O_{\eta,2}(|p|^\alpha)}{C^+_\eta |q|^\alpha + O_{\eta,1}(|q|^\alpha)} \frac{|q|^\alpha}{|p|^\alpha} \right)^{\frac{1}{\alpha}}}{C^-_\eta |p|^\alpha - O_{\eta,2}(|p|^\alpha)} \right),$$

$$\ell(\alpha,p,q) = q^{\alpha-1}, \quad \tilde{\ell}(\alpha,p,q) = |p|^{\alpha-1},$$

where $\eta = \eta(\alpha,p,q)$.

Observe that h and $\tilde{h}$ are differentiable except for $D = \{(1,p,0), (1,0,q), \forall p,q\}$, and they extend continuously to D. Indeed,

$$\lim_{(\alpha,p,q)\to(1,0,0)} h(\alpha,p,q) = B = \lim_{(\alpha,p,q)\to(1,0,0)} \tilde{h}(\alpha,p,q),$$

and so we define $h(1,0,0) = B = \tilde{h}(1,0,0)$. Furthermore,

$$h(\alpha,p,q) \left(\frac{C^+_\eta q^\alpha + O_{\eta,1}(|q|^\alpha)}{q^\alpha} \right)^{\frac{\alpha-1}{\alpha}} = \tilde{h}(\alpha,p,q) \left(\frac{C^-_\eta |p|^\alpha - O_{\eta,2}(|p|^\alpha)}{|p|^\alpha} \right)^{\frac{\alpha-1}{\alpha}}. \tag{2.1}$$

Let us choose $\Lambda > 0$. As $B > 1$, by (A5) there is $\epsilon > 0$ such that $1 \notin [B-\epsilon, B+\epsilon]$. Fix δ_0 and Λ such that if $q, |p| \in (0, \delta_0)$ and $\alpha \in (1-\Lambda, 1)$, then

$$B - \epsilon < h(\alpha,p,q), \tilde{h}(\alpha,p,q) < B + \epsilon\,. \tag{2.2}$$

Shrinking Λ if necessary, we can further assume $\delta_0^{\alpha-1} < B - \epsilon$.

For the proof of Lemma 2.2(a) we shall use the following result which states that for all $\alpha \in (1-\Lambda, 1)$, the projection on the plane $p\,q$ of $graph(h(\alpha,\cdot,\cdot))\cap graph(\ell(\alpha,\cdot,\cdot))$ contains a regular curve C_α joining the vertical sides of $[-\delta_0, 0] \times [0, \delta_0]$.

Let $\mathcal{C}^1([0,1])$ be the set of C^1 maps $\gamma : [0,1] \to [-\delta_0, 0] \times [0, \delta_0]$ endowed with the C^1 topology.

LEMMA 2.3. *There are an open dense set $O \subset (1-\Lambda, 1)$ and a C^1 map*

$$\Gamma : O \times [0,1] \mapsto \mathcal{C}^1([0,1]), \quad (\alpha, t) \mapsto C_\alpha(t),$$

where

$$C_\alpha(t) = (p_\alpha(t), q_\alpha(t))$$

is such that $p_\alpha(0) = 0$, $p_\alpha(1) = -\delta_0$, $h(\alpha, p_\alpha(t), q_\alpha(t)) = [q_\alpha(t)]^{\alpha-1}$, and $0 < q_\alpha(t) < \delta_0$ for all $t \in (0,1)$.

Proof. We first prove that for all α small enough, $graph(h(\alpha,\cdot,\cdot))\cap graph(\ell(\alpha,\cdot,\cdot))$ is transversal. For this we proceed as follows.

By the definition of $h(\alpha, p, q)$ above we have

$$h(\alpha,p,q) = \frac{1 + \left(\frac{C_\eta^+ + O_{\eta,1}(|q|^\alpha)/q^\alpha}{C_\eta^- - O_{\eta,2}(|p|^\alpha)} / \mid p \mid^\alpha\right)^{\frac{1}{\alpha}}}{C_\eta^+ q^\alpha + O_{\eta,1}(|q|^\alpha)}.$$

Recall that $\eta = \eta(\alpha, p, q)$.

Assume that $O_{\eta,i} = 0$ for $i = 1, 2$. Taking the derivative of h with respect to q we obtain

$$\partial_q h(\alpha,p,q) = \frac{(1/\alpha)(C_\eta^+/C_\eta^-)^{1/\alpha-1} \cdot (\partial(C_\eta^+) \cdot C_\eta^- - C_\eta^+ \partial_q(C_\eta^-)) \cdot (C_\eta^-)^{-2} \cdot C_\eta^+}{(C_\eta^+)^2}.$$

From this we have that $\lim_{(\alpha,p,q)\to(1,0,0)} \partial_q h(\alpha,p,q)$ exists, and so $\partial_q h(\alpha,p,q)$ has a lower bound $K > -\infty$. By the definition of $\ell(\alpha,p,q)$ one has

$$\partial_q \ell(\alpha,p,q) = (\alpha-1) q^{(\alpha-2)}.$$

If (α,p,q) satisfies $h(\alpha,p,q) = l(\alpha,p,q)$, then the definition of $\ell(\alpha,\cdot,\cdot)$ with (2.2) imply that

$$q^{\alpha-1} > B - \epsilon.$$

So, for such points (α,p,q) one has

$$\partial_q h(\alpha,p,q) - \partial_q l(\alpha,p,q) > K - (\alpha-1) q^{\alpha-2} > K + \frac{\ln(B-\epsilon)\cdot(B-\epsilon)}{|\ln(q)|\cdot q}. \tag{2.3}$$

Because $B - \epsilon > 1$ it follows that the right-hand term of (2.3) goes to ∞ as $q \to 0^+$. In particular, the left-hand term of (2.3) is nonzero and so $graph(h(\alpha,\cdot,\cdot)) \cap graph(\ell(\alpha,\cdot,\cdot))$ is transversal.

The general case follows from similar computations using that for (α,p,q) such that $h(\alpha,p,q) = l(\alpha,p,q)$ we have $B - \epsilon < q^{(\alpha-1)} < B + \epsilon$ and

$$\lim_{(\alpha,p,q)\to(1,0,0)} \frac{O_{\eta,1}(q^\alpha)}{q^\alpha} = \lim_{(\alpha,p,q)\to(1,0,0)} \partial_q \left(\frac{O_{\eta,1}(q^\alpha)}{q^\alpha}\right) = 0.$$

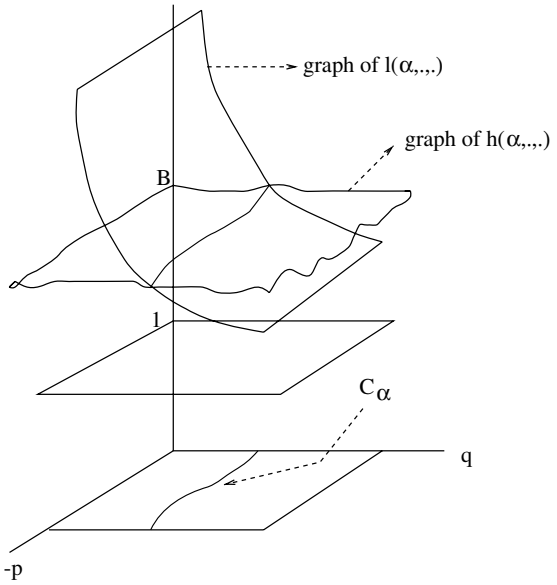

FIG. 2.2. *graph* $(h(\alpha,\cdot,\cdot)) \cap$ *graph* $(l(\alpha,\cdot,\cdot))$.

Returning to the proof of Lemma 2.3, since $graph(h(\alpha,\cdot,\cdot)) \cap graph(\ell(\alpha,\cdot,\cdot))$ is transversal, we have that this intersection is a compact 1-manifold; see Figure 2.2. Although $graph(h(\alpha,\cdot,\cdot)) \cap graph(\ell(\alpha,\cdot,\cdot))$ may be nonconnected, the number of connected components of this intersection is finite. On the other hand, it follows from the intermediate value theorem that $graph(\ell(\alpha,\cdot,\cdot))$ intersects $h(\alpha,\ p(q),q)$ for any curve $q \to p(q)$ with $p(-\delta) = \delta$, $p(0) = 0$. Then, for all $\alpha \in (1-\Lambda, 1)$, the projection onto the plane $p\,q$ of one of the connected components of $graph(h(\alpha,\cdot,\cdot)) \cap graph(\ell(\alpha,\cdot,\cdot))$ is a regular curve C_α joining the vertical sides of $[-\delta_0, 0] \times [0, \delta_0]$. That is,

$$C_\alpha : [0,1] \to [-\delta_0, 0] \times [0, \delta_0], \quad t \mapsto (p_\alpha(t), q_\alpha(t))$$

such that $p_\alpha(0) = 0$, $p_\alpha(1) = -\delta_0$, $h(\alpha, p_\alpha(t), q_\alpha(t)) = [q_\alpha(t)]^{\alpha-1}$, and $0 < q_\alpha(t) < \delta_0$ for all $t \in (0,1)$.

Now observe that by usual transversality arguments, for all α' there is $I_{\alpha'}$ such that for all $\alpha \in I_{\alpha'}$ the curve C_α obtained above can be chosen in such way that the map $(\alpha, t) \in I_{\alpha'} \times [0,1] \mapsto C_\alpha(t) \in [-\delta_0, 0] \times [0, \delta_0]$ is C^1. Hence we can choose an open dense set O in $(1-\Lambda, 1)$ such that the map $(\alpha, t) \in O \times [0,1] \mapsto C_\alpha(t)$ is C^1. All of this together concludes the proof of Lemma 2.3. □

Returning to the proof of Lemma 2.2, observe that we have the following inequalities:

$$\lim_{t\to 0} |p_\alpha(t)|^{\alpha-1} > B + \epsilon \geq \tilde{h}(\alpha, p_\alpha(t), q_\alpha(t)) \geq B - \epsilon > \delta_0^{\alpha-1} = |p_\alpha(1)|^{\alpha-1}.$$

These inequalities and the intermediate value theorem imply that there is $t = t(\alpha)$ such that

$$\tilde{h}(\alpha, p_\alpha(t), q_\alpha(t)) = |p_\alpha(t)|^{\alpha-1}.$$

By Lemma 2.3, $C_\alpha(t) = (p_\alpha(t), q_\alpha(t))$ depends C^1 on (α, t) and so, by the implicit function theorem we get that the map $\alpha \mapsto t(\alpha)$ is C^1.

From now on O is the open dense set given by Lemma 2.3.

Let $a^+(\cdot)$, $a^-(\cdot)$, $p(\cdot)$, $q(\cdot) : O \to \mathbb{R}$ be given by $p(\alpha) = p_\alpha(t_0(\alpha))$, $q(\alpha) = q_\alpha(t_0(\alpha))$, $a^+(\alpha) = p(\alpha)$, and $a^-(\alpha) = q(\alpha)$. These maps are C^1 and satisfy

$$h(\alpha, p(\alpha), q(\alpha)) = (q(\alpha))^{\alpha-1}, \quad \tilde{h}(\alpha, p(\alpha), q(\alpha)) = |p(\alpha)|^{\alpha-1}. \tag{2.4}$$

From this it follows that $(B-\epsilon)^{\frac{1}{\alpha-1}} < q(\alpha) < (B+\epsilon)^{\frac{1}{\alpha-1}}$, and hence we obtain $\lim_{\alpha\to 1^-} q(\alpha) = 0$. Similarly for $p(\alpha)$. Applying these facts in (2.4) we get $\lim_{\alpha\to 1^-} q(\alpha)^{\alpha-1} = B$ and $\lim_{\alpha\to 1^-} |p(\alpha)|^{\alpha-1} = B$.

We claim that $p = p(\alpha)$, $q = q(\alpha)$, $a^+ = a^+(\alpha)$, and $a^- = a^-(\alpha)$ chosen as above satisfy (a) in Lemma 2.2.

Indeed, since $h(\alpha, p, q) = q^{\alpha-1}$ and $\tilde{h}(\alpha, p, q) = |p|^{\alpha-1}$, (2.1) implies that

$$q^{\alpha-1}\left(\frac{C_\eta^+ q^\alpha + O_{\eta,1}(q^\alpha)}{q^\alpha}\right)^{\frac{\alpha-1}{\alpha}} = |p|^{\alpha-1}\left(\frac{C_\eta^- |p|^\alpha - O_{\eta,2}(|p|^\alpha)}{|p|^\alpha}\right)^{\frac{\alpha-1}{\alpha}}.$$

Hence

$$C_\eta^+ q^\alpha + O_{\eta,1}(q^\alpha) = C_\eta^- |p|^\alpha - O_{\eta,2}(|p|^\alpha). \tag{2.5}$$

On the other hand, from the definition of h we get

$$q^{\alpha-1} = q^\alpha \left(\frac{1 + \left(\frac{C_\eta^+ q^\alpha + O_{\eta,1}(q^\alpha)}{C_\eta^- |p|^\alpha - O_{\eta,2}(|p|^\alpha)} \frac{|p|^\alpha}{q^\alpha}\right)^{\frac{1}{\alpha}}}{C_\eta^+ q^\alpha + O_{\eta,1}(q^\alpha)}\right). \tag{2.6}$$

From (2.5) and (2.6) we get $C_\eta^+ q^\alpha + O_{\eta,1}(q^\alpha) = q - p$, implying that $f_\eta(q) = q$ and $f_\eta(p) = p$.

Note that (2.5) implies

$$q^\alpha\left(C_\eta^+ + \frac{O_{\eta,1}(q^\alpha)}{q^\alpha}\right) = |p|^\alpha\left(C_\eta^- - \frac{O_{\eta,2}(|p|^\alpha)}{|p|^\alpha}\right)$$

and so $\lim_{\alpha\to 1^-} \frac{|p(\alpha)|}{q(\alpha)} = \frac{C_{\eta_0}^+}{C_{\eta_0}^-}$.

All of this together finishes the proof of Lemma 2.2(a). The remaining cases are similar and are left to the reader. □

Note 1. Let O be as in Lemma 2.2 and for $\alpha \in O$, let $a^+(\alpha)$ and $a^-(\alpha)$ be the functions defined above. We shall call $\eta = \eta(\alpha, a^+(\alpha), a^-(\alpha))$ for $\alpha \in O$ the *good parameters* of X_η.

3. Rescaling. In this section we perform rescaling techniques [PT]. Keeping the notation p, q in Lemma 2.2 we take suitable parameter-depending change of coordinates in a neighborhood of $[p, q]$ and, at the same time, we normalize the parameter space in a small neighborhood of the good parameters in Notation 2. This yields a new family $g_\alpha(\mu, \nu, \cdot)$ and new good parameters $(\mu(\alpha), \nu(\alpha))$. The goal of this section is to prove that $g_\alpha(\mu, \nu, \cdot)$ converges to a map $g(\mu, \nu, \cdot)$ in the sense to be described below.

To start we consider the parametrized family $\{X_\eta\}$ as in section 2. Recall the notation in that section.

Given $\alpha \in O$ let $a^-(\alpha)$, $a^+(\alpha)$, $p(\alpha)$, and $q(\alpha)$ be as in Lemma 2.2. Define $(\mu(\alpha), \nu(\alpha)) = (\frac{a^+(\alpha)}{q(\alpha)}, \frac{a^-(\alpha)}{q(\alpha)})$ and $(\mu, \nu) = (\frac{a^+}{q(\alpha)}, \frac{a^-}{q(\alpha)})$ in a neighborhood of $(a^+(\alpha), a^-(\alpha))$ onto a neighborhood of $(\mu(\alpha), \nu(\alpha))$, and the family of maps

$$g_\alpha(\mu, \nu, x) = \frac{1}{q(\alpha)} f_\eta(q(\alpha)x), \tag{3.1}$$

where $\eta = \eta(\alpha, q(\alpha)\mu, q(\alpha)\nu)$. We set $\mathrm{Dom}(g_\alpha)$ for the domain of g_α.

Note 2. Observe that for each fixed α, this change of variables renormalizes the parameters $a^\pm$ with $a^+ = q(\alpha)\mu$ and $a^- = q(\alpha)\nu$. Moreover, by Lemma 2.2, $\lim_{\alpha\to 1^-} \mu(\alpha)$ and $\lim_{\alpha\to 1^-} \nu(\alpha)$ exist and we denote them by $\mu(1)$ and $\nu(1)$, respectively.

DEFINITION 3.1. *Let $g : \mathbb{R}^2 \times (\mathbb{R} \setminus \{0\}) \to \mathbb{R}$. We say that $g_\alpha \to g$ in the C^0 topology in compact sets of $\mathbb{R}^3$ as $\alpha \to 1^-$ if*

(a) $\mathrm{Dom}(g_\alpha) \to \mathbb{R}^2 \times (\mathbb{R} \setminus \{0\})$ *as $\alpha \to 1^-$, that is, for all $R > 0$ there is $0 < \alpha_0 < 1$ such that if $\alpha_0 < \alpha < 1$, then $B_R(0) \cap (\mathbb{R}^2 \times (\mathbb{R} \setminus \{0\}) \subset \mathrm{Dom}(g_\alpha)$, where $B_R(0)$ is the ball of radius R centered at $(0, 0, 0)$,*

(b) *for every compact set $K \subset \mathbb{R}^3$ and every $\epsilon > 0$ there is $\delta > 0$ such that if $|\alpha - 1| < \delta$, then*

$$\sup_{y \in K \cap (\mathbb{R}^2 \times (\mathbb{R}\setminus\{0\}))} |g_\alpha(y) - g(y)| < \epsilon.$$

DEFINITION 3.2. *Let $g : \mathbb{R}^2 \times (\mathbb{R} \setminus \{0\}) \to \mathbb{R}$. We say that $g_\alpha \to g$ in the C^1 topology in compact sets of $\mathbb{R}^2 \times (\mathbb{R} \setminus \{0\})$ if*

(a) $\mathrm{Dom}(g_\alpha) \to \mathbb{R}^2 \times (\mathbb{R} \setminus \{0\})$ *as $\alpha \to 1^-$,*

(b) *for every compact set $K \subset \mathbb{R}^2 \times (\mathbb{R} \setminus \{0\})$ and every $\epsilon > 0$ there is $\delta > 0$ such that if $|\alpha - 1| < \delta$, then*

$$\sup_{i \in \{0,1\}, y \in K} |D^i g_\alpha(y) - D^i g(y)| < \epsilon.$$

Note that with these notions, C^1 convergence does not imply C^0 convergence.

We have the following result.

LEMMA 3.3. *Let g_α be as in* (3.1) *and define*

$$g(\mu, \nu, x) = \begin{cases} \mu + \nu^+ C^+_{\eta_0} Bx & \text{if } x > 0, \\ \nu + \nu^- C^-_{\eta_0} Bx & \text{if } x < 0, \end{cases}$$

where $\eta_0 = \eta(1, 0, 0)$. Then

(i) *$g_\alpha \to g$ in the C^0 topology in compact sets of $\mathbb{R}^3$ as $\alpha \to 1^-$, $\alpha \in O$,*

(ii) *$g_\alpha \to g$ in the C^1 topology in compact sets of $\mathbb{R}^2 \times (\mathbb{R} \setminus \{0\})$ as $\alpha \to 1^-$, $\alpha \in O$.*

Moreover, for any $c > \max\{1, C^+_{\eta_0}/C^-_{\eta_0}\}$, there are constants $\Delta_0 > 0$, $0 < K_1 < K_2$ such that for $\alpha \in O \cap [1 - \Delta_0, 1]$ we have

(a) $[-c, c]^2 \times ([-c, c] \setminus \{0\}) \subset \mathrm{Dom}(g_\alpha)$,

(b) $K_1|x|^{\alpha-1} \le \left|\frac{\partial}{\partial x} g_\alpha(\mu, \nu; x)\right| \le K_2|x|^{\alpha-1}$ *for all* $(\mu, \nu, x) \in [-c, c]^2 \times ([-c, c] \setminus \{0\})$.

Proof. First, let us prove that $\mathrm{Dom}(g_\alpha) \to \mathbb{R}^2 \times (\mathbb{R} \setminus \{0\})$. For this, fix $R > 0$. We shall prove that

$$B_R(0) \cap (\mathbb{R}^2 \times (\mathbb{R} \setminus \{0\})) \subset \mathrm{Dom}(g_\alpha)$$

for all α close to 1.

For simplicity set $k = q(\alpha)$, where $q(\alpha)$ is given by Lemma 2.3. Recall $k \to 0$ as $\alpha \to 1^-$.

Let $0 < \epsilon_0$, ϵ_1 so that $(1-\epsilon_1, 1+\epsilon_1) \times (-\epsilon_0, \epsilon_0) \times (-\epsilon_0, \epsilon_0) \subset \mathrm{Dom}(\eta)$.

Given $(\mu, \nu, x) \in B_R(0)$ and α with $|\alpha - 1| < \epsilon_1$, we have $|k\mu|, |k\nu| \le kR < \epsilon_0$ for ϵ_1 small enough. Thus, $(\alpha, k\mu, k\nu) \in \mathrm{Dom}(\eta)$. In a similar way we have that $kx \in \mathrm{Dom} f_{\eta(\alpha,k\mu,k\nu)}$. This proves (a) of Definitions 3.1 and 3.2.

To verify (b) we proceed as follows. First observe that, by definition, see Lemma 2.1, we have

$$g_\alpha(\mu,\nu,x) = k^{-1} f_\eta(kx) = \begin{cases} k^{-1}[k\mu + \nu^+ C_\eta^+ |kx|^\alpha + O_{\eta,1}(|k_\alpha x|^\alpha)], & x > 0, \\ k^{-1}[k\nu - \nu^- C_\eta^- |kx|^\alpha + O_{\eta,2}(|kx|^\alpha)], & x < 0 \end{cases}$$

$$= \begin{cases} \mu + \nu^+ C_\eta^+ k^{\alpha-1}|x|^\alpha + \frac{O_{\eta,1}(|kx|^\alpha)}{k}, & x > 0, \\ \nu - \nu^- C_\eta^- k^{\alpha-1}|x|^\alpha + \frac{O_{\eta,2}(|kx|^\alpha)}{k}, & x < 0. \end{cases} \tag{3.2}$$

As $k^{\alpha-1} \to B$, $k \to 0$, and $\eta \to \eta_0$ when $\alpha \to 1$, we obtain $\nu^\pm C_\eta^\pm k^{\alpha-1}|x|^\alpha \to \nu^\pm C_{\eta_0}^\pm B|x|$ as $\alpha \to 1^-$. On the other hand, by Lemma 2.1 we have $\lim_{x\to 0} \frac{O_{\eta,i}(x)}{|x|} = 0$ uniformly on η. So, $\frac{O_{\eta,i}(|kx|^\alpha)}{k} = \frac{\mathcal{O}_{\eta,i}(|kx|^\alpha)}{|kx|^\alpha} k^{\alpha-1}|x|^\alpha \to 0$ as $\alpha \to 1^-$ for $|x| < R$. All of this together imply (b) of Definition 3.1. So we finish the proof of (i).

To prove (ii) we proceed as follows. We have, for $x \neq 0$,

$$\partial_x g_\alpha(\mu,\nu,x) = \begin{cases} \nu^+ C_\eta^+ \alpha k^{\alpha-1}|x|^{\alpha-1} + O'_{\eta,1}(|kx|^\alpha)\alpha|kx|^{\alpha-1}, & x > 0, \\ \nu^- C_\eta^- \alpha k^{\alpha-1}|x|^{\alpha-1} + O'_{\eta,2}(|kx|^\alpha)\alpha|kx|^{\alpha-1}, & x < 0. \end{cases} \tag{3.3}$$

Now fix a compact set $K \subset \mathbb{R}^2 \times (\mathbb{R} \setminus \{0\})$. Then there are $0 < C_1 = C_1(K) < C_2 = C_2(K)$ such that for all $(\mu, \nu, x) \in K$ we have $C_1 < |x| < C_2$, implying that $|x|^{\alpha-1} \to 1$ as $\alpha \to 1^-$. On the other hand, by Lemma 2.1, $O'_{\eta,i}(0) = 0$ for $i = 1, 2$. Since $k \to 0$ as $\alpha \to 1^-$, $C_1 < |x| < C_2$, we get $kx \to 0$ as $\alpha \to 1$. So, $O'_{\eta,i}(|kx|^\alpha)\alpha|x|^{\alpha-1} \to 0$ as $\alpha \to 1$. Thus $\partial_x g_\alpha(\mu,\nu,x) \to \partial_x g(\mu,\nu,x)$ as $\alpha \to 1^-$ in compact sets of $\mathbb{R}^2 \times (\mathbb{R} \setminus \{0\})$.

On the other hand, since $k = q(\alpha)$, we have $a^+ = k\mu$ and $a^- = k\nu$, see Note 2, and so

$$\partial_\mu g_\alpha(\mu,\nu,x) = \begin{cases} 1 + \nu^+ \partial_{a^+} C_\eta^+ k^\alpha |x|^\alpha + \partial_{a^+} \mathcal{O}_{\eta,1}(|kx|^\alpha), & x > 0, \\ -\nu^- \partial_{a^+} C_\eta^- k^\alpha |x|^\alpha + \partial_{a^+} \mathcal{O}_{\eta,2}(|kx|^\alpha), & x < 0, \end{cases} \tag{3.4}$$

$$\partial_\nu g_\alpha(\mu,\nu,x) = \begin{cases} \nu^+ \partial_{a^-} C_\eta^+ k^\alpha |x|^\alpha + \partial_{a^-} \mathcal{O}_{\eta,1}(|kx|^\alpha), & x > 0, \\ 1 - \nu^- \partial_{a^-} C_\eta^- k^\alpha |x|^\alpha + \partial_{a^-} \mathcal{O}_{\eta,2}(|kx|^\alpha), & x < 0. \end{cases} \tag{3.5}$$

As $C_\eta^\pm$ is C^1, we obtain that $|\partial_{a^+} C_\eta^\pm|$ is uniformly bounded on K. Since $k^{\alpha-1} \to B$ and $k \to 0$ as $\alpha \to 1^-$ we get $k^\alpha |x|^\alpha = k^{\alpha-1} k|x| \to 0$ as $\alpha \to 1^-$ on K. Moreover, $\mathcal{O}_{\eta,i}(0) = 0$ and so $\partial_{a^+} \mathcal{O}_{\eta,i}(|kx|^\alpha) \to 0$ as $\alpha \to 1^-$. Replacing these bounds on (3.4) and (3.5), respectively, we obtain that $\partial_\mu g_\alpha(\mu,\nu,x) \to \partial_\mu g(\mu,\nu,x)$ and $\partial_\nu g_\alpha(\mu,\nu,x) \to \partial_\nu g(\mu,\nu,x)$ for $(\mu,\nu,x) \in K$. This finishes the proof. □

Note 3. Observe that $|g'| = |C_{\eta_0}^\pm B|$. Since $|\nu^\pm C_{\eta_0}^\pm B| > 1$ we obtain that g is an expanding map. Figure 3.1 displays the possible graphics for the map g.

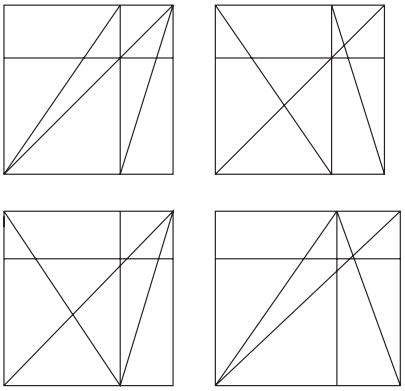

FIG. 3.1. *Possible* $g(\mu(1),\nu(1),.)$.

4. Trapping region. The main goal of this section is to prove Theorem 4.1, which asserts the existence of trapping regions for $g_\alpha(\mu,\nu,\cdot)$, for $\alpha \in O$ close to 1 and (μ,ν) close to $(\mu(\alpha),\nu(\alpha))$ (recall the notation in section 3). A trapping region for $g_\alpha(\mu,\nu,.)$ is a closed interval J such that $g_\alpha(\mu,\nu,J) \subset \mathrm{Int}(J)$, where $\mathrm{Int}(J)$ stands for the interior of J. To prove the existence of such a trapping region we define an auxiliary function $F_\alpha : \mathbb{R}^2 \to \mathbb{R}^2$, which will be used to find them. This function involves the critical and the fixed points of $g_\alpha(\mu,\nu,\cdot)$. We shall prove that F_α is locally one-to-one, i.e., $\det(DF_\alpha(\mu(\alpha),\nu(\alpha))) \neq 0$. Once we have that, the parameters corresponding to maps with trapping regions will be the ones in the inverse image $F_\alpha^{-1}(W_\alpha)$ for certain open set W_α to be defined below, proving Theorem 4.1.

THEOREM 4.1. *There is an open set* $\mathcal{O}^- \subset \mathbb{R}^3$ *such that the following properties hold:*

1. $(\alpha,\mu(\alpha),\nu(\alpha)) \in Cl(\mathcal{O}^-)$ *for all* $\alpha \in O$*;*
2. *if* $\theta = (\alpha,\mu,\nu) \in \mathcal{O}^-$*, then there is a closed interval* $I_\theta \subset \mathbb{R}$ *with* $0 \in \mathrm{Int}\,(I_\theta)$ *such that* $g_\alpha(\mu,\nu,x) \subset \mathrm{Int}\,(I_\theta)$ *for all* $x \in I_\theta$*.*

To prove the theorem we need some terminology. Let O be the set given in the last section. For $\alpha \in O$ let $\mu(\alpha) = \frac{a^+(\alpha)}{q(\alpha)}$, $\nu(\alpha) = \frac{a^-(\alpha)}{q(\alpha)}$, $x(\alpha) = \frac{p(\alpha)}{q(\alpha)}$, and $y(\alpha) = \frac{q(\alpha)}{q(\alpha)} = 1$. Fix $c > \max\{1, C^+_{\eta_0}/C^-_{\eta_0}\}$. Recall that Lemma 2.2 gives us $\lim_{\alpha\to 1} \frac{|p(\alpha|}{q(\alpha)} = \frac{C^+_{\eta_0}}{C^-_{\eta_0}}$. Again, by Lemma 2.2, we have that if $\nu^+ = \nu^- = 1$, then $a^+(\alpha) = p(\alpha)$, $a^-(\alpha) = q(\alpha)$ and so $\mu(\alpha) = \frac{p(\alpha)}{q(\alpha)}$, $\nu(\alpha) = 1$. Then $\lim_{\alpha\to 1} \mu(\alpha) = \lim_{\alpha\to 1} x(\alpha) = \frac{-C^+_{\eta_0}}{C^-_{\eta_0}}$. This implies that $(\mu(\alpha),\nu(\alpha),x(\alpha))$ and $(\mu(\alpha),\nu(\alpha),1)$ belong to $[-c,c]^2 \times ([-c,c] \setminus \{0\})$ and we conclude, by Lemma 3.3(a), that $(\mu(\alpha),\nu(\alpha),x(\alpha))$ and $(\mu(\alpha),\nu(\alpha),1) \in \mathrm{Dom}(g_\alpha)$ for all α close to 1.

In a similar way, for the remaining cases $\nu^+ = \nu^- = -1$, $\nu^+ = -\nu^- = 1$, and $\nu^+ = -\nu^- = -1$, we also obtain $(\mu(\alpha),\nu(\alpha),x(\alpha))$ and $(\mu(\alpha),\nu(\alpha),1) \in \mathrm{Dom}(g_\alpha)$ for all α close to 1.

From now on we assume that $\nu^+ = \nu^- = 1$. In this case

$$g_\alpha(\mu(\alpha),\nu(\alpha),x(\alpha)) = x(\alpha), \quad g_\alpha(\mu(\alpha),\nu(\alpha),1) = 1,$$

$$\partial_x g_\alpha(\mu(\alpha),\nu(\alpha),x(\alpha)) > 1, \quad \partial_x g_\alpha(\mu(\alpha),\nu(\alpha),1) > 1.$$

Next, set $x(\alpha,\mu,\nu)$ and $y(\alpha,\mu,\nu)$ for the C^1 continuations of $x(\alpha)$ and $y(\alpha)$ for (μ,ν) close to $(\mu(\alpha),\nu(\alpha))$ in the (μ,ν) parameter space, i.e., $g_\alpha(x(\alpha,\mu,\nu)) = x(\alpha,\mu,\nu)$

and $g_\alpha(y(\alpha,\mu,\nu)) = y(\alpha,\mu,\nu)$. Observe that $x(\alpha,\mu(\alpha),\nu(\alpha)) = x(\alpha)$ and $y(\alpha,\mu(\alpha),\nu(\alpha)) = y(\alpha) = 1$ for all $\alpha \in O$. As before, set $k = q(\alpha)$.

Define

$$F_\alpha(\mu,\nu) = (\mu - x(\alpha,\mu,\nu), \nu - y(\alpha,\mu,\nu)).$$

We shall prove the following theorem.

THEOREM 4.2. $\det DF_\alpha(u(\alpha),\nu(\alpha)) \neq 0$ *for all* $\alpha \in O$ *close to* 1.

Assuming Theorem 4.2, let us show how to obtain Theorem 4.1.

Observe that Theorem 4.2 implies that F_α is a local diffeomorphism in a neighborhood V_α of $(\mu(\alpha),\nu(\alpha))$ onto a neighborhood of $F_\alpha(\mu(\alpha),\nu(\alpha)) = (0,0)$. Let $W_\alpha = F_\alpha(V_\alpha) \cap A$ where $A = \{(x,y), x > 0, y < 0\}$.

Consider $\theta = (\alpha,\mu,\nu)$ such that $\alpha \in O$ and $(\mu,\nu) \in F_\alpha^{-1}(W_\alpha)$.

We claim that there is a closed interval I_θ satisfying (2) in Theorem 4.1. Indeed, we have the following inequalities:

$$x(\alpha,\mu,\nu) < \mu \quad \text{and} \quad \nu < y(\alpha,\mu,\nu), \text{ since } (\mu,\nu) \in F_\alpha^{-1}(W_\alpha). \tag{4.1}$$

Moreover, by (3.2),

$$\nu = g_\alpha(\mu,\nu,0^-) \quad \text{and} \quad \mu = g_\alpha(\mu,\nu,0^+).$$

Furthermore, since $\nu^+ = \nu^- = 1$, we have that

$$g_\alpha(\mu,\nu,.) \text{ is monotonic in } [x(\alpha,\mu,\nu),0) \text{ and in } (0,y(\alpha,\mu,\nu)], \tag{4.2}$$

and

$$\partial_x g_\alpha(\mu,\nu,\sigma) > 1, \quad \sigma \in \{x(\alpha,\mu,\nu), y(\alpha,\mu,\nu)\}. \tag{4.3}$$

Now, (4.1) and (4.3) imply that there is $\epsilon > 0$ (depending on μ and ν), small such that

$$\nu < y(\alpha,\mu,\nu) - \epsilon \quad \text{and} \quad g_\alpha(\mu,\nu,x(\alpha,\mu,\nu)+\epsilon) > x(\alpha,\mu,\nu) + \epsilon, \tag{4.4}$$

$$x(\alpha,\mu,\nu) + \epsilon < \mu \quad \text{and} \quad g_\alpha(\mu,\nu,y(\alpha,\mu,\nu)-\epsilon) < y(\alpha,\mu,\nu) - \epsilon. \tag{4.5}$$

Define $I_\theta^-(\epsilon) = [x(\alpha,\mu,\nu)+\epsilon, 0)$, $I_\theta^+(\epsilon) = (0, y(\alpha,\mu,\nu)-\epsilon]$, and $I_\theta(\epsilon) = I_\theta^- \cup I_\theta^+$.

On one hand, (4.4) together with (4.2) imply that $g_\alpha(\mu,\nu,I_\theta^-(\epsilon)) \subset \mathrm{Int}(I_\theta(\epsilon))$. On the other hand, (4.5) together with (4.2) imply $g_\alpha(\mu,\nu,I_\theta^+(\epsilon)) \subset \mathrm{Int}(I_\theta(\epsilon))$. Thus, $g_\alpha(\mu,\nu,I_\theta(\epsilon)) \subset \mathrm{Int}(I_\theta(\epsilon))$, proving the claim with $I_\theta = I_\theta(\epsilon)$.

The proof for the remaining cases is similar.

As the existence of a trapping region is an open condition, we conclude that there is $\delta_\theta > 0$ such that if $\|\zeta - \theta\| < \delta_\theta$, then $I_\zeta = I_\theta$ is such that $g_\alpha(I_\zeta) \subset \mathrm{Int}(I_\zeta)$.

Now define

$$\mathcal{O}_0^- = \{\theta = (\alpha,\mu,\nu); \alpha \in O,\ (\mu,\nu) \in F_\alpha^{-1}(W_\alpha)\}$$

and

$$\mathcal{O}^- = \cup_{\theta \in \mathcal{O}_0^-} B(\theta,\delta_\theta),$$

where $B(\theta,\delta_\theta)$ is the ball centered at θ with radius δ_θ. Clearly, $\mathcal{O}^-$ satisfies the required properties. This finishes the proof of Theorem 4.1. □

To prove Theorem 4.2 we shall use the following lemmas.

LEMMA 4.3.
(a) $\lim_{\alpha\to 1^-} \frac{q(\alpha)}{\alpha-1} = \lim_{\alpha\to 1^-} \frac{p(\alpha)}{\alpha-1} = 0.$
(b) $\lim_{\alpha\to 1^-} \frac{q(\alpha)^\alpha}{\alpha-1} = \lim_{\alpha\to 1^-} \frac{|p(\alpha)|^\alpha}{\alpha-1} = 0.$

Proof. (a) As part of the proof of the existence of $q(\alpha)$ and $p(\alpha)$, we obtain that

$$B-\varepsilon < q(\alpha)^{\alpha-1} < B+\varepsilon\,,$$

and from this it follows that

$$\frac{(B-\varepsilon)^{\frac{1}{\alpha-1}}}{|\alpha-1|} < \frac{q(\alpha)}{|\alpha-1|} < \frac{(B+\varepsilon)^{\frac{1}{\alpha-1}}}{|\alpha-1|}\,.$$

Now recall that $q(\alpha)$ and $p(\alpha)$ are defined for $\alpha < 1$. Moreover, since $B-\epsilon > 1$, we get

$$\lim_{\alpha\to 1^-} \frac{(B-\varepsilon)^{\frac{1}{\alpha-1}}}{|\alpha-1|} = \lim_{\alpha\to 1^-} \frac{(B+\varepsilon)^{\frac{1}{\alpha-1}}}{|\alpha-1|} = 0\,.$$

So $\lim_{\alpha\to 1^-} \frac{q(\alpha)}{|\alpha-1|} = 0\,.$

The same argument implies $\lim_{\alpha\to 1^-} \frac{p(\alpha)}{|\alpha-1|} = 0.$

(b) $\lim_{\alpha\to 1^-} \frac{q(\alpha)^\alpha}{|\alpha-1|} = \lim_{\alpha\to 1^-} \frac{q(\alpha)}{|\alpha-1|}\cdot q(\alpha)^{\alpha-1} = 0\cdot B = 0.$ □

LEMMA 4.4. *Set $\eta = \eta(\alpha, p(\alpha), q(\alpha))$. Then*

$$\lim_{\alpha\to 1^-} \frac{1}{\alpha-1}\left[C_\eta^+ C_\eta^- \cdot |p(\alpha)|^{\alpha-1}\cdot q(\alpha)^{\alpha-1} - C_\eta^-\ |p(\alpha)|^{\alpha-1} - C_\eta^+\ q(\alpha)^{\alpha-1}\right] = 0.$$

Proof. Equation (2.6) in Lemma 2.2 implies

$$q(\alpha)^{\alpha-1} = \frac{1+\sqrt[\alpha]{\frac{C_\eta^+ + \frac{O_{n,1}(q(\alpha)^\alpha)}{q(\alpha)^\alpha}}{C_\eta^- - \frac{O_{n,2}(|p(\alpha)|^\alpha)}{|p(\alpha)|^\alpha}}}}{C_\eta^+ + \frac{O_{n,1}(q(\alpha)^\alpha)}{q(\alpha)^\alpha}}.$$

In a similar way we obtain

$$|p(\alpha)|^{\alpha-1} = \frac{1+\sqrt[\alpha]{\frac{C_\eta^- - \frac{O_{n,2}(|p(\alpha)|^\alpha)}{|p(\alpha)|^\alpha}}{C_\eta^+ + \frac{O_{n,1}(q(\alpha)^\alpha)}{q(\alpha)^\alpha}}}}{C_\eta^- - \frac{O_{n,2}(|p(\alpha)|^\alpha)}{|p(\alpha)|^\alpha}}.$$

From these equalities it follows that

$$C_\eta^+ q(\alpha)^{\alpha-1} = 1 + \sqrt[\alpha]{\frac{C_\eta^+ + \frac{O_{n,1}(q(\alpha)^\alpha)}{q(\alpha)^\alpha}}{C_\eta^- - \frac{O_{n,2}(|p(\alpha)|^\alpha)}{|p(\alpha)|^\alpha}}} - \frac{O_{n,1}(q(\alpha)^\alpha)}{q(\alpha)}$$

and

$$C_\eta^- |p(\alpha)|^{\alpha-1} = 1 + \sqrt[\alpha]{\frac{C_\eta^- - \frac{O_{n,2}(|p(\alpha)|^\alpha)}{|p(\alpha)|^\alpha}}{C_\eta^+ + \frac{O_{n,1}(q(\alpha)^\alpha)}{q(\alpha)^\alpha}}} + \frac{O_{n,2}(|p(\alpha)|^\alpha)}{|p(\alpha)|}.$$

In this way we obtain that

$$
\begin{aligned}
(4.6)\qquad & C_\eta^+\, C_\eta^-\ |p(\alpha)|^{\alpha-1}\, q(\alpha)^{\alpha-1} - C_\eta^-\, |p(\alpha)|^{\alpha-1} - C_\eta^+ q(\alpha)^{\alpha-1} \\
&= -\sqrt[\alpha]{\frac{C_\eta^- - \frac{O_{n,2}(|p(\alpha)|^\alpha)}{|p(\alpha)|^\alpha}}{C_\eta^+ + \frac{O_{n,1}(q(\alpha)^\alpha)}{q(\alpha)^\alpha}}} \cdot \frac{O_{n,1}\left(q(\alpha)^\alpha\right)}{q(\alpha)} \\
&\quad + \sqrt[\alpha]{\frac{C_\eta^+ + \frac{O_{n,1}(q(\alpha)^\alpha)}{q(\alpha)^\alpha}}{C_\eta^- - \frac{O_{n,2}(|p(\alpha)|^\alpha)}{|p(\alpha)|^\alpha}}} \cdot \frac{O_{n,2}\left(|p(\alpha)|^\alpha\right)}{p(\alpha)} \\
&\quad - \frac{O_{n,1}\left(q(\alpha)^\alpha\right)}{q(\alpha)} \cdot \frac{O_{n,2}\left(|p(\alpha)|\right)^\alpha}{|p(\alpha)|}\,.
\end{aligned}
$$

But

$$
\left|\frac{O_{n,1}\left(q(\alpha)^\alpha\right)}{q(\alpha)}\right| = \left|O'_{n,1}(x)\right| \cdot q(\alpha)^{\alpha-1}\ C \cdot |x| \cdot q(\alpha)^{\alpha-1}\ ,
$$

where $0 < x < q(\alpha)^\alpha$ and C is a positive constant (mean value theorem).

Thus, by Lemma 4.3

$$
(4.7)\qquad \lim_{\alpha\to 1^-} \left|\frac{O_{n,1}\left(q(\alpha)^\alpha\right)}{(\alpha-1)\cdot q(\alpha)}\right| C \cdot \lim_{\alpha\to 1^-} \frac{q(\alpha)^\alpha}{\alpha-1} \cdot \lim_{\alpha\to 1} q(\alpha)^{\alpha-1} = 0\,.
$$

In the same way we obtain

$$
(4.8)\qquad \lim_{\alpha\to 1^-} \left|\frac{O_{n,2}\left(\left|p_{(\alpha)}\right|^\alpha\right)}{(\alpha-1)\cdot p(\alpha)}\right| = 0\,.
$$

In addition,

$$
\lim_{\alpha\to 1^-} \sqrt[\alpha]{\frac{C_\eta^- - \frac{O_{n,2}(|p(\alpha)|^\alpha)}{|p(\alpha)|^\alpha}}{C_\eta^+ + \frac{O_{n,1}(q(\alpha)^\alpha)}{q(\alpha)^\alpha}}} = \frac{C_{\eta_0}^-}{C_{\eta_0}^+}
$$

and

$$
\lim_{\alpha\to 1^-} \sqrt[\alpha]{\frac{C_\eta^+ + \frac{O_{n,1}(q(\alpha)^\alpha)}{q(\alpha)^\alpha}}{C_\eta^- - \frac{O_{n,2}(|p(\alpha)|^\alpha)}{|p(\alpha)|^\alpha}}} = \frac{C_{\eta_0}^+}{C_{\eta_0}^-}.
$$

Now, taking limit as $\alpha \to 1^-$ and replacing (4.7) and (4.8) in (4.6) we conclude the proof of Lemma 4.4. □

Now we prove Theorem 4.2. For this, recall that $x(\alpha,\mu,\nu)$ and $y(\alpha,\mu,\nu)$ are the continuations of $x(\alpha)$ and $y(\alpha)$ for (μ,ν) close to $(\mu(\alpha),\nu(\alpha))$ in the (μ,ν)-parameter space. Observe that $x(\alpha,\mu(\alpha),\nu(\alpha)) = x(\alpha)$ and $y(\alpha,\mu(\alpha),\nu(\alpha)) = y(\alpha)$.

Since $F_\alpha(\mu,\nu) = (\mu - x(\alpha,\mu,\nu), \nu - y(\alpha,\mu,\nu))$ we have that

$$
DF_\alpha(u(\alpha),\nu(\alpha)) = \begin{pmatrix} 1-\partial_\mu x(\alpha,\mu,\nu) & -\partial_\nu x(\alpha,\mu,\nu) \\ -\partial_\mu y(\alpha,\mu,\nu) & 1-\partial_\nu y(\alpha,\mu,\nu) \end{pmatrix}.
$$

So,

$$\begin{aligned}\det DF_\alpha\left(u\left(\alpha\right),\nu\left(\alpha\right)\right) = &\left(1-\partial_\mu x\left(\alpha,\mu\left(\alpha\right),\nu\left(\alpha\right)\right)\right)\cdot\left(1-\partial_\nu y\left(\alpha,\mu\left(\alpha\right),\nu\left(\alpha\right)\right)\right)\\ &-\partial_\mu y\left(\alpha,\mu\left(\alpha\right),\nu\left(\alpha\right)\right)\cdot\partial_\nu x\left(\alpha,\mu\left(\alpha\right),\nu\left(\alpha\right)\right).\end{aligned}$$

By definition $g_\alpha\left(\mu,\nu,x\left(\alpha,\mu,\nu\right)\right) = x\left(\alpha,\mu,\nu\right)$ and $g_\alpha\left(\mu,\nu,y\left(\alpha,\mu,\nu\right)\right) = y\left(\alpha,\mu,\nu\right)$. Then, by the implicit function theorem we get

$$\partial_\mu x\left(\alpha,\mu\left(\alpha\right),\nu\left(\alpha\right)\right) = \frac{\partial_\mu g_\alpha\left(u\left(\alpha\right),\nu\left(\alpha\right),x\left(\alpha\right)\right)}{1-\partial_x g_\alpha\left(\mu\left(\alpha\right),\nu\left(\alpha\right),x\left(\alpha\right)\right)},$$

$$\partial_\nu x\left(\alpha,\mu\left(\alpha\right),\nu\left(\alpha\right)\right) = \frac{\partial_\nu g_\alpha\left(u\left(\alpha\right),\nu\left(\alpha\right),x\left(\alpha\right)\right)}{1-\partial_x g_\alpha\left(\mu\left(\alpha\right),\nu\left(\alpha\right),x\left(\alpha\right)\right)},$$

$$\partial_\mu y\left(\alpha,\mu\left(\alpha\right),\nu\left(\alpha\right)\right) = \frac{\partial_\mu g_\alpha\left(u\left(\alpha\right),\nu\left(\alpha\right),y\left(\alpha\right)\right)}{1-\partial_x g_\alpha\left(\mu\left(\alpha\right),\nu\left(\alpha\right),y\left(\alpha\right)\right)},$$

$$\partial_\nu y\left(\alpha,\mu\left(\alpha\right),\nu\left(\alpha\right)\right) = \frac{\partial_\nu g_\alpha\left(u\left(\alpha\right),\nu\left(\alpha\right),y\left(\alpha\right)\right)}{1-\partial_x g_\alpha\left(\mu\left(\alpha\right),\nu\left(\alpha\right),y\left(\alpha\right)\right)}.$$

Thus,

$$\begin{aligned}&\det DF_\alpha\left(u\left(\alpha\right),\nu\left(\alpha\right)\right)\\ &= \frac{1-\partial_x g_\alpha\left(u\left(\alpha\right),\nu\left(\alpha\right),x\left(\alpha\right)\right)-\partial_\mu g_\alpha\left(u\left(\alpha\right),\nu\left(\alpha\right),x\left(\alpha\right)\right)}{1-\partial_x g_\alpha\left(u\left(\alpha\right),\nu\left(\alpha\right),x\left(\alpha\right)\right)}\\ &\quad\cdot\frac{\left(1-\partial_x g_\alpha\left(u\left(\alpha\right),\nu\left(\alpha\right),y\left(\alpha\right)\right)-\partial_\nu g_\alpha\left(u\left(\alpha\right),\nu\left(\alpha\right),y\left(\alpha\right)\right)\right)}{1-\partial_x g_\alpha\left(\mu\left(\alpha\right),\nu\left(\alpha\right),y\left(\alpha\right)\right)}\\ &\quad-\frac{\partial_\mu g_\alpha\left(u\left(\alpha\right),\nu\left(\alpha\right),y\left(\alpha\right)\right)}{1-\partial_x g_\alpha\left(u\left(\alpha\right),\nu\left(\alpha\right),x\left(\alpha\right)\right)}\cdot\frac{\partial_\nu g_\alpha\left(u\left(\alpha\right),\nu\left(\alpha\right),y\left(\alpha\right)\right)}{1-\partial_x g_\alpha\left(u\left(\alpha\right),\nu\left(\alpha\right),y\left(\alpha\right)\right)}.\end{aligned}$$

Now we will prove that

$$\lim_{\alpha\to 1^-}\frac{\det DF_\alpha\left(\mu(\alpha),\nu(\alpha)\right)}{\alpha-1} = 2+\frac{C^+_{\eta_0}}{C^-_{\eta_0}}+\frac{C^-_{\eta_0}}{C^+_{\eta_0}}.$$

In fact, note that

$$\begin{aligned}&\det DF_\alpha\left(u\left(\alpha\right),\nu\left(\alpha\right)\right)\\ &= \frac{E}{\left(1-\partial_x g_\alpha\left(u\left(\alpha\right),\nu\left(\alpha\right),x\left(\alpha\right)\right)\right)\left(1-\partial_x g_\alpha\left(u\left(\alpha\right),\nu\left(\alpha\right),y\left(\alpha\right)\right)\right)},\end{aligned}\tag{4.9}$$

where

$$\begin{aligned}E = &\left[1-\partial_x g_\alpha\left(u\left(\alpha\right),\nu\left(\alpha\right),y\left(\alpha\right)\right)-\partial_\mu g_\alpha\left(u\left(\alpha\right),\nu\left(\alpha\right),y\left(\alpha\right)\right)\right]\\ &\cdot\left[1-\partial_x g_\alpha\left(u\left(\alpha\right),\nu\left(\alpha\right),y\left(\alpha\right)\right)-\partial_\nu g_\alpha\left(\partial_x g_\alpha\left(u\left(\alpha\right),\nu\left(\alpha\right),y\left(\alpha\right)\right)\right)\right]\\ &-\partial_\mu g_\alpha\left(u\left(\alpha\right),\nu\left(\alpha\right),y\left(\alpha\right)\right)\cdot\partial_\nu g_\alpha\left(u\left(\alpha\right),\nu\left(\alpha\right),y\left(\alpha\right)\right).\end{aligned}$$

Using the definitions of g_α we obtain

$$\begin{aligned}E = &\Big[1-\alpha C_\eta^- k^{\alpha-1}\left|x\left(\alpha\right)\right|^{\alpha-1}+O_{n,2}^{'}\left(\left|kx\left(\alpha\right)\right|^{\alpha}\right)\left|kx\left(\alpha\right)\right|^{\alpha-1}\\ &+\partial_{a^+}\ C_\eta^-\ \left|kx\left(\alpha\right)\right|^{\alpha}-\partial_{a^+}O_{n,2}\left(\left|kx\left(\alpha\right)\right|^{\alpha}\right)\Big]\\ &\cdot\Big[1-\alpha C_\eta^+ k^{\alpha-1}y\left(\alpha\right)^{\alpha-1}-O_{n,1}^{'}\left(\left|ky\left(\alpha\right)\right|^{\alpha}\right)\left|ky\left(\alpha\right)\right|^{\alpha-1}\\ &-\partial_{a^-}C_\eta^-\left|ky\left(\alpha\right)\right|^{\alpha}-\partial_{a^-}O_{n,1}\left(\left|ky\left(\alpha\right)\right|^{\alpha}\right)\Big]\\ &-\left[1+\partial_{a^+}C_\eta^+\ \left|ky\left(\alpha\right)\right|^{\alpha}+\partial_{a^+}O_{n,1}\left(\left|ky\left(\alpha\right)\right|^{\alpha}\right)\right]\\ &\cdot\left[1-\partial_{a^-}C_\eta^-\left|kx\left(\alpha\right)\right|^{\alpha}+\partial_{a_-}\ O_{n,2}\left(\left|kx\left(\alpha\right)\right|^{\alpha}\right)\right]\ .\end{aligned}$$

Arranging terms we get

$$\begin{aligned}E = &\left[1-\alpha C_\eta^-\left|kx\left(\alpha\right)\right|^{\alpha-1}+A_1\ \left|kx\left(\alpha\right)\right|^{\alpha}\right]\cdot\left[1-\alpha C_\eta^+\left|ky\left(\alpha\right)\right|^{\alpha-1}+A_2\left|ky\left(\alpha\right)\right|^{\alpha}\right]\\ &-\left[1+\partial_{a^+}C_\eta^+\left|ky\left(\alpha\right)\right|^{\alpha-1}+A_3\right]\cdot\left[1-\partial_{a^-}C_\eta'\left|kx\left(\alpha\right)\right|^{\alpha}+A_4\right],\end{aligned}$$

where

$$\begin{aligned}A_1 &= \frac{O_{n,2}'\left(\left|kx\left(\alpha\right)\right|^{\alpha}\right)}{\left|kx\left(\alpha\right)\right|}+\partial_{a^+}C_\eta^- -\frac{\partial_{a^+}O_{n,2}\left(\left|kx\left(\alpha\right)\right|^{\alpha}\right)}{\left|kx\left(\alpha\right)\right|^{\alpha}},\\ A_2 &= -\frac{O_{n,1}'\left(\left|ky\left(\alpha\right)\right|^{\alpha}\right)}{\left|ky\left(\alpha\right)\right|}+\partial_{a-}C_\eta^+ -\frac{\partial_{a^+}O_{n,1}\left(\left|ky\left(\alpha\right)\right|^{\alpha}\right)}{\left|ky\left(\alpha\right)\right|^{\alpha}},\\ A_3 &= \partial_{a^+}\ O_{n,1}\left(\left|ky\left(\alpha\right)\right|^{\alpha}\right),\end{aligned}$$

and $A_4 = \partial_{a^-}\ O_{n,2}\left(\left|kx\left(\alpha\right)\right|^{\alpha}\right)$.

Note that $\lim_{\alpha\to 1^-}A_i$ exists and is a finite number for $1\le i\le 4$.

Recall that $k\,x\left(\alpha\right)=p\left(\alpha\right)$ and $k\,y\left(\alpha\right)=q\left(\alpha\right)$, then

$$\begin{aligned}E = &\left(1-\alpha C_\eta^-\left|p\left(\alpha\right)\right|^{\alpha-1}\right)\cdot\left(1-\alpha C_\eta^+ q\left(\alpha\right)^{\alpha-1}\right)\\ &+\left(1-\alpha C_\eta^-\left|p\left(\alpha\right)\right|^{\alpha-1}\right)A_2\ q\left(\alpha\right)^{\alpha}\\ &+\left(1-\alpha C_\eta^+ q\left(\alpha\right)^{\alpha-1}\right)A_1\ \left|p\left(\alpha\right)\right|^{\alpha}+A_1\ A_2\ q\left(\alpha\right)^{\alpha}\ \left|p\left(\alpha\right)\right|^{\alpha}\\ &-1\ -\ \partial_{a^+}C_\eta^+\ q\left(\alpha\right)^{\alpha}-\ A_3\ +\ \partial_{a^-}\ C_\eta^-\left|p\left(\alpha\right)\right|^{\alpha}-\partial_{a^+}C_\eta^+\ q\left(\alpha\right)^{\alpha}\ \partial_{a^-}C_\eta^-\left|p\left(\alpha\right)\right|^{\alpha}\\ &-A_3\ \partial_{a^-}C_\eta^-\left|p\left(\alpha\right)\right|^{\alpha}-A_4-\partial_{a^+}C_\eta^+\ q\left(\alpha\right)^{\alpha}\ A_4-A_3\ A_4.\end{aligned}$$

So,

$$\begin{aligned}E = &-\alpha\ C_\eta^-\left|p\left(\alpha\right)\right|^{\alpha-1}-\alpha\ C_\eta^+\ q\left(\alpha\right)^{\alpha-1}+\alpha^2\ C_\eta^-\ C_\eta^+\ \left|p\left(\alpha\right)\right|^{\alpha-1}\ q\left(\alpha\right)^{\alpha-1}\\ &+A_5\ q\left(\alpha\right)^{\alpha}+A_6\ \left|p\left(\alpha\right)\right|^{\alpha}+A_7\ q\left(\alpha\right)^{\alpha}\ \left|p\left(\alpha\right)\right|^{\alpha},\end{aligned}$$

where

$$\begin{aligned}A_5 &= \left(1-\alpha\ C_\eta^-\ \left|p\left(\alpha\right)\right|^{\alpha-1}\right)A_2-\partial_{a^+}\ C_\eta^+-\frac{A_3}{q\left(\alpha\right)^{\alpha}}-\partial_{a^+}\ C_\eta^+\ A_4,\\ A_6 &= \left(1-\alpha\ C_\eta^+\ q\left(\alpha\right)^{\alpha-1}\right)A_1+\partial_{a^-}\ C_\eta^- -A_3\ \partial_{a^-}\ C_\eta^- -\frac{A_4}{\left|p\left(\alpha\right)\right|^{\alpha}},\end{aligned}$$

and $A_7 = A_1\, A_2 - \partial_{a^+}\, C_\eta^+\, \partial_{a^-}\, C_\eta^- - \frac{A_3\, A_4}{q(\alpha)^\alpha\, |p(\alpha)|^\alpha}$.

Note that $\lim_{\alpha\to 1^-} A_i$ exists and is a finite number for $5 \le i \le 7$.

Thus,

$$\begin{aligned} E = {} & \left(\alpha^2 - \alpha\right)\left[C_\eta^-\ |p\,(\alpha)|^{\alpha-1} + C_\eta^+\ q\,(\alpha)^{\alpha-1}\right] \\ & + \alpha^2\left[C_\eta^+\ C_\eta^-\ |p\,(\alpha)|^{\alpha-1}\, q\,(\alpha)^{\alpha-1} - C_\eta^-\, |p\,(\alpha)|^{\alpha-1} - C_\eta^+ q\,(\alpha)^{\alpha-1}\right] \\ & + A_5\, q\,(\alpha)^\alpha + A_6\, |p\,(\alpha)|^\alpha + A_7\, q\,(\alpha)^\alpha\, |p\,(\alpha)|^\alpha\,. \end{aligned}$$

Now, taking limits and using Lemmas 4.3 and 4.4 we obtain that

$$\begin{aligned} \lim_{\alpha\to 1^-} \frac{E}{\alpha - 1} = {} & \lim_{\alpha\to 1} \alpha\left[C_\eta^-\ |p\,(\alpha)|^{\alpha-1} + C_\eta^+\ q\,(\alpha)^{\alpha-1}\right] \\ & + \lim_{\alpha\to 1^-} \frac{\alpha^2}{\alpha - 1}\left[C_\eta^+\ C_\eta^-\ |p\,(\alpha)|^{\alpha-1}\ q\,(\alpha)^{\alpha-1} - C_\eta^-\ |p\,(\alpha)|^{\alpha-1} + C_\eta^+\ q\,(\alpha)^{\alpha-1}\right] \\ & + \lim_{\alpha\to 1^-}\left[A_5\, \frac{q\,(\alpha)^\alpha}{\alpha - 1} + A_6\, \frac{|p\,(\alpha)|^\alpha}{\alpha - 1} + A_7\, \frac{q\,(\alpha)^\alpha}{\alpha - 1}\ |p\,(\alpha)|^\alpha\right] \\ & = C_{\eta_0}^-\, B + C_{\eta_0}^+\, B = 2 + \frac{C_{\eta_0}^+}{C_{\eta_0}^-} + \frac{C_{\eta_0}^-}{C_{\eta_0}^+}. \end{aligned}$$

To finish, we have that

$$\begin{aligned} & \lim_{\alpha\to 1^-} \left(1 - \partial_x g_\alpha\left(\mu\,(\alpha), \nu\,(\alpha), x\,(\alpha)\right)\right) \cdot \left(1 - \partial_x g_\alpha\left(\mu\,(\alpha), \nu\,(\alpha), y\,(\alpha)\right)\right) \\ & \qquad = \left(1 - C_{\eta_0}^- B\right)\left(1 - C_{\eta_0}^+ B\right) = 1\,. \end{aligned}$$

So, replacing this last equality in (4.9) we get

$$\lim_{\alpha\to 1^-} \frac{\det DF_\alpha\left(\mu_{(\alpha)}, \nu_{(\alpha)}\right)}{\alpha - 1} = 2 + \frac{C_{\eta_0}^+}{C_{\eta_0}^-} + \frac{C_{\eta_0}^-}{C_{\eta_0}^+}.$$

All of these facts together conclude the proof of Theorem 4.2. □

5. One-dimensional analysis. In this section we will prove that the trapping region obtained in the previous section is contained in the basin of a transitive attractor of $g_\alpha(\mu, \nu)$. Depending on the geometry of the original vector field we can obtain a different kind of such attractor. When $\nu^+ = \nu^- = 1$, the attractor is the interval $[\mu, \nu]$. If $\nu^+ = \nu^- = -1$ we have three alternatives: (1) the attractor is given by the interval $[\nu, g_\alpha(\mu, \nu, \mu)]$, (2) $[\nu, \mu]$, or (3) $[g_\alpha(\mu, \nu, \mu), \mu]$, depending on $\nu < g_\alpha(\mu, \nu, \mu) < \mu \le g_\alpha(\mu, \nu, \nu)$, $\nu < g_\alpha(\mu, \nu, \mu) < g_\alpha(\mu, \nu, \nu) < \mu$, or $g_\alpha(\mu, \nu, \mu) \le \nu < g_\alpha(\mu, \nu, \nu) < \mu$, respectively. For $\nu^+ = -\nu^- = 1$ we have two possibilities: (1) $[\mu, g_\alpha(\mu, \nu, \mu)]$ when $\mu \le \nu$ and (2) $[\nu, g_\alpha(\mu, \nu, \nu)]$ when $\nu < \mu$. In the same way, for $\nu^+ = -\nu^- = -1$, we have (1) $[g_\alpha(\mu, \nu, \mu), \nu]$ or (2) $[g_\alpha(\mu, \nu, \nu), \mu]$, when $\mu \le \nu$ or $\nu < \mu$, respectively (see Figure 5.1).

The proof below is similar to the one given in [Rob4, Theorem A] and we are including it here for completeness.

THEOREM 5.1. *Fix $\lambda > 1$ and let $n_0 \in \mathbb{N}$ be such that*

$$\frac{\lambda^{n_0}}{2} > 1.$$

Let $a < 0 < b$ and $f : [a, b]^* \to [a, b]$, $[a, b]^* = [a, b] \setminus \{0\}$ a map such that $f(0_-) = \mu$, $f(0_+) = \nu$. Assume also that the following conditions hold:

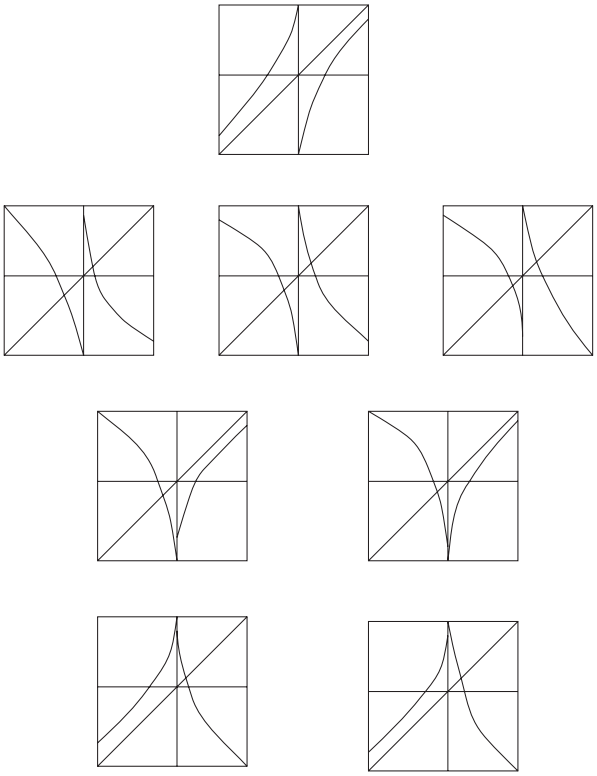

FIG. 5.1. *Possible attractors for* $g_\alpha(\mu,\nu,.)$.

(H0) $[a,b] = convexhull\{\mu,\nu,f(\mu),f(\nu)\}$,
(H1) f *has a single singularity at* $x=0$,
(H2) f *is differentiable in* $[a,b]^*$ *and* $|f'(x)| \geq \lambda$ *for all* $x \in [a,b]^*$,
(H3) *there are* $x<0<y$ *such that* $f(x)=f(y)=0$,
(H4) *there is* $\epsilon>0$ *such that if* $f(0_+)=\nu$ *and* $f(0_-)=\mu$, *then* $\{f^i(\mu), f^i(\nu), f^i(a), f^i(b) : 0 \leq i \leq n_0\} \subset [a,a+\epsilon] \cup [b-\epsilon,b]$ *for all* $1 \leq i \leq n_0$.

Then, for ϵ *sufficiently small* f *is transitive.*

Note 4. For $n_0=2$ this lemma reduces to a well-known result in [W].

Proof. First of all, hypothesis (H0) shows that $f([x,0))$ or $f((0,y])$ must contain $(a,0]$ or $[0,b)$. Now, fix and open interval $I \subset [a,b]$. We claim that one of the following alternatives hold:

(A) $f(I)$ is an interval and $|f(I)| \geq \lambda |I|$,
(B) there is $n \in \mathbb{N}$ such that $f^n(I)$ contains $(a,0]$ or $[0,b)$,
(C) there is an interval $J \subset f^{n_0}(I)$ such that $|J| > \frac{\lambda^{n_0}}{2}|I|$.

Indeed, if $0 \notin I$, then (A) holds. Thus, we can assume that $0 \in I$. Then $I^* = I \setminus \{0\}$ splits into $I^* = I^+ \cup I^-$ where $I^+ = I \cap (0,b]$ and $I^- = I \cap [a,0)$.

Let $\tilde{I} = I^+$ or I^- such that $|f(\tilde{I}| \geq \lambda/2\ |I|$.

We have either
(a) $f(\tilde{I}) \supset (a,0]$,
(b) $f(\tilde{I}) \supset [0,b)$,
(c) $f(\tilde{I}) = (x,\delta)$ with $0<x<\delta$ where $\delta \in \{\mu,\nu\}$, or
(d) $f(\tilde{I}) = (\delta,x)$ with $\delta<x<0$ where $\delta \in \{\mu,\nu\}$.

In cases (a) and (b) we have (B) for I with $n = 1$.

Assume case (c).

In this case we have two possibilities, namely,

(c1) there is $1 \leq i = i(I) < n_0$ such that i is the first iterate such that $0 \in f^i(f(\tilde{I}))$,

(c2) $0 \notin f^i(f(\tilde{I}))$ for all $1 \leq i \leq n_0$.

If (c2) holds, then

$$|f^{n_0}(\tilde{I})| = |f^{n_0-1}(f(\tilde{I}))| \geq \lambda^{n_0-1} \ |f(\tilde{I})| \geq \lambda^{n_0-1} \ \lambda/2 \ |I| \geq \lambda^{n_0}/2 \ |I|.$$

If (c1) holds, then $f^i(f(\tilde{I}))$ is an interval containing both 0 and $f^i(\delta)$.

As $f^i(f(\tilde{I}))$ is an interval we conclude that $f^i(f(\tilde{I}))$ contains an interval containing 0 and $f^i(\delta)$, and this implies that it contains one of the intervals $[x, 0]$ or $[0, y]$. From here, $f^{i+1}(f(\tilde{I}))$ contains both $[x, 0]$ and $[0, y]$. Hence,

$$f^{i+2}(f(\tilde{I})) = f(f^{i+1}(f(\tilde{I}))) \supset (a, 0]$$

or

$$f^{i+2}(f(\tilde{I})) = f(f^{i+1}(f(\tilde{I}))) \supset [0, b).$$

Thus, (B) holds with $n = i + 2$.

Finally, assume (d). The proof is similar to the previous case.

To finish the proof we proceed as follows.

First we claim that for all open intervals $I \subset [a, b]$ there is $N \in \mathbb{N}$ such that $f^N(I)$ contains either $(a, 0)$ or $(0, b)$.

The proof goes by contradiction. Assume that such an N does not exist. Then, either (A) or (C) holds.

Fix $I_1 = I$. Then there is $N_1 \in \mathbb{N}$ such that $f^{N_1}(I_1)$ contains an interval I_2 such that $|I_2| \geq \min\{\lambda, \frac{\lambda^{n_0}}{2}\} \ |I_1|$. Set $k = \min\{\lambda, \frac{\lambda^{n_0}}{2}\}$. Note that (B) is not true for I_2 (because $I_2 \subset f^{N_1}(I_2)$). Then, there is $N_2 \in \mathbb{N}$ such that $f^{N_2}(I_2)$ contains an interval I_3 such that $|I_3| \geq k \ |I_2| \geq k^2 \ |I_1|$.

In this way we construct a sequence of intervals $I_1, I_2, I_3, \ldots$ such that

$$|I_i| \geq k^i \ |I_1|.$$

This yields a contradiction since $k > 1$. This proves the claim.

Now we finish the proof arguing in the following way.

If $f^N(I) \supset (a, 0)$ or $f^N(I) \supset (0, b)$, then $f^N(I) \cup f^{N+1}(I) \supset (a, b)$. This implies that f is transitive and the theorem follows. □

Note 5. It is not difficult to prove, using (H4), that the map f in Theorem 5.1 is *leo* (locally eventually onto), i.e., for any interval $J \subset I$, there is $n \geq 1$ such that $f^n(J) = I$. Clearly a leo map is transitive.

Proof of Theorem 1.1. Let X_η satisfying (A1)–(A7), and f_η be the one-parameter family associated with X_η given by Lemma 2.1 in section 2. Recall that $\eta = \eta(\alpha, a^+, a^-)$. Next, as in section 3 we obtain $g_\alpha(\mu, \nu, x)$ which is a renormalization of f_η.

From now on we work with $g_\alpha(\mu, \nu, x)$.

By (A5), we can apply Theorem 4.1. Let $\mathcal{O}^- \subset \mathbb{R}^3$ be the open set given by Theorem 4.1. For each $\alpha \in O$ close enough to 1, where O is given by Lemma 2.2, consider $(\alpha, \mu(\alpha), \nu(\alpha))$, with $(\mu(\alpha), \nu(\alpha))$ defined at the beginning of section 3.

By the definition of $\mathcal{O}^-$ we have that $(\alpha, \mu(\alpha), \nu(\alpha)) \in \mathrm{Cl}(\mathcal{O}^-)$.

By Theorem 4.1(2) we have that for all $\theta \in \mathcal{O}^-$ there is I_θ such that $g_\alpha(\mu, \nu, I_\theta) \subset \mathrm{Int}(I_\theta)$. Set $\tilde{I}_\theta \subset I_\theta$ the $convexhull\{\mu, \nu, g_\alpha(\mu, \nu, \mu), g_\alpha(\mu, \nu, \nu)\}$. Taking (μ, ν) close to $(\mu(\alpha), \nu(\alpha))$, the map $g_\alpha(\mu, \nu, \cdot)$ satisfies the hypotheses of Theorem 5.1. Indeed, (H0) follows from the definition of $\tilde{I}_\theta$, (H1) is straightforward by the definition of g_α, and (H2), (H3), and (H4) follow from Lemma 3.3 for (μ, ν) close to $(\mu(\alpha), \nu(\alpha))$. Hence, the restriction of $g_\alpha(\mu, \nu, .)$ to $\tilde{I}_\theta$ is transitive, and thus we conclude that $g_\alpha(\mu, \nu, .)$ is transitive in $\tilde{I}_\theta$. All of these facts together conclude the proof of Theorem 1.1. □

REFERENCES

[ABS] V. S. Afraimovich, Bykov, and L. P. Shil'nikov, *On the appearance and structure of the Lorenz attractor*, Dokl. Acad. Sci. USSR, 234 (1977), pp. 336–339.

[ACL] V. Afraimovich, S. Chow, and W. Liu, *Lorenz-type attractors from co-dimension one bifurcation*, J. Dynam. Differential Equations, 7 (1995), pp. 375–407.

[AH] V. S. Afraimovich and S.-B. Hsu, *Lectures on Chaotic Dynamical Systems*, AMS/IP Stud. Adv. Math. 28, AMS, Providence, RI, 2003.

[AL] L. Alsedá and J. Llibre, *Wandering intervals for Lorenz maps with bounded nonlinearity*, in Dynamical Systems and Ergodic Theory, Warsaw, 1986, Banach Center Publ. 23, PWN, Warsaw, 1989, pp. 83–89.

[BPV] Ch. Bonatti, A. Pumarino, and M. Viana, *Lorenz attractors with arbitrary expanding dimension*, C. R. Acad. Sci. Paris Sér. I Math., 325 (1997), pp. 883–888.

[BW] J. S. Birman and R. F. Williams, *Knotted periodic orbits in dynamical systems. I. Lorenz's equations*, Topology 22 (1983), pp. 47–82.

[BL] V. A. Bouichenko and G. A. Leonov, *The Hausdorff dimension of attractors of the Lorenz system*, Differ. Uravn., 25 (1989), pp. 1999–2000, 2023 (in Russian).

[Bu] L. A. Bunimovich, *Statistical properties of the Lorenz model*, Izv. Vyssh. Uchebn. Zaved. Radiofiz., 28 (1985), pp. 1472–1473 (in Russian).

[CPRV] M. J. Costa, M. J. Pacifico, A. Rovella, and M. Viana, *Persistence of global spiral attractors*, in preparation.

[GW] J. Guckenheimer and R. F. Williams, *Structural stability of Lorenz attractors*, Inst. Hautes Études Sci. Publ. Math., 50 (1979), pp. 59–72.

[dMP] W. de Melo and J. Palis, *Geometric Theory of Dynamical Systems. An Introduction*, Springer-Verlag, New York, 1982.

[DY] Y. Ding and W. Fan, *The asymptotic periodicity of Lorenz maps*, Acta Math. Sci., Ser. B Engl. Ed. 19 (1999), pp. 114–120.

[DKO] F. Dumortier, H. Kokubu, and H. Oka, *A degenerate singularity generating geometric Lorenz attractors*, Ergodic Theory Dynam. Systems, 15 (1995), pp. 833–856.

[GH] R. W. Ghrist and P. J. Holmes, *An ODE whose solutions contain all knots and links*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 6 (1996), pp. 779–800.

[HS] J. H. Hubbard and C. Sparrow, *The classification of topologically expansive Lorenz maps*, Comm. Pure Appl. Math., 43 (1990), pp. 431–443.

[KS] G. Keller and M. St. Pierre, *Topological and measurable dynamics of Lorenz maps*, in Ergodic Theory, Analysis and Efficient Simulations of Dynamical Systems, Springer, Berlin, 2001, pp. 333–361.

[KKO] M. Kisaka, H. Kokubu, and H. Oka, *Supplement to homoclinic doubling bifurcation in vector fields*, in Dynamical Systems, Santiago, 1990, Pitman Res. Notes Math. Ser. 285, Longman Sci. Tech., Harlow, 1993, pp. 92–116.

[Kl] N. Klinshpont, *A topological invariant of the Lorenz attractor*, Uspekhi Mat. Nauk, 47 (1992), pp. 195–196 (in Russian); translation in Russian Math. Surveys, 47 (1992), pp. 221–223.

[Ko1] M. Komuro, *Expansive properties of Lorenz attractors*, in The Theory of Dynamical Systems and Its Applications to Nonlinear Problems, World Scientific, Singapore, 1984, pp. 4–26.

[Ko2] M. Komuro, *Lorenz attractors do not have the pseudo-orbit tracing property*, J. Math. Soc. Japan, 37 (1985), pp. 489–514.

[LM1] R. Labarca and C. Moreira, *Bifurcation of the essential dynamics of Lorenz maps and applications to Lorenz-like flows: Contributions to the study of the expanding case*, Bol. Soc. Brasil Mat. (N.S.), 32 (2001), pp. 107–144.

[LM2] R. Labarca and C. Moreira, *Bifurcation of the essential dynamics of Lorenz maps on the*

real line and the bifurcation scenario for the linear family, Sci. Ser. A. Math. Sci. (N.S.), 7 (2001), pp. 13–29.

[Le1] G. A. Leonov, *Formulas for the Lyapunov dimension of Hénon and Lorenz attractors*, Algebra i Analiz, 13 (2001), pp. 155–170 (in Russian); translation in St. Petersburg Math. J., 13 (2002), 453–464.

[Le2] G. A. Leonov, *Asymptotic behavior of the solutions of the Lorenz system*, Differ. Uravn., 24 (1988), pp. 804–809, 917 (in Russian); translation in Differential Equations, 24 (1988), pp. 527–531 .

[Lo] E. N. Lorenz, *Deterministic nonperiodic flow,* J. Atmospheric Sci., 20 (1963), pp. 130–141.

[Lo84] E. N. Lorenz, *The local structure of a chaotic attractor in four dimensions*, Phys. D, 13 (1984), pp. 90–104.

[LV] S. Luzzatto and M. Viana, *Positive Lyapunov exponents for Lorenz-like families with criticalities*, in Geometrie Complexe et Systemes Dynamiques, Orsay, 1995, Asterisque 261, 2000, pp. 201–237.

[MdeM] M. Martens and W. de Melo, *Universal models for Lorenz maps*, Ergodic Theory Dynam. Systems, 21 (2001), pp. 833–860.

[Me1] R. J. Metzger, *Stochastic stability for contracting Lorenz maps and flows*, Comm. Math. Phys., 212 (2000), pp. 277–296.

[Me2] R. J. Metzger, *Sinai-Ruelle-Bowen measures for contracting Lorenz maps and flows*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 17 (2000), pp. 247–276.

[Mo] C. Morales, *Lorenz attractors through saddle-node bifurcations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 13 (1996), pp. 589–617.

[MPP1] C. A. Morales, M. J. Pacifico, and E. R. Pujals, *Strange attractors across the boundary of hyperbolic systems*, Comm. Math. Phys., 211 (2000), pp. 527–558.

[MPP2] C. A. Morales, M. J. Pacifico, and E. R. Pujals, *On C^1 robust transitive sets for three dimensional flows*, C. R. Acad. Sci. Paris, Série 1 326 (1998), pp. 81–86.

[MPP3] C. A. Morales, M. J. Pacifico, and E. R. Pujals, *Robust transitive singular sets for 3-flows are partially hyperbolic attractors or repellers*, Ann. of Math. (2), 160 (2004), pp. 375–432.

[MPu] C. A. Morales and E. R. Pujals, *Singular strange attractors on the boundary of Morse-Smale systems*, Ann. Sci. École Norm. Sup. (4), 30 (1997), pp. 693–717.

[MSV] E. Munoz, B. San Martin, and J. Vera, *Nonhyperbolic persistent attractors near the Morse-Smale boundary*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 20 (2003), pp. 867–888.

[N] V. Naudot, *Strange attractor in the unfolding of an inclination-flip homoclinic orbit*, Ergodic Theory Dynam. Systems, 16 (1996), pp. 1071–1086.

[PRV] M. J. Pacifico, A. Rovella, and M. Viana, *Infinite-modal maps with global chaotic behavior* Ann. of Math. (2), 148 (1998), 441–484.

[PT] J. Palis and F. Takens, *Hyperbolicity and Sensitive Chaotic Dynamics at Homoclinic Bifurcations. Fractal Dimensions and Infinitely Many Attractors*, Cambridge Stud. Adv. Math. 35, Cambridge University Press, Cambridge, UK, 1993.

[P] Ya. Pesin, *Dynamical systems with generalized hyperbolic attractors: Hyperbolic, ergodic and topological properties*, Ergodic Theory Dynam. Systems, 12 (1992), pp. 123–151.

[PS] Ya. B. Pesin and Ya. G. Sinai, *Hyperbolicity and stochasticity of dynamical systems*, in Mathematical Physics Reviews, Vol. 2, Soviet Sci. Rev. Sect. C: Math. Phys. Rev. 2, Harwood Academic, Chur, 1981, pp. 53–115.

[Rob1] C. Robinson, *Nonsymmetric Lorenz attractors from a homoclinic bifurcation*, SIAM J. Math. Anal., 32 (2000), pp. 119–141.

[Rob2] C. Robinson, *Homoclinic bifurcation to a transitive attractor of Lorenz type* II, SIAM J. Math. Anal., 23 (1992), pp. 1255–1268.

[Rob3] C. Robinson, *Homoclinic bifurcation to a transitive attractor of Lorenz type*, Nonlinearity, 2 (1989), pp. 495–518.

[Rob4] C. Robinson, *Transitivity and invariant measures for the geometric model of the Lorenz equations*, Ergodic Theory Dynam. Systems, 4 (1984), pp. 605–611.

[Rov] A. Rovella, *The dynamics of perturbations of the contracting Lorenz attractor*, Bol. Soc. Brasil. Mat. (N.S.), 24 (1993), pp. 233–259.

[Ry] M. R. Rychlik, *Lorenz attractors through Silnikov-type bifurcation*, I. Ergodic Theory Dynam. Systems, 10 (1990), pp. 793–821.

[S] E. A. Sataev, *Invariant measures for hyperbolic maps with singularities*, Russ. Math. Surveys, 471 (1992), pp. 191–251.

[Sp] C. Sparrow, *The Lorenz Equations: Bifurcations, Chaos and Strange Attractors*, Appl. Math. Sci. 41, Springer, Berlin, 1982.

[Sh] A. L. Shilnikov, *On bifurcations of the Lorenz attractor in the Shimizu-Morioka model*, Phys. D, 62 (1993), pp. 338–346.

[ST] L. P. Shilnikov and D. V. Turaev, *An example of a wild strange attractor*, Mat. Sb., 189 (1998), pp. 137–160 (in Russian); translation in Sb. Math., 189 (1998), pp. 291–314.

[SM] T. Shimizu and N. Morioka, *On the bifurcation of a symmetric limit cycle to an asymmetric one in a simple model*, Phys. Lett., A76 (1980), pp. 201–204.

[S] M. Shub, *Global Stability of Dynamical Systems*, Springer, New York, 1987.

[Tu1] W. Tucker, *The Lorenz attractor exists*, C. R. Acad. Sci. Paris Sér. I Math., 328 (1999), pp. 1197–1202.

[Tu2] W. Tucker, *A rigorous ODE solver and Smale's* 14*th problem*, Found. Comput. Math., 2 (2002), pp. 53–117.

[V] M. Viana, *What's new on Lorenz strange attractor?,* Math. Intelligencer, 22 (2000), pp. 6–19.

[W] R. Williams, *The structure of Lorenz attractors*, Inst. Hautes Études Sci. Publ. Math., 50 (1979), pp. 73–99.

[Ya] L. S. Young, *On the prevalence of horseshoes*, Trans. Amer. Math. Soc., 263 (1981), pp. 75–88.

SIAM J. MATH. ANAL.
Vol. 36, No. 6, pp. 1862–1886

TRAFFIC FLOW ON A ROAD NETWORK*

G. M. COCLITE†, M. GARAVELLO‡, AND B. PICCOLI§

Abstract. This paper is concerned with a fluidodynamic model for traffic flow. More precisely, we consider a single conservation law, deduced from the conservation of the number of cars, defined on a road network that is a collection of roads with junctions. The evolution problem is underdetermined at junctions; hence we choose to have some fixed rules for the distribution of traffic plus optimization criteria for the flux. We prove existence of solutions to the Cauchy problem and we show that the Lipschitz continuous dependence by initial data does not hold in general, but it does hold under special assumptions.

Our method is based on a wave front tracking approach [A. Bressan, *Hyperbolic Systems of Conservation Laws. The One-dimensional Cauchy Problem*, Oxford University Press, Oxford, UK, 2000] and works also for boundary data and time-dependent coefficients of traffic distribution at junctions, including traffic lights.

Key words. scalar conservation laws, traffic flow

AMS subject classifications. 90B20, 35L65

DOI. 10.1137/S0036141004402683

1. Introduction.

This paper deals with a fluidodynamic model of heavy traffic on a road network. More precisely, we consider the conservation law formulation proposed by Lighthill and Whitham [14] and Richards [15]. This nonlinear framework is based simply on the conservation of cars and is described by the equation

$$\rho_t + f(\rho)_x = 0, \tag{1.1}$$

where $\rho = \rho(t,x) \in [0, \rho_{max}]$, $(t,x) \in \mathbb{R}_+ \times \mathbb{R}$, is the *density* of cars, $v = v(t,x)$ is the *velocity*, and $f(\rho) = v\,\rho$ is the *flux*. This model is appropriate for revealing shocks formation, as it is natural for conservation laws, whose solutions may develop discontinuities in finite time even for smooth initial data (see [5]). In most cases one assumes that v is a function of ρ only and that the corresponding flux is a concave function. We make this assumption; moreover, we let f have a unique maximum $\sigma \in]0, \rho_{max}[$ and for notational simplicity we assume $\rho_{max} = 1$.

Here we deal with a network of roads, as in [12]. This means that we have a finite number of roads modeled by intervals $[a_i, b_i]$ (with one of the two endpoints possibly infinite) that meet at some junctions. For endpoints that do not touch a junction (and are not infinite), we assume given boundary data and solve the corresponding boundary problem, as in [1, 2, 4]. The key role is played by junctions, at which the system is underdetermined even after prescribing the conservation of cars, that can

*Received by the editors May 19, 2004; accepted for publication (in revised form) September 3, 2004; published electronically June 16, 2005.
http://www.siam.org/journals/sima/36-6/40268.html
†Dipartimento di Matematica, Università di Bari, Via E. Orabona 4, 70125 Bari, Italy (giusepc@math.uio.no).
‡Dipartimento di Matematica e Applicazioni, Università di Milano Bicocca, Via R. Cozzi 53 - Edificio U5, 20125 Milano, Italy (mauro.garavello@unimib.it). This author was supported by SISSA-ISAS and by "Istituto per le Applicazioni del Calcolo 'Mauro Picone' ".
§Istituto per le Applicazioni del Calcolo "M. Picone," Viale del Policlinico 137, 00161 Roma, Italy (piccoli@iac.rm.cnr.it).

be written as the Rankine–Hugoniot relation

$$\sum_{i=1}^{n} f(\rho_i(t,b_i)) = \sum_{j=n+1}^{n+m} f(\rho_j(t,a_j)), \tag{1.2}$$

where ρ_i, $i = 1,\dots,n$, are the car densities on incoming roads, while ρ_j, $j = n+1,\dots,n+m$, are the car densities on outgoing roads. In [12], the Riemann problem, that is, the problem with constant initial data on each road, is solved maximizing a concave function of the fluxes, and existence of weak solutions for Cauchy problems with suitable initial data of bounded variation is proved. In this paper we assume the following:

(A) There are some prescribed preferences of drivers, that is, the traffic from incoming roads is distributed on outgoing roads according to fixed coefficients.

(B) Respecting (A), drivers choose so as to maximize fluxes.

To deal with rule (A), we fix a traffic distribution matrix

$$A \doteq \{\alpha_{ji}\}_{j=n+1,\dots,n+m,\ i=1,\dots,n} \in \mathbb{R}^{m\times n},$$

such that

$$0 < \alpha_{ji} < 1, \qquad \sum_{j=n+1}^{n+m} \alpha_{ji} = 1, \tag{1.3}$$

for each $i = 1,\dots,n$ and $j = n+1,\dots,n+m$, where α_{ji} is the percentage of drivers arriving from the ith incoming road who take the jth outgoing road. Notice that with only the rule (A) Riemann problems are still underdetermined. This choice represents a situation in which drivers have a final destination, and hence distribute on outgoing roads according to a fixed law but maximize the flux whenever possible. We are able to solve uniquely Riemann problems, under suitable conditions on the matrix A, and then to construct solutions to Cauchy problems for networks with simple junctions, i.e., junctions with two incoming roads and two outgoing ones. Our main technique is the use of a front tracking algorithm and the control of the total variation of the flux. We refer the reader to [5] for the general theory of conservation laws and for a discussion of wave front tracking algorithms.

The main difficulty in solving systems of conservation laws is the control of the total variation; see [5]. It is easy to see that for a single conservation law the total variation is decreasing; however, in our case it may increase due to interaction of waves with junctions.

There is a natural lack of symmetry for *big waves* (i.e., waves crossing the value σ; see Definition 5.8) and *bad data* (see Definition 5.8) at junctions, since the role of incoming roads is different from that of outgoing ones. Similarly, for scalar conservation laws with discontinuous coefficients, one has to use a definition of strength for discontinuities of the coefficient, seen as waves, that is not symmetric but depends on the sign of the jump in the solution; see [13, 16, 17]. This is enough to control the total variation in that case; on the contrary, our problem is more delicate. In fact, the variation can still increase due to interactions of waves with junctions (and there is no bound on the number of interactions and of the size of magnification; see Appendix B). The bounded quantity is the total variation of the flux. We prove this fact for junctions with only two incoming roads and two outgoing ones. Unfortunately the total variation of the flux is not equivalent to the total variation of ρ, since $f'(\sigma) = 0$,

and so it is not sufficient to prove existence of solutions. Therefore some compactness argument is used together with a bound of big waves near junctions.

Our techniques are quite flexible, so we can deal with time-dependent coefficients for rule (A). In particular, we can model traffic lights and, in this case, the control of total variation is extremely delicate. An arbitrarily small change in the coefficients can produce waves whose strength is bounded away from zero. Still, it is possible to consider periodic coefficients, a case of particular interest for applications. We can also deal with roads having different fluxes: these can be treated in the same way with the necessary notational modifications.

There is an interesting ongoing discussion on hydrodynamic models for heavy traffic flow. In particular, some models using systems of two conservation laws have been proposed; see [3, 8, 10, 11]. We do no treat this aspect.

The paper is organized as follows. In section 2 we give the definition of weak entropic solution and, following rules (A) and (B), we introduce an admissibility condition at junctions. In section 3 we prove the existence and uniqueness of admissible solutions for the Riemann problem in a junction, then using this we describe the construction of the approximants for the Cauchy problem (see section 4). In section 5 we prove the bound on the total variation of the flux and existence of admissible solutions for the Cauchy problem with suitable initial data. In section 6 we prove with a counterexample that the Lipschitz continuous dependence with respect to initial data does not hold in general, but we also show that this property holds under special assumptions. In section 7 we describe what happens when there are traffic lights and time-dependent coefficients. Appendix A contains an example of flux variation increase that does not occur for junctions with only two incoming and two outgoing roads. Finally, in Appendix B we show that the interaction of a small wave with a junction can produce a uniformly big wave.

2. Basic definitions. We consider a network of roads that is modeled by a finite collection of intervals $I_i = [a_i, b_i] \subset \mathbb{R}$, $i = 1, \dots, N$, $a_i < b_i$, possibly with either $a_i = -\infty$ or $b_i = +\infty$, on which we consider (1.1). Hence the data are given by a finite collection of functions ρ_i defined on $[0, +\infty[\times I_i$.

On each road I_i we want ρ_i to be a weak entropic solution, that is, for every function $\varphi : [0, +\infty[\times I_i \to \mathbb{R}$ smooth with compact support on $]0, +\infty[\times]a_i, b_i[$

$$\int_0^{+\infty} \int_{a_i}^{b_i} \Big(\rho_i \frac{\partial \varphi}{\partial t} + f(\rho_i) \frac{\partial \varphi}{\partial x} \Big) dx dt = 0, \tag{2.1}$$

and for every $k \in \mathbb{R}$ and every $\tilde{\varphi} : [0, +\infty[\times I_i \to \mathbb{R}$ smooth, positive with compact support on $]0, +\infty[\times]a_i, b_i[$

$$\int_0^{+\infty} \int_{a_i}^{b_i} \Big(|\rho_i - k| \frac{\partial \tilde{\varphi}}{\partial t} + \operatorname{sgn}\,(\rho_i - k)(f(\rho_i) - f(k)) \frac{\partial \tilde{\varphi}}{\partial x} \Big) dx dt \geq 0. \tag{2.2}$$

It is well known that, for (1.1) on $\mathbb{R}$ and for all initial data in L^∞, there exists a unique weak entropic solution depending in a continuous way on the initial data in L^1_{loc}.

We assume that the roads are connected by some junctions. Each junction J is given by a finite number of incoming roads and a finite number of outgoing roads; thus we identify J with $((i_1, \dots, i_n), (j_1, \dots, j_m))$, where the first n-tuple indicates the set of incoming roads and the second m-tuple indicates the set of outgoing roads.

We assume that each road can be an incoming road for at most one junction and outgoing for at most one junction.

Hence the complete model is given by a couple $(\mathcal{I}, \mathcal{J})$, where $\mathcal{I} = \{I_i : i = 1, \dots, N\}$ is the collection of roads and $\mathcal{J}$ is the collection of junctions.

Fix a junction J with incoming roads, say $I_1, \dots, I_n$, and outgoing roads, say $I_{n+1}, \dots, I_{n+m}$. A weak solution at J is a collection of functions $\rho_l : [0, +\infty[\times I_l \to \mathbb{R}$, $l = 1, \dots, n+m$, such that

$$\sum_{l=0}^{n+m} \left(\int_0^{+\infty} \int_{a_l}^{b_l} \left(\rho_l \frac{\partial \varphi_l}{\partial t} + f(\rho_l) \frac{\partial \varphi_l}{\partial x} \right) dx dt \right) = 0, \tag{2.3}$$

for every φ_l, $l = 1, \dots, n+m$ smooth having compact support in $]0, +\infty[\times]a_l, b_l]$ for $l = 1, \dots, n$ (incoming roads) and in $]0, +\infty[\times [a_l, b_l[$ for $l = n+1, \dots, n+m$ (outgoing roads) that are also *smooth across the junction*, i.e.,

$$\varphi_i(\cdot, b_i) = \varphi_j(\cdot, a_j), \qquad \frac{\partial \varphi_i}{\partial x}(\cdot, b_i) = \frac{\partial \varphi_j}{\partial x}(\cdot, a_j), \qquad i = 1, \dots, n,\ j = n+1, \dots, n+m.$$

Remark 1. Let $\rho = (\rho_1, \dots, \rho_{n+m})$ be a weak solution at the junction such that each $x \to \rho_i(t, x)$ has bounded variation. We can deduce that ρ satisfies the *Rankine–Hugoniot condition* at the junction J, namely,

$$\sum_{i=1}^{n} f(\rho_i(t, b_i-)) = \sum_{j=n+1}^{n+m} f(\rho_j(t, a_j+)), \tag{2.4}$$

for almost every $t > 0$.

Rules (A) and (B) can be given explicitly only for solutions with bounded variation, as in the next definition.

DEFINITION 2.1. *Let $\rho = (\rho_1, \dots, \rho_{n+m})$ be such that $\rho_i(t, \cdot)$ is of bounded variation for every $t \geq 0$. Then ρ is an admissible weak solution of* (1.1) *related to the matrix A, satisfying* (1.3), *at the junction J if and only if the following properties hold:*

(i) *ρ is a weak solution at the junction J;*
(ii) *$f(\rho_j(\cdot, a_j+)) = \sum_{i=1}^{n} \alpha_{ji} f(\rho_i(\cdot, b_i-))$, for each $j = n+1, \dots, n+m$;*
(iii) *$\sum_{i=1}^{n} f(\rho_i(\cdot, b_i-))$ is maximum subject to* (ii).

For every road $I_i = [a_i, b_i]$, if $a_i > -\infty$ and I_i is not the outgoing road of any junction, or $b_i < +\infty$ and I_i is not the incoming road of any junction, then boundary data $\psi_i : [0, +\infty[\to \mathbb{R}$ are given. In this case we ask ρ_i to satisfy $\rho_i(t, a_i) = \psi_i(t)$ (or $\rho_i(t, b_i) = \psi_i(t)$) in the sense of [4]. The treatment of boundary data in the sense of [4] can be done in the same way as in [1, 2]; thus we treat the case without boundary data. All the stated results hold also for the case with boundary data with obvious modifications.

Our aim is to solve the Cauchy problem on $[0, +\infty[$ for given initial and boundary data as in next definition.

DEFINITION 2.2. *Given $\bar{\rho}_i : I_i \to \mathbb{R}$, $i = 1, \dots, N$, L^∞ functions, a collection of functions $\rho = (\rho_1, \dots, \rho_N)$, with $\rho_i : [0, +\infty[\times I_i \to \mathbb{R}$ continuous as functions from $[0, +\infty[$ into L^1_{loc}, is an admissible solution if ρ_i is a weak entropic solution to* (1.1) *on I_i, $\rho_i(0, x) = \bar{\rho}_i(x)$ a.e., at each junction ρ is a weak solution and is an admissible weak solution in case of bounded variation.*

On the flux f we make the following assumption:

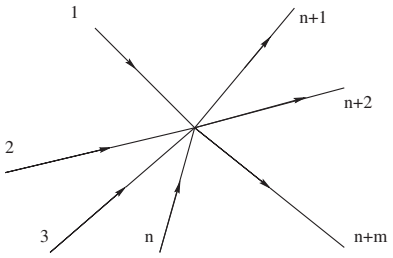

FIG. 1. *A junction.*

($\mathcal{F}$) $f : [0,1] \to \mathbb{R}$ is a smooth strictly concave function (i.e., $f'' \le -c < 0$ for some $c > 0$) such that $f(0) = f(1) = 0$. Therefore there exists a unique $\sigma \in]0,1[$ such that $f'(\sigma) = 0$ (that is, σ is a strict maximum).

3. The Riemann problem. In this section we study Riemann problems. For a scalar conservation law a Riemann problem is a Cauchy problem for initial data of Heaviside type, that is, piecewise constant with only one discontinuity. One looks for centered solutions, i.e., $\rho(t,x) = \phi(\frac{x}{t})$, which are the building blocks of solutions to the Cauchy problem via wave front tracking algorithm. These solutions are formed by continuous waves called rarefactions and by traveling discontinuities called shocks. The speeds of waves are related to the values of f'; see [5].

Analogously, we call the Riemann problem for a junction the Cauchy problem corresponding to an initial datum that is constant on each road. Then, for the whole network, since solutions on each road I_i can be constructed in the same way as for the scalar conservation law, it suffices to describe the solution at junctions. Because of finite propagation speed, it is enough to study the Riemann problem for a single junction.

Consider a junction J in which there are n roads with incoming traffic, m roads with outgoing traffic, and a traffic distribution matrix A. For simplicity we indicate by

$$(t,x) \in \mathbb{R}_+ \times I_i \mapsto \rho_i(t,x) \in [0,1], \quad i = 1, \dots, n, \tag{3.1}$$

the densities of the cars on the roads with incoming traffic and by

$$(t,x) \in \mathbb{R}_+ \times I_j \mapsto \rho_j(t,x) \in [0,1], \quad j = n+1, \dots, n+m \tag{3.2}$$

those on the roads with outgoing traffic; see Figure 1.

We need the following notation.

DEFINITION 3.1. *Let $\tau : [0,1] \to [0,1]$ be the map such that*

1. *$f(\tau(\rho)) = f(\rho)$ for every $\rho \in [0,1]$;*
2. *$\tau(\rho) \neq \rho$ for every $\rho \in [0,1] \setminus \{\sigma\}$.*

Clearly, τ is well defined and satisfies

$$0 \le \rho \le \sigma \Longleftrightarrow \sigma \le \tau(\rho) \le 1, \qquad \sigma \le \rho \le 1 \Longleftrightarrow 0 \le \tau(\rho) \le \sigma.$$

To state the main result of this section we need some assumption on the matrix A satisfied under generic conditions. Let $\{e_1, \dots, e_n\}$ be the canonical basis of $\mathbb{R}^n$ and, for every subset $V \subset \mathbb{R}^n$, indicate by $V^\perp$ its orthogonal. Define for every $i = 1, \dots, n$, $H_i = \{e_i\}^\perp$, i.e., the coordinate hyperplane orthogonal to e_i, and for

every $j = n+1, \dots, n+m$ let $\alpha_j = (\alpha_{j1}, \dots, \alpha_{jn}) \in \mathbb{R}^n$ and define $H_j = \{\alpha_j\}^\perp$. Let $\mathcal{K}$ be the set of indices $k = (k_1, \dots, k_\ell)$, $1 \le \ell \le n-1$, such that $0 \le k_1 < k_2 < \cdots < k_\ell \le n+m$, and for every $k \in \mathcal{K}$ set

$$H_k = \bigcap_{h=1}^{\ell} H_{k_h}.$$

Letting $\mathbf{1} = (1, \dots, 1) \in \mathbb{R}^n$, we assume

(C) for every $k \in \mathcal{K}$, $\mathbf{1} \notin H_k^\perp$.

Remark 2. Condition (C) is a technical condition, which allows us to have uniqueness in the maximization problem described in Theorem 3.2. From (C) we immediately derive $m \ge n$. Otherwise, since by definition, $\mathbf{1} = \sum_{j=n+1}^{n+m} \alpha_j$, we get $\mathbf{1} \in H_k^\perp$, where

$$H_k = \bigcap_{j=n+1}^{n+m} H_j.$$

Moreover if $n \ge 2$, then (C) implies that, for every $j \in \{n+1, \dots, n+m\}$ and for all distinct elements $i, i' \in \{1, \dots, n\}$, it holds that $\alpha_{ji} \ne \alpha_{ji'}$. Otherwise, without loss of generality, we may suppose that $\alpha_{n+1,1} = \alpha_{n+1,2}$. If we consider

$$H = \left(\bigcap_{2<j\le n} H_j \right) \bigcap H_{n+1},$$

then, by (C), there exists an element $(x_1, x_2, 0, \dots, 0) \in H$ such that $x_1 + x_2 \ne 0$ and $\alpha_{n+1,1}(x_1 + x_2) = 0$.

In the case of a simple junction J with two incoming roads and two outgoing ones, condition (C) is completely equivalent to the fact that, for every $j \in \{3, 4\}$, $\alpha_{j1} \ne \alpha_{j2}$.

Remark 3. Notice that the matrix A could have identical lines. For example the matrix

$$A = \begin{pmatrix} \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{3} & \frac{1}{2} & \frac{3}{5} \end{pmatrix}$$

satisfies condition (C).

THEOREM 3.2. *Consider a junction J, assume that the flux $f : [0,1] \to \mathbb{R}$ satisfies $(\mathcal{F})$, and that the matrix A satisfies condition* (C). *For every $\rho_{1,0}, \dots, \rho_{n+m,0} \in [0,1]$, there exists a unique admissible centered weak solution, in the sense of Definition* 2.1, *$\rho = (\rho_1, \dots, \rho_{n+m})$ of* (1.1) *at the junction J such that*

$$\rho_1(0, \cdot) \equiv \rho_{1,0}, \dots, \rho_{n+m}(0, \cdot) \equiv \rho_{n+m,0}.$$

Moreover, there exists a unique $(n+m)$-tuple $(\hat\rho_1, \dots, \hat\rho_{n+m}) \in [0,1]^{n+m}$ such that

$$\hat\rho_i \in \begin{cases} \{\rho_{i,0}\} \cup \,]\tau(\rho_{i,0}), 1] & \text{if } 0 \le \rho_{i,0} \le \sigma, \\ [\sigma, 1] & \text{if } \sigma \le \rho_{i,0} \le 1, \end{cases} \qquad i = 1, \dots, n, \tag{3.3}$$

and

$$\hat\rho_j \in \begin{cases} [0,\sigma] & \text{if } 0 \le \rho_{j,0} \le \sigma, \\ \{\rho_{j,0}\} \cup [0, \tau(\rho_{j,0})[& \text{if } \sigma \le \rho_{j,0} \le 1, \end{cases} \qquad j = n+1, \dots, n+m. \tag{3.4}$$

By fixed $i \in \{1, \dots, n\}$*, if* $\rho_{i,0} \le \hat\rho_i$*, we have*

$$\rho_i(t,x) = \begin{cases} \rho_{i,0} & \text{if } x < \frac{f(\hat\rho_i) - f(\rho_{i,0})}{\hat\rho_i - \rho_{i,0}} t + b_i,\ t \ge 0, \\ \hat\rho_i & \text{if } x > \frac{f(\hat\rho_i) - f(\rho_{i,0})}{\hat\rho_i - \rho_{i,0}} t + b_i,\ t \ge 0, \end{cases} \tag{3.5}$$

and, if $\hat\rho_i < \rho_{i,0}$*,*

$$\rho_i(t,x) = \begin{cases} \rho_{i,0} & \text{if } x \le f'(\rho_{i,0})t + b_i,\ t \ge 0, \\ \big(f'\big)^{-1}\big((x-b_i)/t\big) & \text{if } f'(\rho_{i,0})t + b_i \le x \le f'(\hat\rho_i)t + b_i,\ t \ge 0, \\ \hat\rho_i & \text{if } x > f'(\hat\rho_i)t + b_i,\ t \ge 0. \end{cases} \tag{3.6}$$

By fixed $j \in \{n+1, \dots, n+m\}$*, if* $\rho_{j,0} \le \hat\rho_j$*, we have*

$$\rho_j(t,x) = \begin{cases} \hat\rho_j & \text{if } x \le f'(\hat\rho_j)t + a_j,\ t \ge 0, \\ \big(f'\big)^{-1}\big((x-a_j)/t\big) & \text{if } f'(\hat\rho_j)t + a_j \le x \le f'(\rho_{j,0})t + a_j,\ t \ge 0, \\ \rho_{j,0} & \text{if } x > f'(\rho_{j,0})t + a_j,\ t \ge 0, \end{cases} \tag{3.7}$$

and, if $\hat\rho_j < \rho_{j,0}$*,*

$$\rho_j(t,x) = \begin{cases} \hat\rho_j & \text{if } x < \frac{f(\rho_{j,0}) - f(\hat\rho_j)}{\rho_{j,0} - \hat\rho_j} t + a_j,\ t \ge 0, \\ \rho_{j,0} & \text{if } x > \frac{f(\rho_{j,0}) - f(\hat\rho_j)}{\rho_{j,0} - \hat\rho_j} t + a_j,\ t \ge 0. \end{cases} \tag{3.8}$$

Proof. Define the map

$$E : (\gamma_1, \dots, \gamma_n) \in \mathbb{R}^n \longmapsto \sum_{i=1}^{n} \gamma_i$$

and the sets

$$\Omega_i \doteq \begin{cases} [0, f(\rho_{i,0})] & \text{if } 0 \le \rho_{i,0} \le \sigma, \\ [0, f(\sigma)] & \text{if } \sigma \le \rho_{i,0} \le 1, \end{cases} \qquad i = 1, \dots, n,$$

$$\Omega_j \doteq \begin{cases} [0, f(\sigma)] & \text{if } 0 \le \rho_{j,0} \le \sigma, \\ [0, f(\rho_{j,0})] & \text{if } \sigma \le \rho_{j,0} \le 1, \end{cases} \qquad j = n+1, \dots, n+m,$$

$$\Omega \doteq \Big\{ (\gamma_1, \dots, \gamma_n) \in \Omega_1 \times \cdots \times \Omega_n \big| A \cdot (\gamma_1, \dots, \gamma_n)^T \in \Omega_{n+1} \times \cdots \times \Omega_{n+m} \Big\}.$$

The set Ω is closed, convex, and not empty. Moreover, by (C), $\nabla E = \mathbf{1}$ is not orthogonal to any nontrivial subspace contained in a supporting hyperplane of Ω; hence there exists a unique vector $(\hat\gamma_1, \dots, \hat\gamma_n) \in \Omega$ such that

$$E(\hat\gamma_1, \dots, \hat\gamma_n) = \max_{(\gamma_1, \dots, \gamma_n) \in \Omega} E(\gamma_1, \dots, \gamma_n).$$

For every $i \in \{1, \dots, n\}$, we choose $\hat{\rho}_i \in [0,1]$ such that

$$f(\hat{\rho}_i) = \hat{\gamma}_i, \quad \hat{\rho}_i \in \begin{cases} \{\rho_{i,0}\} \cup]\tau(\rho_{i,0}), 1] & \text{if } 0 \le \rho_{i,0} \le \sigma, \\ [\sigma, 1] & \text{if } \sigma \le \rho_{i,0} \le 1. \end{cases}$$

By $(\mathcal{F})$, $\hat{\rho}_i$ exists and is unique. Let

$$\hat{\gamma}_j \doteq \sum_{i=1}^{n} \alpha_{ji} \hat{\gamma}_i, \qquad j = n+1, \dots, n+m,$$

and $\hat{\rho}_j \in [0,1]$ be such that

$$f(\hat{\rho}_j) = \hat{\gamma}_j, \quad \hat{\rho}_j \in \begin{cases} [0, \sigma] & \text{if } 0 \le \rho_{j,0} \le \sigma, \\ \{\rho_{j,0}\} \cup [0, \tau(\rho_{j,0})[& \text{if } \sigma \le \rho_{j,0} \le 1. \end{cases}$$

Since $(\hat{\gamma}_1, \dots, \hat{\gamma}_n) \in \Omega$, $\hat{\rho}_j$ exists and is unique for every $j \in \{n+1, \dots, n+m\}$. Solving the Riemann problem (see [5, Chap. 6]) on each road, the claim is proved. ∎

4. The wave front tracking algorithm. Once the solution to a Riemann problem is provided, we are able to construct piecewise constant approximations via a wave front tracking algorithm. The construction is very similar to that for scalar conservation laws (see [5]); hence we briefly describe it.

Let $\bar{\rho} = (\rho_1, \dots, \rho_N)$ be a piecewise constant map defined on the road network. We want to construct a weak solution of (1.1) with initial condition $\rho(0, \cdot) \equiv \bar{\rho}$. We begin by solving the Riemann problems on each road in correspondence to the jumps of $\bar{\rho}$ and the Riemann problems at junctions determined by the values of $\bar{\rho}$ (see Theorem 3.2). We split each rarefaction wave into a rarefaction fan formed by rarefaction shocks that are discontinuities traveling with the Rankine–Hugoniot speed. We always split rarefaction waves, inserting the value σ (if it is in the range of the rarefaction). Moreover, we let any rarefaction shock with endpoint σ have velocity zero.

When a wave interacts with another we simply solve the new Riemann problem. Instead, when a wave reaches a junction, we solve the Riemann problem at the junction. The number of waves may increase only for interactions of waves at junctions. Since the speeds of waves are bounded, there are finitely many waves on the network at each time $t \ge 0$. We call the obtained function *an approximate wave front tracking solution*. Given a general initial data, we approximate it by a sequence of piecewise constant functions and construct the corresponding approximate solutions. If they converge in L^1_{loc}, then the limit is a weak entropic solution on each road; see [5] for a proof.

5. Estimates on flux variation and existence of solutions. This section is devoted to the estimation of the total variation of the flux along an approximate wave front tracking solution and to the construction of solutions to the Cauchy problem. From now on, we assume that every junction has exactly two incoming roads and two outgoing ones. This hypothesis is crucial because, as shown in Appendix A, the presence of more complicated junctions causes additional increases in the total variation of the flux. The case where junctions have at most two incoming roads and at most two outgoing roads can be treated in the same way. So, for each junction J, the matrix A, defined in the introduction, takes the form

$$A = \begin{pmatrix} \alpha & \beta \\ 1-\alpha & 1-\beta \end{pmatrix}, \tag{5.1}$$

where $\alpha, \beta \in]0,1[$ and $\alpha \neq \beta$, so that (C) is satisfied.

From now on we fix an approximate wave front tracking solution ρ, defined on the road network.

DEFINITION 5.1. *For every road I_i, $i = 1, \dots, N$, we indicate by*

$$(\rho^\theta_-, \rho^\theta_+), \quad \theta \in \Theta = \Theta(\rho, t, i), \quad \Theta \text{ finite set},$$

the discontinuities on road I_i at time t, and by $x^\theta(t)$, $\lambda^\theta(t)$, $\theta \in \Theta$, respectively, their positions and velocities at time t. We also refer to the wave θ to indicate the discontinuity $(\rho^\theta_-, \rho^\theta_+)$.

For each discontinuity $(\rho^\theta_-, \rho^\theta_+)$ at time $\bar t$ on road I_i, we call $y^\theta(t)$, $t \in [\bar t, t_\theta]$, the trace of the wave so defined. We start with $y^\theta(\bar t) = x^\theta(\bar t)$ and we continue up to the first interaction with another wave or a junction. If at time $\tilde t$ an interaction with a wave or a junction occurs, then either a single new wave $(\rho^{\tilde\theta}_-, \rho^{\tilde\theta}_+)$ on road I_i is produced or no wave is produced. In the latter case we set $t_\theta = \tilde t$; otherwise we set $y^\theta(\tilde t) = x^{\tilde\theta}(\tilde t)$ and follow $x^{\tilde\theta}(t)$ for $t \geq t^{\tilde\theta}$ up to the next interaction, and so on.

We start by proving some technical lemmas.

LEMMA 5.2. *Fix a junction J and an incoming road I_i. Let θ be a wave on road I_i, originated at time $\bar t$ from J with a flux decrease, i.e., $x^\theta(\bar t) = b_i$, $\lambda^\theta(\bar t) < 0$, and $f(\rho^\theta_+) < f(\rho^\theta_-)$. Let y^θ be the traced wave and assume that there exists $\tilde t$, the first time y^θ interacts with J after $\bar t$. Then either y^θ interacts with another junction on $]\bar t, \tilde t[$ or, letting $\theta_1, \dots, \theta_l$ be the waves interacting with y^θ at times $t_m \in]\bar t, \tilde t[$, $m = 1, \dots, l$ $(t_1 < t_2 < \dots < t_l)$, we have*

$$\begin{aligned}&\left|f(\rho(\tilde t - \varepsilon, y^\theta(\tilde t - \varepsilon)+)) - f(\rho(\tilde t - \varepsilon, y^\theta(\tilde t - \varepsilon)-))\right| \\ &\leq \sum_{m=1}^{l} \left|f(\rho(t_m - \varepsilon, x^{\theta_m}(t_m - \varepsilon)+)) - f(\rho(t_m - \varepsilon, x^{\theta_m}(t_m - \varepsilon)-))\right| - \left|f(\rho^\theta_-) - f(\rho^\theta_+)\right|\end{aligned}$$

for $\varepsilon > 0$ small enough. This means that the initial flux variation along y^θ is canceled. The same conclusion holds for an outcoming road I_j.

Proof. Consider the wave $(\rho^\theta_-, \rho^\theta_+)$ as in the statement. Then it is a shock with negative velocity and $\rho^\theta_+ > \max\{\rho^\theta_-, \tau(\rho^\theta_-)\}$. If y^θ interacts with another junction, then there is nothing to prove. So, we assume that y^θ does not interact with another junction. At time t_1, the wave θ_1 interacts with y^θ. We analyze first the case of interaction from the left of y^θ. We have the following two possibilities:

1. $\rho^{\theta_1}_- \in [0, \tau(\rho^\theta_+)]$. In this case we have total cancellation of the flux variation and so
$$\left|f(\rho^\theta_+) - f(\rho^{\theta_1}_-)\right| = \left|f(\rho^{\theta_1}_-) - f(\rho^\theta_-)\right| - \left|f(\rho^\theta_-) - f(\rho^\theta_+)\right|.$$
Therefore the claim easily follows.
2. $\rho^{\theta_1}_- \in]\tau(\rho^\theta_+), \rho^\theta_+]$. In this case the wave y^θ after the time interaction t_1 is of the same type of y^θ before t_1, i.e.,
$$\max\{\rho(t_1, y^\theta(t_1)-), \tau(\rho(t_1, y^\theta(t_1)-))\} < \rho(t_1, y^\theta(t_1)+).$$

We consider now the case of interaction from the right of y^θ. It is clear that $\rho^{\theta_1}_+ \in]\rho^\theta_-, 1]$. If, moreover, $f(\rho^{\theta_1}_+) \geq f(\rho^\theta_-)$, then we have total cancellation of the flux and we conclude as before. If instead $f(\rho^{\theta_1}_+) < f(\rho^\theta_-)$, then the wave y^θ after the time t_1 is of the same type of y^θ before t_1.

We repeat this argument at each interaction time t_m. If at some t_m we have total cancellation of the flux, then we are done. Therefore we may suppose that at each t_m total cancellation of the flux does not occur. Since the type of the wave y^θ does not change, we have

$$\max\{\rho(\tilde{t}-\tilde{\varepsilon}, y^\theta(\tilde{t}-\tilde{\varepsilon})-), \tau(\rho(\tilde{t}-\tilde{\varepsilon}, y^\theta(\tilde{t}-\tilde{\varepsilon})-))\} < \rho(\tilde{t}-\tilde{\varepsilon}, y^\theta(\tilde{t}-\tilde{\varepsilon})+)$$

for $\tilde{\varepsilon} > 0$ small enough, and hence the speed $\lambda^\theta(\tilde{t}-\tilde{\varepsilon})$ is negative, which contradicts the fact that y^θ interacts with J at time $\tilde{t}$. □

LEMMA 5.3. *Fix a junction J and an incoming road I_i. Let θ be a wave on road I_i, originated at time $\bar{t}$ from J by a flux increase, i.e., $x^\theta(\bar{t}) = b_i$, $\lambda^\theta(\bar{t}) < 0$, and $f(\rho^\theta_+) > f(\rho^\theta_-)$. Let y^θ be the traced wave and assume that there exists $\tilde{t}$, the first time y^θ interacts with J after $\bar{t}$. Then y^θ interacts with other junctions in $]\bar{t}, \tilde{t}[$, or cancels the flux variation, or produces a flux decrease at J at $\tilde{t}$, i.e.,*

$$f(\rho(\tilde{t}-\varepsilon, y^\theta(\tilde{t}-\varepsilon)-)) < f(\rho(\tilde{t}-\varepsilon, y^\theta(\tilde{t}-\varepsilon)+)),$$

for $\varepsilon > 0$ small enough. The same holds for outgoing roads.

Proof. Since $\lambda^\theta(\bar{t}) < 0$ and $f(\rho^\theta_+) > f(\rho^\theta_-)$, then $\rho^\theta_- > \sigma$. Moreover, the wave $(\rho^\theta_-, \rho^\theta_+)$ is a rarefaction fan; hence $\sigma < \rho^\theta_+ < \rho^\theta_-$.

If an interaction on the right with a wave θ_1 occurs, then $\rho^{\theta_1}_+ \in]\rho^\theta_-, 1]$ and we have total cancellation of the flux variation. Therefore we may suppose that an interaction on the left with a wave θ_1 occurs. In this case we have two possibilities:

1. $\rho^{\theta_1}_- \in [0, \tau(\rho^\theta_+)[$;
2. $\rho^{\theta_1}_- \in [\tau(\rho^\theta_+), \rho^\theta_+[$.

In the latter case we have total cancellation of the flux variation and so we are done. In the first case, instead, the type of the wave changes, since

$$0 < \rho^{\theta_1}_- < \tau(\rho^\theta_+) \le \sigma \le \rho^\theta_+ < 1.$$

The speed of the wave y^θ after this interaction is positive and if there are no more interactions, then we have the claim since $f(\rho^{\theta_1}_-) < f(\rho^\theta_+)$. Thus we suppose that an interaction with a wave θ_2 occurs. If it is an interaction from the left, then the possibilities are as follows:

1. $\rho^{\theta_2}_- \in [0, \tau(\rho^\theta_+)[$. We do not have total cancellation of the flux variation, but the type of the wave does not change and the situation is identical to the previous one.
2. $\rho^{\theta_2}_- \in [\tau(\rho^\theta_+), \sigma[$. We have total cancellation of the flux variation and so we are done.

If it is an interaction from the right, then the possibilities are as follows:

1. $\rho^{\theta_2}_+ \in [\sigma, \tau(\rho^{\theta_1}_-)[$. We do not have total cancellation of the flux variation, but the type of the wave does not change.
2. $\rho^{\theta_2}_+ \in [\tau(\rho^{\theta_1}_-), 1]$. We have total cancellation of the flux variation and so we are done.

The conclusion now easily follows repeating this argument. If at each interaction we do not have total cancellation of the flux variation, then we necessarily have that

$$f(\rho(\tilde{t}-\varepsilon, y^\theta(\tilde{t}-\varepsilon)-)) < f(\rho(\tilde{t}-\varepsilon, y^\theta(\tilde{t}-\varepsilon)+))$$

for $\varepsilon > 0$ small enough, which concludes the proof. □

LEMMA 5.4. *Fix a junction J. If a wave interacts with the junction J from an incoming road at time $\bar t$, then*

$$\mathit{Tot.Var.}\big(f(\rho(\bar t+,\cdot))\big) = \mathit{Tot.Var.}\big(f(\rho(\bar t-,\cdot))\big). \tag{5.2}$$

Proof. For simplicity let us assume that I_1, I_2 are the incoming roads and I_3, I_4 are the outgoing ones. Let $(\rho_{1,0},\dots,\rho_{4,0})$ be an equilibrium configuration at the junction J. We assume that the wave is coming from the first road and that it is given by the values $(\rho_1,\rho_{1,0})$. Let us define the incoming flux

$$f^{in}(y) \doteq \begin{cases} f(y) & \text{if } 0 \le y \le \sigma, \\ f(\sigma) & \text{if } \sigma \le y \le 1, \end{cases} \tag{5.3}$$

and the outgoing flux

$$f^{out}(y) \doteq \begin{cases} f(\sigma) & \text{if } 0 \le y \le \sigma, \\ f(y) & \text{if } \sigma \le y \le 1. \end{cases} \tag{5.4}$$

Clearly, since the wave on the first road has positive velocity, we have

$$0 \le \rho_1 < \sigma. \tag{5.5}$$

Let $(\hat\rho_1,\dots,\hat\rho_4)$ be the solution of the Riemann problem in the junction J with initial data $(\rho_1,\rho_{2,0},\rho_{3,0},\rho_{4,0})$ (see Theorem 3.2). By definition, $\big(f(\rho_{1,0}),f(\rho_{2,0})\big)$ is the maximum point of the map E on the domain

$$\Omega_0 \doteq \Big\{(\gamma_1,\gamma_2) \in \Omega_{1,0}\times\Omega_{2,0}\big| A\cdot(\gamma_1,\gamma_2)^T \in \Omega_{3,0}\times\Omega_{4,0}\Big\},$$

and $\big(f(\hat\rho_1),f(\hat\rho_2)\big)$ is the maximum point of the map E on the domain

$$\hat\Omega \doteq \Big\{(\gamma_1,\gamma_2) \in \Omega_{1}\times\Omega_{2,0}\big| A\cdot(\gamma_1,\gamma_2)^T \in \Omega_{3,0}\times\Omega_{4,0}\Big\},$$

where

$$\Omega_{j,0} \doteq \begin{cases} [0,f^{in}(\rho_{j,0})] & \text{if } j=1,2, \\ [0,f^{out}(\rho_{j,0})] & \text{if } j=3,4, \end{cases}$$

and, by (5.5),

$$\Omega_1 \doteq [0,f^{in}(\rho_1)] = [0,f(\rho_1)].$$

It is also clear that

$$\big(f(\rho_{1,0}),f(\rho_{2,0})\big) \in \partial\Omega_0, \qquad \big(f(\hat\rho_1),f(\hat\rho_2)\big) \in \partial\hat\Omega.$$

For simplicity we use the notation (5.1).

We distinguish two cases. First we suppose that

$$f(\rho_1) < f(\rho_{1,0}) \tag{5.6}$$

(equality cannot happen in the previous equation because the wave would have velocity zero). Then $\hat\Omega \subset \Omega_0$ and

$$f(\hat\rho_1) \le f(\rho_1), \qquad f(\hat\rho_1)+f(\hat\rho_2) \le f(\rho_{1,0})+f(\rho_{2,0}). \tag{5.7}$$

We claim that

$$f(\rho_{2,0}) \le f(\hat\rho_2), \quad f(\hat\rho_3) \le f(\rho_{3,0}), \quad f(\hat\rho_4) \le f(\rho_{4,0}). \tag{5.8}$$

The points $\big(f(\rho_{1,0}), f(\rho_{2,0})\big)$, $\big(f(\hat\rho_1), f(\hat\rho_2)\big)$ are on the boundaries of Ω_0, $\hat\Omega$, respectively, where the function E attains the maximum; hence each one is at least on one of the curves

$$\alpha\gamma_1 + \beta\gamma_2 = f^{out}(\rho_{3,0}), \qquad (1-\alpha)\gamma_1 + (1-\beta)\gamma_2 = f^{out}(\rho_{4,0}), \qquad \gamma_2 = f^{in}(\rho_{2,0}).$$

Let us assume that the two points are on the same curve, with the other cases being similar:

$$\alpha\gamma_1 + \beta\gamma_2 = f^{out}(\rho_{3,0}). \tag{5.9}$$

Observe that the map E is increasing on the curve

$$\gamma_1 \mapsto \left(\gamma_1, \frac{f^{out}(\rho_{3,0})}{\beta} - \frac{\alpha}{\beta}\gamma_1\right);$$

otherwise we contradict the maximality of E at $\big(f(\rho_{1,0}), f(\rho_{2,0})\big)$. Thus $\alpha < \beta$, $\hat\rho_1 = \rho_1$, the first two inequalities in (5.8) hold, and

$$f(\hat\rho_1) = f(\rho_1), \qquad f(\hat\rho_2) > f(\rho_{2,0}), \qquad f(\hat\rho_3) = f(\rho_{3,0}) = f^{out}(\rho_{3,0}). \tag{5.10}$$

On the other hand, by (5.7), we have

$$\begin{aligned} f(\hat\rho_4) &= (1-\alpha)f(\hat\rho_1) + (1-\beta)f(\hat\rho_2) \\ &\le (1-\alpha)\big(f(\rho_{1,0}) + f(\rho_{2,0}) - f(\hat\rho_2)\big) + (1-\beta)f(\hat\rho_2) \\ &= (1-\alpha)\big(f(\rho_{1,0}) + f(\rho_{2,0})\big) + \big(\alpha-\beta\big)f(\hat\rho_2) \\ &\le (1-\alpha)\big(f(\rho_{1,0}) + f(\rho_{2,0})\big) + \big(\alpha-\beta\big)f(\rho_{2,0}) = f(\rho_{4,0}). \end{aligned}$$

Thus (5.8) holds. Using the Rankine–Hugoniot condition (2.4) at the junction J, and using (5.8) and (5.10), we get

$$\begin{aligned} &\text{Tot.Var.}\big(f(\rho(\bar t+,\cdot))\big) \\ &= |f(\hat\rho_1) - f(\rho_1)| + |f(\hat\rho_2) - f(\rho_{2,0})| + |f(\hat\rho_3) - f(\rho_{3,0})| + |f(\hat\rho_4) - f(\rho_{4,0})| \\ &= \big(f(\hat\rho_2) - f(\rho_{2,0})\big) + \big(f(\rho_{3,0}) - f(\hat\rho_3)\big) + \big(f(\rho_{4,0}) - f(\hat\rho_4)\big) \\ &= f(\rho_{1,0}) - f(\hat\rho_1) = f(\rho_{1,0}) - f(\rho_1) = \text{Tot.Var.}\big(f(\rho(\bar t-,\cdot))\big). \end{aligned}$$

Suppose now that

$$f(\rho_{1,0}) < f(\rho_1);$$

then $\rho_{1,0} < \rho_1 < \sigma$ and $\Omega_0 \subset \hat\Omega$. Assuming again that both maximum points of the function E are on the curve (5.9), we have

$$f(\hat\rho_1) = f(\rho_1), \quad f(\hat\rho_2) \le f(\rho_{2,0}), \quad f(\rho_{3,0}) = f(\hat\rho_3), \quad f(\rho_{4,0}) \le f(\hat\rho_4).$$

By the Rankine–Hugoniot condition at the junction J (see (2.4)), we have

$$\begin{aligned}&\text{Tot.Var.}\big(f(\rho(\bar{t}+,\cdot))\big)\\ &= |f(\hat{\rho}_1) - f(\rho_1)| + |f(\hat{\rho}_2) - f(\rho_{2,0})| + |f(\hat{\rho}_3) - f(\rho_{3,0})| + |f(\hat{\rho}_4) - f(\rho_{4,0})|\\ &= \big(f(\rho_{2,0}) - f(\hat{\rho}_2)\big) + \big(f(\hat{\rho}_3) - f(\rho_{3,0})\big) + \big(f(\hat{\rho}_4) - f(\rho_{4,0})\big)\\ &= f(\hat{\rho}_1) - f(\rho_{1,0}) = f(\rho_1) - f(\rho_{1,0}) = \text{Tot.Var.}\big(f(\rho(\bar{t}-,\cdot))\big).\end{aligned}$$

This concludes the proof. ∎

LEMMA 5.5. *Consider a network $(\mathcal{I},\mathcal{J})$. We have*

$$\textit{Tot. Var.}\big(f(\rho(0+,\cdot))\big) \le \textit{Tot. Var.}\big(f(\rho(0,\cdot))\big) + 2Rf(\sigma),$$

where R is the total number of roads of the network.

Proof. At time $t = 0$ we can have an instantaneous increase of the total variation of the flux due to the waves generated by the Riemann problems in the junctions. Clearly, this increase can be estimated by the maximum number of waves generated in the junctions ($\le 2R$) times the maximum variation of the flux on each road ($\le f(\sigma)$). ∎

We are now ready to prove the following.

LEMMA 5.6. *Consider a road network $(\mathcal{I},\mathcal{J})$. For some $K > 0$, we have*

$$\begin{aligned}\textit{Tot. Var.}\big(f(\rho(t+,\cdot))\big) &\le e^{Kt}\,\textit{Tot. Var.}\big(f(\rho(0+,\cdot))\big)\\ &\le e^{Kt}\big(\textit{Tot. Var.}\big(f(\rho(0,\cdot))\big) + 2Rf(\sigma)\big),\end{aligned}$$

for each $t \ge 0$.

Proof. Fix a junction J. Notice that there exists a constant C_J, depending on the coefficients of the matrix A at J, so that each interaction of a wave with J causes an increase in the flux variation by at most a factor of C_J. More precisely, if $\text{Tot.Var.}_f^{\pm}$ is the flux variation of waves before and after the interaction, then $\text{Tot.Var.}_f^{+} \le C_J \text{Tot.Var.}_f^{-}$.

Consider a wave θ interacting with the junction J. Then from Lemma 5.4 the flux variation can increase only if the wave is coming from an outgoing road. Let $\theta_1,\dots,\theta_4$ be the waves so produced. Thanks to Lemma 5.2, waves produced by a flux decrease cannot interact with the junction J without canceling the flux variation or reaching another junction. Moreover, by Lemma 5.3, every θ_i can return to junction J (without interacting with other junctions) only with a decrease of the flux. Now notice that a wave with decreasing flux interacting with J always produces a flux decrease on outgoing roads. Hence, waves θ_i may return to the junction only with decreasing flux, thus, by Lemma 5.2, producing other waves that cannot return to the junction, unless they cancel their flux variation or interact with other junctions. Finally, each wave flux variation can be magnified at most twice by a factor C_J interacting only with junction J and not with other junctions.

Now let δ be the minimum length of a road, i.e., $\delta = \min_{i\in\mathcal{I}}(b_i - a_i)$, and $\hat{\lambda}$ be the maximum speed of a wave, i.e., $\hat{\lambda} = \max\{f'(0), |f'(1)|\}$. Then each wave takes at least time $\delta/\hat{\lambda}$ to go from one junction to another.

Finally, recalling that the total variation of the flux may only decrease for interactions on roads, we get that a magnification of flux variation of a factor $C_{\mathcal{J}} = \max_{J\in\mathcal{J}} C_J^2$ may occur only once on each time interval of length $\delta/\hat{\lambda}$. We thus get

$$\begin{aligned}\text{Tot.Var.}\big(f(\rho(t+,\cdot))\big) &\le C_{\mathcal{J}}^{\frac{t\hat{\lambda}}{\delta}}\,\text{Tot.Var.}\big(f(\rho(0+,\cdot)\big)\\ &= e^{Kt}\text{Tot.Var.}\big(f(\rho(0+,\cdot))\big),\end{aligned}$$

where $K = \hat{\lambda}\log(C_{\mathcal{J}})/\delta$. ∎

DEFINITION 5.7. *Consider a road network $(\mathcal{I},\mathcal{J})$ and an approximate wave front tracking solution ρ. For every road I_i, we define two curves $Y_-^{i,\rho}(t)$, $Y_+^{i,\rho}(t)$, called* boundary of external flux (BEF)*, in the following way. We set the initial condition $Y_-^{i,\rho}(0) = a_i$, $Y_+^{i,\rho}(0) = b_i$ (if $a_i = -\infty$, then $Y_-^{i,\rho} \equiv -\infty$ and if $b_i = +\infty$, then $Y_+^{i,\rho} \equiv +\infty$). We let $Y_\pm^{i,\rho}(t)$ follow the generalized characteristic as defined in* [9]*, letting $Y_-^{i,\rho}(t) = a_i$ (resp., $Y_+^{i,\rho}(t) = b_i$) if the generalized characteristic reaches the boundary and $f'(\rho(t,a_i)) < 0$ (resp., $f'(\rho(t,b_i)) > 0$). (In this way $Y_\pm^{i,\rho}(t)$ may coincide with a_i or b_i for some time intervals.) Let $\bar{t}$ be the first time $\bar{t}$ such that $Y_-^{i,\rho}(\bar{t}) = Y_+^{i,\rho}(\bar{t})$ (possibly $\bar{t} = +\infty$); then we let $Y_\pm^{i,\rho}$ be defined on $[0,\bar{t}]$. Finally, we define the sets*

$$D_1^i(\rho) = \Big\{(t,x) : t \in [0,\bar{t}) : Y_-^{i,\rho}(t) \le x \le Y_+^{i,\rho}(t)\Big\}$$

and

$$D_2^i(\rho) = [0,+\infty) \times [a_i,b_i] \setminus D_1^i(\rho).$$

Clearly, $Y_\pm^i(t)$ bounds the set on which the data are not influenced by the other roads through the junctions.

DEFINITION 5.8. *Fix an approximate wave front tracking solution ρ, a road I_i, $i = 1,\dots,N$, and a junction J. A wave θ in I_i is said to be a big wave if*

$$sgn(\rho_-^\theta - \sigma)\cdot sgn(\rho_+^\theta - \sigma) \le 0,$$

where $sgn(0) = 0$.

We say that an incoming road I_i has a bad datum at J at time $t > 0$ if

$$\rho_i(t,b_i-) \in [0,\sigma[.$$

We say that an outgoing road I_j has a bad datum at J at time $t > 0$ if

$$\rho_j(t,a_j+) \in]\sigma,1].$$

LEMMA 5.9. *For every $t \ge 0$, there exist at most two big waves on*

$$\big\{x : (t,x) \in D_2^i(\rho)\big\} \subseteq [a_i,b_i].$$

Proof. A big wave can originate at time t on road I_i from J only if road I_i has a bad datum at J at time t. If this happens, then road I_i does not have a bad datum at J up to the time in which a big wave is absorbed from I_i. Then we reach the conclusion. ∎

THEOREM 5.10. *Fix a road network $(\mathcal{I},\mathcal{J})$. Given $C > 0$ and $T > 0$, there exists an admissible solution defined on $[0,T]$ for all initial data $\bar{\rho} \in \mathrm{cl}\{\rho : TV(\rho) \le C\}$, where* cl *indicates the closure in L^1_{loc}.*

Proof. We fix a sequence of initial data $\bar{\rho}_\nu$ piecewise constant such that $TV(\bar{\rho}_\nu) \le C$ for every $\nu \ge 0$ and $\bar{\rho}_\nu \to \bar{\rho}$ in L^1_{loc} as $\nu \to +\infty$. For each $\bar{\rho}_\nu$ we consider an approximate wave front tracking solution ρ_ν such that $\rho_\nu(0,x) = \bar{\rho}_\nu(x)$ and rarefactions are split in rarefaction shocks of size $\frac{1}{\nu}$.

For every road I_i, we notice that on $D_1^i(\rho_\nu)$, ρ_ν is not influenced by other roads and so the estimates of [5] hold. Since the curves $Y_\pm^{i,\rho_\nu}$ are uniformly Lipschitz continuous,

they converge, up to a subsequence, to a limit curve, and hence the regions $D_1^i(\rho_\nu)$ "converge" to a limit region D_1^i. Then $\rho_\nu \to \rho$ in L^1_{loc} on D_1^i with ρ an admissible solution to the Cauchy problem.

On $D_2^i := [0, +\infty[\times [a_i, b_i] \setminus D_1^i$, we have that, up to a subsequence, $\rho_\nu \rightharpoonup^* \rho$ weak* on L^1 and, using Theorem 2.4 of [5] and Lemma 5.6, $f(\rho_\nu) \to \bar{f}$ in L^1 for some $\bar{f}$. By Lemma 5.9, there are at most two big waves on D_2^i for every time, hence splitting the domain D_2^i in a finite number of pieces, where we can invert the function f, getting $\rho_\nu \to f^{-1}(\bar{f})$ in L^1. Together with $\rho_\nu \rightharpoonup^* \rho$ weak* on L^1, we conclude that $\rho_\nu \to \rho$ strongly in L^1.

The other requirements of the definition of admissible solution are clearly satisfied. □

6. Lipschitz continuous dependence: A counterexample and two special cases. In this section we assume that every junction has exactly two incoming roads and two outgoing ones and for every junction we follow the notation (5.1). We present a counterexample to the Lipschitz continuous dependence by initial data with respect to the L^1-norm. The continuous dependence by initial data with respect the L^1-norm remains an open problem. The counterexample is constructed using shifts of waves as in the spirit of [6, 7], to which we refer the reader for general theory.

We show that, for every $C > 0$, it is possible to choose two piecewise constant initial data, which are exactly the same except for a shift ξ of a discontinuity, such that the L^1-distance of the two corresponding solutions increases by the multiplicative factor C. Obviously, the L^1-distance of the initial data is finite and given by $|\xi\, \Delta\rho|$, where ξ is the shift and $\Delta\rho$ is the jump across the corresponding discontinuity. From now on, we consider a junction J, satisfying condition (C), with I_1, I_2 as incoming roads and I_3, I_4 as outgoing ones. Moreover, we suppose that the entries of the matrix A satisfy $\alpha < \beta$.

First we need some technical lemmas. The first one is well known; we report the proof for the reader's convenience.

LEMMA 6.1. *Let us consider in a road two waves, with speeds λ_1 and λ_2, respectively, that interact together at a certain time $\bar{t}$ producing a wave with speed λ_3. If the first wave is shifted by ξ_1 and the second wave by ξ_2, then the shift of the resulting wave is given by*

$$\xi_3 = \frac{\lambda_3 - \lambda_2}{\lambda_1 - \lambda_2}\xi_1 + \frac{\lambda_1 - \lambda_3}{\lambda_1 - \lambda_2}\xi_2. \tag{6.1}$$

Moreover, we have that

$$\Delta\rho_3\, \xi_3 = \Delta\rho_1\, \xi_1 + \Delta\rho_2\, \xi_2, \tag{6.2}$$

where $\Delta\rho_i$ are the signed strengths of the corresponding waves.

Proof. We suppose that ρ_l and ρ_m are the left and right values of the wave with speed λ_1 and ρ_m and ρ_r are the left and right values of the wave with speed λ_2; see Figure 2.

So $\Delta\rho_1 = \rho_m - \rho_l$, $\Delta\rho_2 = \rho_r - \rho_m$, and $\Delta\rho_3 = \rho_r - \rho_l$. The two wave fronts have, respectively,

$$x = \lambda_1 t + x_{1,0}, \qquad x = \lambda_2 t + x_{2,0},$$

where $x_{1,0}$ and $x_{2,0}$ are the initial positions of the wave fronts with speed λ_1 and λ_2,

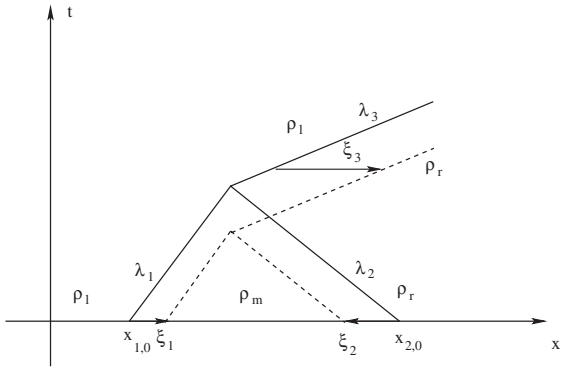

FIG. 2. *Shifts of waves.*

respectively. Therefore they interact at the point

$$(\bar{x},\bar{t}) = \left(\lambda_1 \frac{x_{1,0}-x_{2,0}}{\lambda_2-\lambda_1} + x_{1,0}, \frac{x_{1,0}-x_{2,0}}{\lambda_2-\lambda_1}\right).$$

If we consider the shifts, then the two wave fronts interact at the point

$$(\tilde{x},\tilde{t}) = \left(x_{1,0}+\xi_1+\lambda_1\frac{(x_{2,0}+\xi_2)-(x_{1,0}+\xi_1)}{\lambda_1-\lambda_2}, \frac{(x_{2,0}+\xi_2)-(x_{1,0}+\xi_1)}{\lambda_1-\lambda_2}\right),$$

and consequently (6.1) holds. Multiplying equation (6.1) by $\Delta\rho_3 = \Delta\rho_1 + \Delta\rho_2$, we easily deduce (6.2). ∎

LEMMA 6.2. *Let us consider a junction J with incoming roads I_1 and I_2 and outgoing roads I_3 and I_4. If a wave on a road I_i ($i \in \{1,\dots,4\}$) interacts with J without producing waves in the same road I_i and if ξ_i is the shift of the wave in I_i, then the shift ξ_j produced in a different road I_j ($j \in \{1,\dots,4\}\setminus\{i\}$) satisfies*

$$\xi_j\left(\rho_j^+ - \rho_j^-\right) = \frac{\Delta\gamma_j}{\Delta\gamma_i}\xi_i\left(\rho_i^+ - \rho_i^-\right), \tag{6.3}$$

where $\Delta\gamma_l$ ($l \in \{i,j\}$) represents the variation of the flux in the road I_l and ρ_l^-, ρ_l^+ ($l \in \{i,j\}$) are the states at J in the road I_l, respectively, before and after the interaction.

Proof. For simplicity let us consider the case $i = 1$ and $j = 3$, with the other cases being identical. Applying the shift ξ_1 to the wave (ρ_1^+, ρ_1^-), the interaction of the wave with J is shifted in time by

$$-\xi_1\frac{\rho_1^+-\rho_1^-}{f(\rho_1^+)-f(\rho_1^-)} = -\xi_1\frac{\rho_1^+-\rho_1^-}{\Delta\gamma_1}.$$

The shift in time in I_3 must be the same and so

$$\xi_1\frac{\rho_1^+-\rho_1^-}{\Delta\gamma_1} = \xi_3\frac{\rho_3^+-\rho_3^-}{\Delta\gamma_3},$$

which concludes the lemma. ∎

Remark 4. It is easy to understand that the coefficient of multiplication $\Delta\gamma_j/\Delta\gamma_i$ in the previous lemma depends by the entries of the matrix A. For example, under the same hypotheses of the previous lemma, if a wave in the I_1 road interacts with J producing a variation of the flux $\Delta\gamma_1$, and if no wave is produced in I_1 and I_2, then

$$\Delta\gamma_3 = \alpha\Delta\gamma_1, \qquad \Delta\gamma_4 = (1-\alpha)\Delta\gamma_1.$$

Consequently, in this case

$$\frac{\Delta\gamma_3}{\Delta\gamma_1} = \alpha, \quad \frac{\Delta\gamma_4}{\Delta\gamma_1} = 1-\alpha.$$

The following lemma is the first step needed to show that the Lipschitz dependence by initial data does not hold in our setting. More precisely, we show that there exists a simple configuration of waves and of shifts, which, after some interactions with J, produces an increase in the L^1-distance and takes a similar configuration.

LEMMA 6.3. *There exists an initial datum given by* $(\rho_{1,0}, \rho_{2,0}, \rho_{3,0}, \rho_{4,0})$ *that is an equilibrium configuration at* J*, a wave* $(\bar\rho_2, \rho_{2,0})$ *on road* I_2*, waves* $(\rho_{3,0}, \rho_3^*)$ *with shift* $\xi_{3,0}$*, and* $(\rho_3^*, \bar\rho_3)$ *on road* I_3 *such that the following occur in chronological order:*

1. *The initial distance in* L^1 *is* $\xi_{3,0}\,|\rho_{3,0} - \rho_3^*|$;
2. *the wave* $(\rho_{3,0}, \rho_3^*)$ *in* I_3 *with shift* $\xi_{3,0}$ *interacts with* J;
3. *waves are produced only in* I_2 *and* I_4;
4. *the wave on road* I_2 *interacts with* $(\bar\rho_2, \rho_{2,0})$ *producing a new wave*;
5. *the new wave from road* I_2 *interacts with* J;
6. *waves are produced only in* I_3 *and* I_4;
7. *in* I_4 *the* L^1*-distance after the interactions is equal to*

$$2\frac{1-\beta}{\beta}\,|\xi_{3,0}\,(\rho_3^* - \rho_{3,0})|\,,$$

and the L^1*-distance on road* I_3 *is equal to* $\xi_{3,0}\,|\rho_{3,0} - \rho_3^*|$.

Proof. Let $(\rho_{1,0}, \rho_{2,0}, \rho_{3,0}, \rho_{4,0})$ be an equilibrium configuration in J such that

$$0 < \rho_{1,0} < \sigma, \quad 0 < \rho_{2,0} < \sigma, \quad 0 < \rho_{3,0} < \sigma, \quad 0 < \rho_{4,0} < \sigma.$$

In road I_3, we consider a wave with negative speed $(\rho_{3,0}, \rho_3^*)$ with shift $\xi_{3,0}$. Since $(\rho_{3,0}, \rho_3^*)$ has negative speed, then $\rho_3^* > \tau(\rho_{3,0})$. Initially the L^1-distance of the two solutions is given by $|\xi_{3,0}(\rho_{3,0} - \rho_3^*)|$. When this wave interacts with J, new waves are produced in I_2 and I_4, which is possible, since $\alpha < \beta$. Therefore the new solution to the Riemann problem at J is given by

$$(\rho_{1,0}, \hat\rho_2, \hat\rho_3, \hat\rho_4)\,,$$

where $\tau(\rho_{2,0}) < \hat\rho_2 < 1$, $0 < \hat\rho_4 < \rho_{4,0}$. Moreover, some shifts $\hat\xi_2$ and $\hat\xi_4$ are produced in roads I_2 and I_4, respectively, where obviously $\hat\xi_2$ has the same sign of $\xi_{3,0}$ while $\hat\xi_4$ has opposite sign. By Lemma 6.2, we have

$$\begin{cases} \hat\xi_2(\hat\rho_2 - \rho_{2,0}) = \frac{1}{\beta}\xi_{3,0}(\rho_3^* - \rho_{3,0}), \\ \hat\xi_4(\hat\rho_4 - \rho_{4,0}) = \frac{1-\beta}{\beta}\xi_{3,0}(\rho_3^* - \rho_{3,0}). \end{cases}$$

If $0 < \bar\rho_2 < \tau(\hat\rho_2)$, then the wave $(\bar\rho_2, \rho_{2,0})$ in the road I_2 with shift $\bar\xi_2 = 0$ interacts with the wave $(\rho_{2,0}, \hat\rho_2)$ producing a wave $(\bar\rho_2, \hat\rho_2)$ with positive speed and with shift $\tilde\xi_2$. In this case,

$$\tilde\xi_2(\hat\rho_2 - \bar\rho_2) = \hat\xi_2(\hat\rho_2 - \rho_{2,0}) = \frac{1}{\beta}\xi_{3,0}(\rho_3^* - \rho_{3,0}).$$

Then, after the interaction of the wave $(\bar\rho_2, \hat\rho_2)$ with J, the new solution of the Riemann problem at J is given by

$$(\rho_{1,0}, \bar\rho_2, \hat\rho_3, \bar\rho_4)\,,$$

where $0 < \hat\rho_3 < \tau(\rho_3^*)$ and $0 < \bar\rho_4 < \hat\rho_4$. So in the roads I_3 and I_4 new shifts $\hat\xi_3$ and $\bar\xi_4$ are created, where

$$\begin{cases} \hat\xi_3(\rho_3^* - \hat\rho_3) = \beta\tilde\xi_2(\hat\rho_2 - \bar\rho_2) = \xi_{3,0}(\rho_3^* - \rho_{3,0}), \\ \bar\xi_4(\hat\rho_4 - \bar\rho_4) = (1-\beta)\tilde\xi_2(\hat\rho_2 - \bar\rho_2) = \frac{1-\beta}{\beta}\xi_{3,0}(\rho_3^* - \rho_{3,0}). \end{cases}$$

Now, if $\tau(\hat\rho_3) < \bar\rho_3 < 1$, then the wave $(\rho_3^*, \bar\rho_3)$ with shift $\bar\xi_3 = 0$ interacts in I_3 with the wave $(\hat\rho_3, \rho_3^*)$ producing a wave $(\hat\rho_3, \bar\rho_3)$ with negative speed and with shift $\tilde\xi_3$ such that

$$\tilde\xi_3(\bar\rho_3 - \hat\rho_3) = \hat\xi_3(\rho_3^* - \hat\rho_3) = \xi_{3,0}(\rho_3^* - \rho_{3,0}).$$

If the two waves on road I_4 do not interact, and this occurs when choosing appropriately the position of waves, then in the road I_4 the L^1-distance is

$$2\frac{1-\beta}{\beta}\left|\xi_{3,0}(\rho_3^* - \rho_{3,0})\right|,$$

and so the lemma is proved. □

Applying repeatedly Lemma 6.3, we produce a counterexample to the Lipschitz continuous dependence by initial data, as the next proposition shows.

PROPOSITION 6.4. *Let $C > 0$, J be a junction, and $(\rho_{1,0}, \dots, \rho_{4,0})$ be an equilibrium configuration as in Lemma* 6.3. *There exist two piecewise constant initial data satisfying the equilibrium configuration at J such that the L^1-distance between the corresponding two solutions increases by the multiplication factor C.*

Proof. Let n be big enough so that

$$\left(1 + 2n\frac{1-\beta}{\beta}\right) > C.$$

We want to define an initial data that provides the desired increase. We choose ρ_3^* and two finite sequences $(\bar\rho_2^i)$, $(\bar\rho_3^i)$, $i = 1, \dots, n$, so that, letting $\hat\rho_2^i$, $\hat\rho_3^i$ be the states determined as in Lemma 6.3, we have

$$\begin{cases} \rho_3^* \in\,]\tau(\rho_{3,0}), 1], \\ \bar\rho_2^i \in [0, \tau(\hat\rho_2^i)[, \qquad i = 1, \dots, n, \\ \bar\rho_3^i \in\,]\tau(\hat\rho_3^i), 1], \qquad i = 1, \dots, n. \end{cases}$$

It is easy to check that these sequences can be defined by induction.

The piecewise constant initial data in I_3 are given by

$$\begin{cases} \rho_{3,0} & \text{if } 0 < x < x^*, \\ \rho_3^* & \text{if } x^* < x < \hat{x}_1, \\ \hat{\rho}_3^1 & \text{if } \hat{x}_1 < x < \hat{x}_2, \\ \vdots & \dots \\ \bar{\rho}_3^n & \text{if } \tilde{x}_n < x, \end{cases}$$

where the values x^*, $\hat{x}_1, \dots, \hat{x}_n$ are to be determined in what follows. If $\xi_{3,0}$ denotes the shift of the wave $(\rho_{3,0}, \rho_3^*)$ and if no more shifts are present, then the L^1-distance of initial data are given by

$$|\xi_{3,0}| \, (\rho_3^* - \rho_{3,0}) \, .$$

The initial data on I_2 are

$$\begin{cases} \rho_{2,0} & \text{if } \tilde{x}_1 < x < 0, \\ \hat{\rho}_2^1 & \text{if } \tilde{x}_2 < x < \tilde{x}_1, \\ \vdots & \dots \\ \hat{\rho}_2^n & \text{if } x < \tilde{x}_n, \\ \vdots & \dots, \end{cases}$$

where $\tilde{x}_1, \dots, \tilde{x}_n$ are to be chosen appropriately.

The speed of the wave $(\rho_{3,0}, \rho_3^*)$ is given by the Rankine–Hugoniot condition

$$\frac{f(\rho_{3,0}) - f(\rho_3^*)}{\rho_{3,0} - \rho_3^*},$$

and consequently the time needed to go to junction J is

$$\bar{T} = -\frac{(\rho_{3,0} - \rho_3^*) \, x^*}{f(\rho_{3,0}) - f(\rho_3^*)}.$$

Clearly we adjust $\bar{T}$, choosing x^*. Applying n times Lemma 6.3 and adjusting the interaction times by choosing appropriately $\bar{x}_i$, $\tilde{x}_i$, $i \in \{1, \dots, n\}$, we can create $2n$ waves on road I_4 that do not interact together before the end of these n cycles and so we deduce that, at the end, the L^1-distance of the two solutions is given by

$$\left(1 + 2n \frac{1-\beta}{\beta}\right) |\xi_{3,0}(\rho_3^* - \rho_{3,0})| \, ,$$

which concludes the proof. □

Remark 5. The process described in the proof of Proposition 6.4 cannot be infinitely repeated. In fact, the sequences $\bar{\rho}_2^i$, $\bar{\rho}_3^i$ are monotonic and so $\bar{\rho}_3^{i+1} - \bar{\rho}_3^i \sim \frac{\bar{\rho}_3^1}{n}$ as n goes to infinity. Then the corresponding shifts on I_3 tend to infinity, letting waves interact with each other on road I_4. Therefore, with this method, it is not possible to produce a blow-up of the L^1-distance in finite time.

In some special cases the Lipschitz continuous dependence holds, as we show in the next subsections.

6.1. Network with only one junction. We consider a road network with only one junction J and with I_1, I_2 incoming roads and I_3, I_4 outgoing roads. We define

$$\mathcal{D} := \left\{ \bar\rho = (\bar\rho_1, \dots, \bar\rho_4) \in L^\infty(I_1 \times \cdots \times I_4) \cap L^1(I_1 \times \cdots \times I_4) : \bar\rho_j \in [0, \sigma], j = 3, 4 \right\}.$$

The following theorem holds.

THEOREM 6.5. *There exists a Lipschitz continuous semigroup $S : [0, +\infty[\times \mathcal{D} \to \mathcal{D}$ so that, for every $\bar\rho \in \mathcal{D}$, $\rho(t, x) = S(t, \bar\rho)(x)$ is an admissible solution with $\rho(0, x) = \bar\rho(x)$.*

Before proving the theorem, we consider the following lemma.

LEMMA 6.6. *Let $T > 0$ and let $\rho, \tilde\rho$ be two approximate wave front tracking solutions (AWFTS) connected by shifts such that $\rho(0, \cdot) \in \mathcal{D}$ and $\tilde\rho(0, \cdot) \in \mathcal{D}$. Then, for every $t \in [0, T]$, we have*

$$\|\rho(t, \cdot) - \tilde\rho(t, \cdot)\|_{L^1} = \sum_{\theta \in \Theta(t)} \left|\xi^\theta \Delta\rho^\theta\right| = \|\rho(0, \cdot) - \tilde\rho(0, \cdot)\|_{L^1},$$

where $\Theta(t)$ denotes the set of the jumps $\Delta\rho^\theta$ of $\rho(t, \cdot)$ with shifts ξ^θ.

Proof. We note first that $\mathcal{D}$ is invariant with respect to approximate wave front tracking solutions. Since $\rho_j \in [0, \sigma]$ for every $j \in \{3, 4\}$, each wave on I_3 and I_4 has positive speed and so shifts on outgoing roads cannot propagate themselves on other roads. The conclusion easily follows from Lemmas 6.2 and 5.4. ∎

Proof of Theorem 6.5. For every $T > 0$, by Theorem 5.10, a solution exists for every initial data in $\mathcal{D}$. By fixed $\rho, \tilde\rho \in \mathcal{D}$, we denote by $\rho_\nu, \tilde\rho_\nu$ two approximate wave front tracking solutions. As in [5, 6], to control the norm $\|\rho_\nu(t, \cdot) - \tilde\rho_\nu(t, \cdot)\|_{L^1}$, $t \in [0, T]$, it is enough to control the lengths of the shifts. Therefore, by Lemma 6.6, we obtain

$$\|\rho_\nu(t, \cdot) - \tilde\rho_\nu(t, \cdot)\|_{L^1} \le \|\rho_\nu(0, \cdot) - \tilde\rho_\nu(0, \cdot)\|_{L^1}$$

for every $t \in [0, T]$. Passing to the limit in the last expression, we finish the proof. ∎

6.2. Finite number of big waves, bad data, and interactions. Here we want to show a more general result about the Lipschitz continuity with respect to initial data. We omit the proof of this result, since it can be done with the same techniques as in the last subsection.

Let us consider a road network $(\mathcal{I}, \mathcal{J})$.

DEFINITION 6.7. *Let us fix an approximate wave front tracking solution ρ. For every junction J and for every incoming road I_i, the function $b_\rho(J, i, \cdot)$ is defined on $[0, T]$ by*

$$b_\rho(J, i, t) = \begin{cases} 0 & \text{if } \rho_i(t, b_i-) \in [\sigma, 1], \\ 1 & \text{if } \rho_i(t, b_i-) \in [0, \sigma[. \end{cases}$$

If ρ_ν is a sequence of AWFTS, then we say that the sequence ρ_ν has the property (H) if the following hold:

H1. There exists $M \in \mathbb{N}$ such that the function $b_{\rho_\nu}(J, i, \cdot)$ has at most M discontinuities for every $J \in \mathcal{J}$, for every $i \in \{1, \dots, N\}$, and for every $\nu \ge 0$.

H2. There exists $\delta > 0$ such that

$$|\rho_\nu(t, a_i+) - \sigma| > \delta$$

and

$$|\rho_\nu(t,b_i-)-\sigma|>\delta$$

for every $J\in\mathcal{J}$, for every $i\in\{1,\dots,N\}$, for every $\nu\geq 0$, and for every $t\in[0,T]$.

The following proposition holds.

PROPOSITION 6.8. *By fixed $T>0$, we consider a solution ρ defined on $[0,T]$ such that, for every $t\in[0,T]$, $\rho(t,\cdot)$ is a bounded variation function. Given $\eta>0$, $\delta>0$, and $M\in\mathbb{N}$, we define*

$$\begin{aligned}\mathcal{D}_\rho^\eta(\delta,M):=\{\bar\rho\in L^1_{loc}:\ &\exists(\rho_\nu)_{\nu\in\mathbb{N}}\ \textit{sequence of AWFTS satisfying (H)}\\ &\textit{with parameters } \delta \textit{ and } M,\\ &\rho_\nu(0,\cdot)\to\bar\rho(\cdot)\ \textit{in } L^1_{loc},\ \textit{Tot.Var.}(\rho_\nu(0,\cdot)-\rho(0,\cdot))<\eta\}.\end{aligned}$$

If there exist $0<\eta'<\eta$, $\delta>0$ and $M\in\mathbb{N}$ such that

$$\mathcal{D}:=\operatorname{cl}\{\tilde\rho:Tot.Var.(\rho-\tilde\rho)<\eta'\}\subseteq\mathcal{D}_\rho^\eta(\delta,M),$$

then there exists a Lipschitz continuous semigroup S of solutions defined on $[0,T]\times\mathcal{D}$.

Remark 6. We expect the existence of η,η',δ,M, as in Proposition 6.8, if we have $\eta<\delta$ and if we assume that big waves of ρ have velocity bounded away from zero.

7. Time-dependent traffic. In this section we consider a model of traffic including traffic lights and time-dependent traffic. The latter means that the choice of drivers at junctions may depend on the period of the day; for instance, during the morning the traffic flows toward some specific parts of the network and during the evening it may flow back. This means that the matrix A may depend on time t.

Consider a single junction J as in section 3 with two incoming roads I_1, I_2 and two outgoing ones I_3 and I_4. Let $\alpha=\alpha(t)$, $\beta=\beta(t)$ be two piecewise constant functions such that

$$0<\alpha(t)<1,\qquad 0<\beta(t)<1,\qquad \alpha(t)\neq\beta(t) \tag{7.1}$$

for each $t\geq 0$. Moreover, let $\chi_1=\chi_1(t),\chi_2=\chi_2(t)$ be piecewise constant maps such that

$$\chi_1(t)+\chi_2(t)=1,\qquad \chi_i(t)\in\{0,1\},\qquad i=1,2,$$

for each $t\geq 0$. The two maps represent traffic lights, with the value 0 corresponding to the red light and the value 1 to the green light.

DEFINITION 7.1. *Consider $\rho=\big(\rho_1,\dots,\rho_4\big)$ with bounded variation. We say that ρ is a solution at junction J if it satisfies* (i), (iii) *of Definition* 2.1 *and if the following property holds:*

(iv) $f(\rho_3(t,a_3+))=\alpha(t)\chi_1(t)f(\rho_1(t,b_1-))+\beta(t)\chi_2(t)f(\rho_2(t,b_2-))$ *and* $f(\rho_4(t,a_4+))=\big(1-\alpha(t)\big)\chi_1(t)f(\rho_1(t,b_1+))+\big(1-\beta(t)\big)\chi_2(t)f(\rho_2(t,b_2+))$ *for each $t>0$.*

The construction of the solution can be done as in section 5. However, the total variation of $f(\rho)$ does not depend continuously on the total variation of the maps $\alpha(\cdot)$, $\beta(\cdot)$. Indeed, let us suppose that there are no traffic lights, i.e., $\chi_i\equiv 1$, and let

$$\alpha(t)=\begin{cases}\eta_1 & \text{if } 0\leq t\leq\bar t,\\ \eta_2 & \text{if } \bar t\leq t\leq T,\end{cases}\qquad \beta(t)=\begin{cases}\eta_2 & \text{if } 0\leq t\leq\bar t,\\ \eta_1 & \text{if } \bar t\leq t\leq T,\end{cases}$$

where $0 < \eta_2 < \eta_1 < \frac{1}{2}$ and $0 < \bar{t} < T$. Consider the initial data $(\rho_{1,0}, \rho_{2,0}, \rho_{3,0}, \rho_{4,0})$, where

$$f(\rho_{1,0}) = f(\rho_{4,0}) = f(\sigma), \qquad f(\rho_{2,0}) = f(\rho_{3,0}) = \frac{\eta_1}{1-\eta_2} f(\sigma),$$

and

$$\sigma < \rho_{2,0} < 1, \qquad 0 < \rho_{3,0} < \sigma.$$

This is an equilibrium configuration and hence the solution of the Riemann problem for $0 \le t \le \bar{t}$. At time $t = \bar{t}$ we have to solve a new Riemann problem. Let $(\hat{\rho}_1, \hat{\rho}_2, \hat{\rho}_3, \hat{\rho}_4)$ be the new solution. We have

$$f(\hat{\rho}_2) = f(\hat{\rho}_4) = f(\sigma), \qquad f(\hat{\rho}_1) = f(\hat{\rho}_3) = \frac{\eta_1}{1-\eta_2} f(\sigma).$$

Now, if $\eta_1 \to \eta_2$, then

$$\text{Tot.Var.}\big(\alpha; [0,T]\big) \longrightarrow 0, \qquad \text{Tot.Var.}\big(\beta; [0,T]\big) \longrightarrow 0,$$

but

$$(f(\rho_{1,0}),\ f(\rho_{2,0})) \to \left(f(\sigma),\ \frac{\eta_2}{1-\eta_2} f(\sigma)\right), (f(\hat{\rho}_1),\ f(\hat{\rho}_2)) \to \left(\frac{\eta_2}{1-\eta_2} f(\sigma),\ f(\sigma)\right),$$

and hence $\text{Tot.Var.}\big(f(\rho); [0,T]\big)$ is bounded away from zero.

Appendix A. Total variation of the fluxes. Let J be a junction with three incoming roads and three outgoing ones. We show with an example that the total variation of the flux may increase if a wave arrives at J from an incoming road. Let us suppose that the matrix A is given by

$$A \doteq \begin{pmatrix} \frac{1}{2} - \varepsilon & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{2} + \varepsilon \\ \frac{1}{6} + \varepsilon & 0 & \frac{1}{6} - \varepsilon \end{pmatrix}, \tag{A.1}$$

with $\varepsilon > 0$. Notice that the matrix A satisfies condition (C) for every $\varepsilon > 0$ small enough.

Let us choose $\rho_1,\ \rho_{1,0}, \ldots,\ \rho_{6,0} \in [0,1]$ such that

$$\rho_{1,0} = \rho_{4,0} = \rho_{5,0} = \sigma, \quad \sigma < \rho_{2,0} < 1, \quad \sigma < \rho_{3,0} < 1, \quad 0 < \rho_{6,0} < \sigma, \quad 0 < \rho_1 < \sigma,$$

$$f(\rho_{2,0}) = \frac{1 + 36\varepsilon + 36\varepsilon^2}{3(1+6\varepsilon)}, \qquad f(\rho_{3,0}) = \frac{1-6\varepsilon}{1+6\varepsilon}, \qquad f(\rho_{6,0}) = \frac{1}{6} + \varepsilon + \frac{(1-6\varepsilon)^2}{6(1+6\varepsilon)}.$$

We assume $f(\sigma) = 1$. Then $(\rho_{1,0}, \ldots,\ \rho_{6,0})$ is an equilibrium configuration and ρ, given by

$$\rho_1(0,x) = \begin{cases} \rho_{1,0} & \text{if } x_1 \le x \le b_1, \\ \rho_1 & \text{if } x < x_1, \end{cases} \qquad \rho_i(0,\cdot) \equiv \rho_{i,0}, \quad i = 2, \ldots, 6,$$

is a solution (see Figure 3). Moreover the point $\big(f(\rho_{1,0}), f(\rho_{2,0}), f(\rho_{3,0})\big)$ is given by the intersection of the planes

$$\left(\frac{1}{2} - \varepsilon\right)\gamma_1 + \frac{1}{2}\gamma_2 + \frac{1}{3}\gamma_3 = 1, \qquad \frac{1}{3}\gamma_1 + \frac{1}{2}\gamma_2 + \left(\frac{1}{2} + \varepsilon\right)\gamma_3 = 1, \qquad \gamma_1 = 1.$$

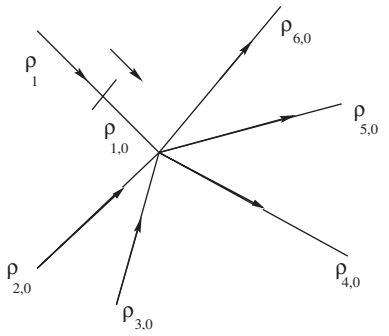

FIG. 3. *Configuration at J.*

At some time, say $\bar{t}$, the wave $(\rho_1, \rho_{1,0})$ interacts with the junction. Let $(\hat{\rho}_1, \dots, \hat{\rho}_6)$ be the solution of the Riemann problem at the junction for the data $(\rho_1, \rho_{2,0}, \dots, \rho_{6,0})$. If $f(\rho_1)$ is sufficiently near 1, then we have

$$\begin{aligned} f(\hat{\rho}_1) &= f(\rho_1), & f(\hat{\rho}_2) &= 2 - \frac{5 - 36\varepsilon^2}{3(1+6\varepsilon)} f(\rho_1), \\ f(\hat{\rho}_3) &= \frac{1-6\varepsilon}{1+6\varepsilon} f(\rho_1), & f(\hat{\rho}_4) &= f(\hat{\rho}_5) = 1, \\ f(\hat{\rho}_6) &= \frac{1+36\varepsilon^2}{3(1+6\varepsilon)} f(\rho_1). \end{aligned}$$

Therefore

$$\text{Tot.Var.}\big(f(\rho(\bar{t}-,\cdot))\big) = 1 - f(\rho_1)$$

and

$$\text{Tot.Var.}\big(f(\rho(\bar{t}+,\cdot))\big) = \frac{3(1-2\varepsilon)}{1+6\varepsilon}\big(1 - f(\rho_1)\big) > 2\,\text{Tot.Var.}\big(f(\rho(\bar{t}-,\cdot))\big).$$

Appendix B. Total variation of the densities. Consider a junction J with two incoming roads and two outgoing ones that we parameterize with the intervals $]-\infty, b_1]$, $]-\infty, b_2]$, $[a_3, +\infty[$, $[a_4, +\infty[$, respectively. We suppose that $0 < \beta < \alpha < 1/2$, where α and β are the entries of the matrix A as in (5.1).

Define a solution ρ by

(B.1)
$$\rho_1(0,x) = \begin{cases} \rho_{1,0} & \text{if } x_1 \le x \le b_1, \\ \rho_1 & \text{if } x < x_1, \end{cases} \quad \rho_2(0,x) = \rho_{2,0},\ \rho_3(0,x) = \rho_{3,0},\ \rho_4(0,x) = \rho_{4,0},$$

where ρ_1, $\rho_{1,0}$, $\rho_{2,0}$, $\rho_{3,0}$, $\rho_{4,0}$ are constants such that

(B.2)
$$\sigma < \rho_{2,0} < 1, \quad \sigma < \rho_{3,0} < 1, \quad 0 \le \rho_1 < \sigma, \quad \rho_{1,0} = \rho_{4,0} = \sigma,$$

$$f(\rho_{1,0}) = f(\rho_{4,0}) = f(\sigma), \quad f(\rho_{2,0}) = f(\rho_{3,0}) = \frac{\alpha}{1-\beta} f(\sigma),$$

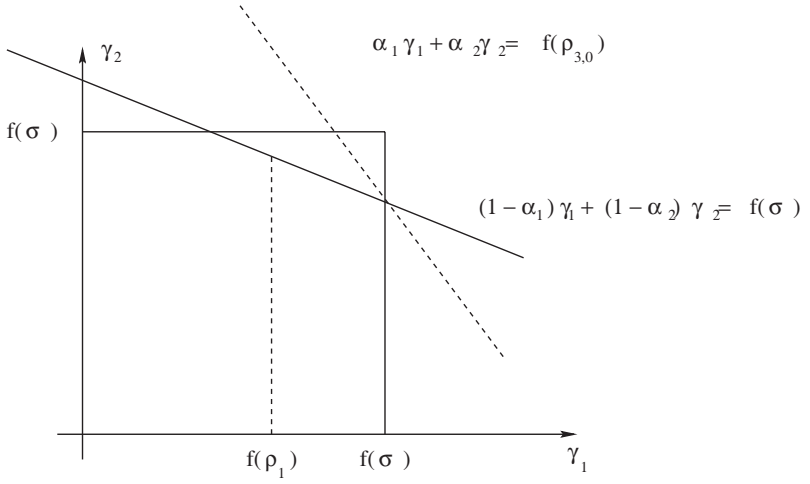

FIG. 4. *Solution to the Riemann problem at J.*

so $(\rho_{1,0},\ \rho_{2,0},\ \rho_{3,0},\ \rho_{4,0})$ is an equilibrium configuration.

After some time the wave $(\rho_1, \rho_{1,0})$ interacts with the junction. Let $(\hat{\rho}_1, \hat{\rho}_2, \hat{\rho}_3, \hat{\rho}_4)$ be the solution of the Riemann problem in the junction for the data $(\rho_1, \rho_{2,0}, \rho_{3,0}, \rho_{4,0})$; see Figure 4. By (B.1) and (B.2),

$$f(\hat{\rho}_1) = f(\rho_1), \quad f(\hat{\rho}_2) = \frac{f(\sigma) - (1-\alpha) f(\rho_1)}{1-\beta},$$

$$f(\hat{\rho}_3) = \frac{\alpha - \beta}{1-\beta} f(\rho_1) + \frac{\beta}{1-\beta} f(\sigma), \quad f(\hat{\rho}_4) = f(\sigma),$$

and

$$0 < \hat{\rho}_3 < \sigma \leq \hat{\rho}_2 < 1. \tag{B.3}$$

Therefore, if $\rho_1 \to \rho_{1,0} = \sigma$, then

$$f(\hat{\rho}_3) \longrightarrow \frac{\alpha}{1-\beta} f(\sigma) = f(\rho_{3,0}),$$

and, by (B.3) and (B.2), we have $\hat{\rho}_3 \to \tau(\rho_{3,0})$. Then, we are able to create on the third road a wave with strength bounded away from zero using an arbitrarily small wave on the first one.

Acknowledgments. The authors would like to thank Prof. Rinaldo M. Colombo for useful discussions.

REFERENCES

[1] D. AMADORI, *Initial-boundary value problems for systems of conservation laws*, NoDEA Nonlinear Differential Equations Appl., 4 (1997), pp. 1–42.

[2] F. ANCONA AND A. MARSON, *Scalar non-linear conservation laws with integrable boundary data*, Nonlinear Anal., 35 (1999), pp. 687–710.

[3] A. Aw and M. Rascle, *Resurrection of "second order" models of traffic flow*, SIAM J. Appl. Math., 60 (2000), pp. 916–938.
[4] C. Bardos, A. Y. Le Roux, and J. C. Nédélec, *First order quasilinear equations with boundary conditions*, Comm. Partial Differential Equations, 4 (1979), pp. 1017–1034.
[5] A. Bressan, *Hyperbolic Systems of Conservation Laws. The One-dimensional Cauchy Problem*, Oxford University Press, Oxford, UK, 2000.
[6] A. Bressan, G. Crasta, and B. Piccoli, *Well-posedness of the Cauchy Problem for $n \times n$ Systems of Conservation Law*, Mem. Amer. Math. Soc. 146, no. 694, AMS, Providence, RI, 2000.
[7] A. Bressan and A. Marson, *A variational calculus for discontinuous solutions of systems of conservation laws*, Comm. Partial Differential Equations, 20 (1995), pp. 1491–1552.
[8] R. M. Colombo, *Hyperbolic phase transitions in traffic flow*, SIAM J. Appl. Math., 63 (2002), pp. 708–721.
[9] C. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Springer-Verlag, Berlin, 2000.
[10] J. M. Greenberg, *Extension and amplifications of a traffic model of Aw and Rascle*, SIAM J. Appl. Math., 62 (2001), pp. 729–745.
[11] J. M. Greenberg, A. Klar, and M. Rascle, *Congestion on multilane highways*, SIAM J. Appl. Math., 63 (2003), pp. 818–833.
[12] H. Holden and N. H. Risebro, *A mathematical model of traffic flow on a network of unidirectional roads*, SIAM J. Math. Anal., 26 (1995), pp. 999–1017.
[13] C. Klingenberg and N. H. Risebro, *Convex conservation laws with discontinuous coefficients. Existence, uniqueness and asymptotic behavior*, Comm. Partial Differential Equations, 20 (1995), pp. 1959–1990.
[14] M. J. Lighthill and G. B. Whitham, *On kinetic waves. II. Theory of traffic flows on long crowded roads*, Proc. Roy. Soc. London Ser. A, 229 (1955), pp. 317–345.
[15] P. I. Richards, *Shock waves on the highway*, Oper. Res., 4 (1956), pp. 42–51.
[16] B. Temple, *Global solutions of the Cauchy problem for a class of 2×2 nonstrictly hyperbolic conservation laws*, Adv. Appl. Math., 3 (1982), pp. 335–375.
[17] J. D. Towers, *A difference scheme for conservation laws with a discontinuous flux: The nonconvex case*, SIAM J. Numer. Anal., 39 (2001), pp. 1197–1218.

SIAM J. MATH. ANAL.
Vol. 36, No. 6, pp. 1887–1928

SIZE EFFECTS ON QUASI-STATIC GROWTH OF CRACKS*

ALESSANDRO GIACOMINI†

Abstract. We perform an analysis of the size effect for quasi-static growth of cracks in isotropic linearly elastic bodies under antiplanar shear. In the framework of the variational model proposed by Francfort and Marigo in [*J. Mech. Phys. Solids*, 46 (1998), pp. 1319–1342], we prove that if the size of the body tends to infinity, and even if the surface energy is of cohesive form, under suitable boundary displacements the crack propagates following the Griffith's functional.

Key words. variational models, energy minimization, free discontinuity problems, crack propagation, quasi-static evolution, brittle fracture

AMS subject classifications. 35R35, 35J85, 35J25, 74R10

DOI. 10.1137/S0036141004439362

1. Introduction.

A well-known fact in fracture mechanics is that *ductility* is also influenced by the size of the structure, and in particular the structure tends to become brittle if its size increases (see, for example, [8] and the references therein). The aim of this paper is to capture this fact for the problem of quasi-static growth of cracks in linearly elastic bodies in the framework of the variational theory of crack propagation formulated by Francfort and Marigo in [14].

The model proposed in [14] is inspired by the classical Griffith's criterion (see [16], [17], [18]) and it determines the evolution of the crack through a competition between volume and surface energies. Let us illustrate the theory and the variant we investigate in the case of *generalized antiplanar shear*.

Let $\Omega \subseteq \mathbb{R}^N$ be open, connected, bounded, and with Lipschitz boundary. A crack $\Gamma \subseteq \overline{\Omega}$ is any rectifiable set, and a displacement u is any function defined almost everywhere in Ω whose set of discontinuities $S(u)$ is contained in Γ (we will make precise the functional setting later). The total energy of the configuration (u, Γ) is given by

$$\int_{\Omega\setminus\Gamma} |\nabla u|^2 \, dx + \mathcal{H}^{N-1}(\Gamma). \tag{1.1}$$

The first term in (1.1) implies that we assume to apply linearized elasticity in the unbroken part of Ω. The second term can be considered as the work done to create Γ.

As suggested in [14], more general surface energies can be considered in (1.1), especially those of Barenblatt's type [5], and here we consider energies of the form

$$\int_{\Gamma} \varphi(|[u]|(x)) \, d\mathcal{H}^{N-1}(x), \tag{1.2}$$

where $[u](x) := u^+(x) - u^-(x)$ is the difference of the traces of u on both sides of Γ, and $\varphi : [0, +\infty[\to [0, +\infty[$ (which depends on the material) is such that $\varphi(0) = 0$.

*Received by the editors January 7, 2004; accepted for publication (in revised form) July 25, 2004; published electronically June 16, 2005.
http://www.siam.org/journals/sima/36-6/43936.html
†Dipartimento di Matematica, Università di Brescia, Via Valotti 9, 25133 Brescia, Italy (alessandro.giacomini@ing.unibs.it).

In order to get a physical interpretation of (1.2), let us set $\sigma := \varphi'$: we interpret $\sigma(|[u]|(x))$ as density of forces in x that act between the two lips of the crack Γ whose displacements are $u^+(x)$ and $u^-(x)$, respectively. Typically σ is decreasing, and $\sigma(s) = 0$ for $s \geq \bar{s}$: this means that the interaction between the two lips of the crack decreases as the opening increases, and disappears when the opening is greater than a critical length $\bar{s}$. As a consequence, φ is increasing and concave, and $\varphi(s)$ is constant for $s \geq \bar{s}$. We will then consider φ increasing, concave, with $\varphi(0) = 0$, $a = \varphi'(0) < +\infty$, and $\lim_{s\to+\infty} \varphi(s) = 1$. We can interpret

$$\int_\Gamma \varphi(|[u]|(x))\, d\mathcal{H}^{N-1}(x)$$

as the work made to create Γ with an opening given by $[u]$. Assuming linearized elasticity to hold in $\Omega \setminus \Gamma$, we consider a total energy of the form

$$\|\nabla u\|^2 + \int_\Gamma \varphi(|[u]|)\, d\mathcal{H}^{N-1}, \tag{1.3}$$

where $\|\cdot\|$ denotes the L^2 norm. The problem of irreversible quasi-static growth of cracks in the cohesive case can be addressed through a *time discretization process* in analogy to what is proposed in [14] for the energy (1.1).

Let $g(t)$ be a time-dependent boundary displacement defined on $\partial_D\Omega \subseteq \partial\Omega$ with $t \in [0,T]$. Let $\delta > 0$ and let $I_\delta := \{0 = t_0^\delta < t_1^\delta < \cdots < t_{N_\delta}^\delta = T\}$ be a subdivision of $[0,T]$ with $\max(t_{i+1}^\delta - t_i^\delta) < \delta$, and let $g_i^\delta := g(t_i^\delta)$. At time $t = 0$ we consider u_0^δ as a minimum of

$$\|\nabla u\|^2 + \int_{S^{g(0)}(u)} \varphi(|[u]|)\, d\mathcal{H}^{N-1}. \tag{1.4}$$

Here $S^{g(0)}(u) := S(u) \cup \{x \in \partial_D\Omega : u(x) \neq g(0)(x)\}$, and for all $x \in \partial_D\Omega$ we consider $[u](x) := g(x) - \tilde{u}(x)$, where $\tilde{u}$ is the trace of u on $\partial\Omega$. We define the crack Γ_0^δ at time $t = 0$ as $S^{g(0)}(u_0^\delta)$. We also set $\psi_0^\delta := |[u_0^\delta]|$ on Γ_0^δ. The presence of $S^{g(0)}(u)$ in (1.4) indicates that the points at which the boundary displacement is not attained are considered as a part of the crack.

Supposing to have constructed Γ_i^δ and ψ_i^δ at time t_i^δ, we consider a minimum u_{i+1}^δ of the problem

$$\|\nabla u\|^2 + \int_{S^{g_{i+1}^\delta}(u)\cup\Gamma_i^\delta} \varphi(|[u]| \vee \psi_i^\delta)\, d\mathcal{H}^{N-1}, \tag{1.5}$$

where $|[u]| \vee \psi_i^\delta := \max\{|[u]|, \psi_i^\delta\}$, and define $\Gamma_{i+1}^\delta := \Gamma_i^\delta \cup S^{g_{i+1}^\delta}(u_{i+1}^\delta)$ and $\psi_{i+1}^\delta := \psi_i^\delta \vee |[u_{i+1}^\delta]|$ on Γ_{i+1}^δ.

Notice that problem (1.5) takes into account an *irreversibility condition* in the growth of the crack. Indeed, while on $S^{g_{i+1}^\delta}(u) \setminus \Gamma_i^\delta$ the surface energy which comes in minimization of (1.5) is exactly as in (1.2), on $S^{g_{i+1}^\delta}(u) \cap \Gamma_i^\delta$ the surface energy involved takes into account the previous work made on Γ_i^δ. The surface energy is of the form of (1.2) only if $|[u]| > \psi_i^\delta$, that is, only if the opening is increased. If $|[u]| \leq \psi_i^\delta$, no energy is gained, that is, displacements of this form along the crack are in a sense surface-energy-free. Notice finally that the irreversibility condition involves only the modulus of $[u]$: this is an assumption which is reasonable since we are considering only antiplanar displacements. Clearly more complex irreversibility conditions can be

formulated, involving, for example, a partial release of energy: the one we study is the first straightforward extension of the irreversibility condition given in [14] for the energy (1.1).

The *discrete-in-time evolution* of the crack relative to the subdivision I_δ, and the boundary datum $g(t)$ is given by $\{(u_i^\delta, \Gamma_i^\delta, \psi_i^\delta) : i = 0, \dots, N_\delta\}$.

The *irreversible quasi-static evolution* of the crack relative to the boundary datum $g(t)$ is obtained as a limit for $\delta \to 0$ of $(u^\delta(t), \Gamma^\delta(t), \psi^\delta(t))$, where $u^\delta(t) := u_i^\delta$, $\Gamma^\delta(t) := \Gamma_i^\delta$, and $\psi^\delta(t) := \psi_i^\delta$ for $t_i^\delta \le t < t_{i+1}^\delta$.

This program has been studied in detail in several papers in the case $\varphi \equiv 1$, that is, for energy of the form (1.1). A first mathematical formulation has been given by Dal Maso and Toader in [12], where the authors consider the case of dimension $N = 2$ and cracks which are compact and with a uniform bound on the number of connected components. This analysis has been extended to the case of plane elasticity by Chambolle in [9]. In [13] Francfort and Larsen consider the general N-dimensional case, and remove the bound on the number of the connected components of the cracks: the key point is to introduce a weak formulation of the problem considering displacements in the space SBV (see section 2). Finally Dal Maso, Francfort, and Toader [11] treat the case of finite elasticity not restricted to antiplanar shear, with volume energy depending on the full gradient under suitable growth conditions, and in presence of body and traction forces: the appropriate functional space for the displacements is now $GSBV$ (see, for example, [4] for a precise definition).

In all these papers (see [12], [9], [13], [11]), the analysis of the limit reveals three basic properties (irreversibility, static equilibrium, and nondissipativity; see Theorem 2.2) which are taken as definition of irreversible quasi-static growth in brittle fractures: the time discretization procedure is considered as a privileged way to get an existence result.

In the case of energy (1.3), several difficulties arise in the analysis of the discrete-in-time evolution, and in the analysis as $\delta \to 0$. In section 3, we prove that the functional space we need in order to apply the direct method of the Calculus of Variations in the step-by-step minimizations (1.4), (1.5) is the space of functions with bounded variation BV (see section 2): we thus consider a relaxed version of the problems, namely,

$$\int_\Omega f(\nabla u)\, dx + \int_\Gamma \varphi(|[u]| \vee \psi)\, d\mathcal{H}^{N-1} + a|D^c u|(\Omega),$$

where $a = \varphi'(0)$, f is defined in (3.3), and $D^c u$ indicates the Cantorian part of the derivative of u. An existence result for *discrete-in-time evolution* in this context of BV space is given in Proposition 3.1.

The analysis for $\delta \to 0$ presents several difficulties, the main one being the stability of the minimality property of the discrete-in-time evolutions. The main purpose of this paper is to prove that these difficulties disappear as the size of the reference configuration increases, thanks to the fact that the body response tends to become more and more brittle in spite of the presence of cohesive forces on the cracks. More precisely, we prove this fact for the discrete evolutions in $\Omega_h := h\Omega$ for h large and under suitable boundary displacements. The idea is to rescale displacements and cracks to the fixed configuration Ω, and take advantage from the form of the problem in this new setting. The boundary displacements on $\partial_D \Omega_h := h\partial_D \Omega$ will be taken of the form

$$g_h(t,x) := h^\alpha g\left(t, \frac{x}{h}\right), \qquad g \in AC([0,T]; H^1(\Omega)),$$
$$\|g(t)\|_\infty \le C, \qquad t \in [0,T],\ x \in \Omega_h,$$

where $\alpha > 0$ and $C > 0$. We indicate by $(u^{\delta,h}(t), \Gamma^{\delta,h}(t), \psi^{\delta,h}(t))$ the piecewise constant interpolation of the discrete-in-time evolution of the crack in Ω_h relative to the boundary displacement g_h. Let us moreover set for every $t \in [0,T]$

$$\mathcal{E}^{\delta,h}(t) := \int_{h\Omega} f(\nabla u^{\delta,h}(t))\,dx + \int_{\Gamma^{\delta,h}(t)} \varphi(\psi^{\delta,h}(t))\,d\mathcal{H}^{N-1} + a|D^c u^{\delta,h}(t)|(\Omega_h).$$

In the case $\alpha = \frac{1}{2}$, we make the following rescaling:

$$v^{\delta,h}(t,x) := \frac{1}{\sqrt{h}} u^{\delta,h}(t,hx), \qquad K^{\delta,h}(t) := \frac{1}{h}\Gamma^{\delta,h}(t), \qquad \gamma^{\delta,h}(t) := \frac{1}{\sqrt{h}}\psi^{\delta,h}(t,hx),$$

where $t \in [0,T]$ and $x \in \Omega$. The main result of the paper is the following (see Theorem 4.1 for a more precise statement).

THEOREM 1.1. *If $\delta \to 0$ and $h \to +\infty$, there exists a quasi-static crack evolution $\{t \to (v(t), K(t))\}$ in Ω relative to the boundary displacement g in the sense of* [13] *(see Theorem* 2.2*) such that for all $t \in [0,T]$ we have*

$$\nabla v^{\delta,h}(t) \rightharpoonup \nabla v(t) \qquad \text{weakly in } L^1(\Omega; \mathbb{R}^N).$$

Moreover, for all $t \in [0,T]$ we have

$$\frac{1}{h^{N-1}}\mathcal{E}^{\delta,h}(t) \to \|\nabla v(t)\|^2 + \mathcal{H}^{N-1}(K(t));$$

in particular $h^{-N+1}|D^c u^{\delta,h}(t)|(\Omega_h) \to 0$,

$$\frac{1}{h^{N-1}}\int_{\Omega_h} f(\nabla u^{\delta,h}(t))\,dx \to \|\nabla v(t)\|^2,$$

and

$$\frac{1}{h^{N-1}}\int_{\Gamma^{\delta,h}(t)} \varphi(\psi^{\delta,h}(t))\,d\mathcal{H}^{N-1} \to \mathcal{H}^{N-1}(K(t)).$$

Theorem 1.1 proves that as the size of the reference configuration increases, the response of the body in the problem of quasi-static growth of cracks tends to become brittle, so that energy (1.1) can be considered. Moreover, we have convergence results for the volume and surface energies involved.

The particular value $\alpha = \frac{1}{2}$ comes out because a problem of quasi-static evolution has been considered. In fact if we consider an infinite plane with a crack segment of length l and subject to a uniform stress σ at infinity, following Griffith's theory, the crack propagates quasi-statically if $\sigma = \frac{K_{IC}}{\sqrt{\pi l}}$, where K_{IC} is the critical *stress intensity factor* (depending on the material). So if the crack has length hl, the stress rescales as $\frac{1}{\sqrt{h}}$. This is precisely what we are prescribing in the case $\alpha = \frac{1}{2}$: in fact the stress that intuitively we prescribe at the boundary can be reconstructed from ∇u_h and rescales precisely as $\frac{1}{\sqrt{h}}$.

For the proof of Theorem 1.1, the first step is to recognize that

$$(v^{\delta,h}(t), K^{\delta,h}(t), \gamma^{\delta,h}(t))$$

is a discrete-in-time evolution relative to the boundary displacement g for a total energy of the form

$$\int_\Omega f_h(\nabla u)\,dx + \int_\Gamma \varphi_h(|[u]| \vee \gamma)\,d\mathcal{H}^{N-1} + a\sqrt{h}|D^c u|(\Omega),$$

where $\varphi_h(s) \nearrow 1$ for all $s \in [0, +\infty[$, and $f_h(\xi) \nearrow |\xi|^2$ for all $\xi \in \mathbb{R}^N$. From the fact that $\varphi_h \nearrow 1$ we recognize that the structure tends to become brittle. Bound on total energy for the discrete-in-time evolution is available, so that compactness in the space BV can be applied: it turns out that the limits of the displacements are of class SBV with gradient in $L^2(\Omega; \mathbb{R}^N)$. Limits for the cracks are constructed through a Γ-convergence procedure (see Proposition 5.2). The main point in order to see that $(v(t), K(t))$ is a quasi-static crack growth is to recover the static equilibrium condition (see point (c) of Theorem 2.2)

$$\|\nabla v(t)\|^2 \le \|\nabla v\|^2 + \mathcal{H}^{N-1}(S^{g(t)}(v) \setminus K(t)), \qquad v \in SBV(\Omega)$$

from the minimality properties satisfied by $(v^{\delta,h}(t), K^{\delta,h}(t), \gamma^{\delta,h}(t))$. This is done in Proposition 5.5 by means of a refined version of the Transfer of Jump of [13]: the main difference here is that we have to deal with BV functions and we have to transfer the jump on the part of $K^{\delta,h}(t)$ where $\psi^{\delta,h}(t)$ is greater than a given small constant.

We also consider the cases $\alpha \in]0, \frac{1}{2}[$ and $\alpha > \frac{1}{2}$. It turns out that in the case $\alpha \in]0, \frac{1}{2}[$, the body is not solicited enough to create a crack, that is, Ω_h tends to behave elastically: more precisely we prove (Theorem 4.2) that setting

$$v^{\delta,h}(t,x) := \frac{1}{h^\alpha} u^{\delta,h}(t, hx) \tag{1.6}$$

for all $t \in [0,T]$, we have that $v^{\delta,h}(t)$ converges to the displacement of the elastic problem in Ω under boundary displacement given by $g(t)$.

In the case $\alpha > \frac{1}{2}$ we have that the body tends brutally toward *rupture*: in fact in Theorem 4.3 we prove that $v^{\delta,h}(0)$ given by (1.6) converges to a piecewise constant function v in Ω, so that $S^{g(0)}(v)$ disconnects Ω. This phenomenon is a consequence of the variational approach based on the search for global minimizers: as the size of Ω_h increases, cracks carry an energy of order h^{N-1}, while nonrigid displacements carry an energy of greater order: in this way crack is preferred to deformation.

The paper is organized as follows. In section 2 we recall some basic definitions and introduce the functional setting for the problem. In section 3 we deal with the problem of discrete-in-time evolutions for cracks in the cohesive case. The main theorems are listed in section 4, while in section 5 we prove some results which will be employed in their proofs to which sections 6, 7, and 8 are devoted. In the appendix we prove a relaxation result which is used in the problem of the discrete-in-time evolution of cracks.

2. Preliminaries. In this section we state the notation and recall the preliminary results employed in the rest of the paper.

2.1. Basic notation. We will employ the following basic notation:

- Ω is an open, connected, and bounded subset of $\mathbb{R}^N$ with Lipschitz boundary;
- $\partial_D\Omega$ is a subset of $\partial\Omega$ open in the relative topology;
- $\mathcal{H}^{N-1}$ is the $(N-1)$-dimensional Hausdorff measure;
- we say that $A \tilde{\subseteq} B$ if $A \subseteq B$ up to a set of $\mathcal{H}^{N-1}$-measure zero; similarly we say that $A \tilde{=} B$ if $A = B$ up to a set of $\mathcal{H}^{N-1}$-measure zero;

- $\Gamma \subseteq \Omega$ is rectifiable if there exists a sequence of C^1 manifolds $(M_i)_{i\in\mathbb{N}}$ such that $\Gamma \tilde{\subseteq} \cup_i M_i$;
- for all $A \subseteq \mathbb{R}^N$, $|A|$ denotes the Lebesgue measure of A;
- for all $A \subseteq \mathbb{R}^N$, 1_A denotes the characteristic function of A;
- if μ is a Borel measure on $\mathbb{R}^N$ and A is a Borel subset of $\mathbb{R}^N$, $\mu \llcorner A$ denotes the restriction of μ to A, i.e., $(\mu \llcorner A)(B) := \mu(B \cap A)$ for all Borel sets $B \subseteq \mathbb{R}^N$;
- $\|u\|_\infty$ and $\|u\|$ denote the sup-norm and the L^2 norm of u, respectively;
- if $u, g \in BV(\Omega;\mathbb{R}^m)$, $S^g(u) := S(u) \cup \{x \in \partial_D\Omega : u(x) \neq g(x)\}$;
- if $x, y \in \mathbb{R}$, $x \vee y := \max\{x, y\}$ and $x \wedge y := \min\{x, y\}$.

2.2. Functions of bounded variation. For the general theory of functions of bounded variation, we refer the reader to [4]; here we recall some basic definitions and theorems we need in what follows. We say that $u \in BV(A)$ if $u \in L^1(A)$, and its distributional derivative Du is a bounded vector-valued Radon measure on A. In this case it turns out that the set $S(u)$ of points $x \in A$ which are not Lebesgue points of u is rectifiable, that is, there exists a sequence of C^1 manifolds $(M_i)_{i\in\mathbb{N}}$ such that $S(u) \subseteq \cup_i M_i$ up to a set of $\mathcal{H}^{N-1}$-measure zero. As a consequence $S(u)$ admits a normal $\nu_u(x)$ at $\mathcal{H}^{N-1}$-a.e. $x \in S(u)$. Moreover, for $\mathcal{H}^{N-1}$-a.e. $x \in S(u)$, there exist $u^+(x), u^-(x) \in \mathbb{R}$ such that

$$\lim_{r\to 0} \frac{1}{|B_r^\pm(x)|} \int_{B_r^\pm(x)} |u(y) - u^\pm(x)|\, dy = 0,$$

where $B_r^+(x) := \{y \in B_r(x) : (y-x)\cdot\nu_u(x) > 0\}$ (similarly for $B_r^-(x)$), and $B_r(x)$ is the ball with center x and radius r. It turns out that Du can be represented as

$$Du(A) = \int_A \nabla u(x)\, dx + \int_{A\cap S(u)} (u^+(x) - u^-(x))\nu_u(x)\, d\mathcal{H}^{N-1}(x) + D^c u(A),$$

where ∇u denotes the approximate gradient of u and $D^c u$ is the Cantor part of Du. $BV(A)$ is a Banach space with respect to the norm $\|u\|_{BV(A)} := \|u\|_{L^1(A)} + |Du|(A)$.

We will often use the following result: if A is bounded and Lipschitz, and if $(u_k)_{k\in\mathbb{N}}$ is a bounded sequence in $BV(A)$, then there exist a subsequence $(u_{k_h})_{h\in\mathbb{N}}$ and $u \in BV(A)$ such that

$$\begin{aligned} u_{k_h} &\to u \quad \text{strongly in } L^1(A), \\ Du_{k_h} &\overset{*}{\rightharpoonup} Du \quad \text{weakly* in the sense of measures.} \end{aligned} \tag{2.1}$$

We say that $u_k \overset{*}{\rightharpoonup} u$ weakly* in $BV(A)$ if (2.1) holds.

We say that $u \in SBV(A)$ if $u \in BV(A)$ and $D^c u = 0$. The space $SBV(A)$ is called the space of *special functions of bounded variation*. Note that if $u \in SBV(A)$, then the singular part of Du is concentrated on $S(u)$.

The space SBV is very useful when dealing with variational problems involving volume and surface energies because of the following compactness and lower semicontinuity result due to Ambrosio [1], [2], [3].

THEOREM 2.1. *Let A be an open and bounded subset of $\mathbb{R}^N$, and let $(u_k)_{k\in\mathbb{N}}$ be a sequence in $SBV(A)$. Assume that there exist $q > 1$ and $C \in [0;+\infty[$ such that*

$$\int_A |\nabla u_k|^q\, dx + \mathcal{H}^{N-1}(S(u_k)) + ||u_k||_\infty \leq C$$

for every $k \in \mathbb{N}$. Then there exist a subsequence $(u_{k_h})_{h\in\mathbb{N}}$ and a function $u \in SBV(A)$ such that

$$\begin{aligned} u_{k_h} &\to u \quad \textit{strongly in } L^1(A),\\ \nabla u_{k_h} &\rightharpoonup \nabla u \quad \textit{weakly in } L^1(A;\mathbb{R}^N),\\ \mathcal{H}^{N-1}(S(u)) &\le \liminf_{h\to+\infty} \mathcal{H}^{N-1}(S(u_{k_h})). \end{aligned} \tag{2.2}$$

In the rest of the paper, we will say that $u_k \rightharpoonup u$ weakly in $SBV(A)$ if u_k and u satisfy (2.2). The following fact, which can be derived from Ambrosio's theorem, will also be useful: if $u_k \rightharpoonup u$ weakly in $SBV(A)$ and if $\mathcal{H}^{N-1}\llcorner S(u_k) \overset{*}{\rightharpoonup} \mu$ weakly* in the sense of measures, then $\mathcal{H}^{N-1}\llcorner S(u) \le \mu$ as measures.

Finally in the context of fracture problems we will use the following notation: if A is Lipschitz, and if $\partial_D A \subseteq \partial A$, then for all $u, g \in BV(A)$ we set

$$S^g(u) := S(u) \cup \{x \in \partial_D A \,:\, u(x) \neq g(x)\}, \tag{2.3}$$

where the inequality on $\partial_D A$ is intended in the sense of traces. Moreover, we set for all $x \in S(u)$

$$[u](x) := u^+(x) - u^-(x),$$

and for all $x \in \partial_D A$ we set $[u](x) := u(x) - g(x)$, where the traces of u and g on ∂A are used.

2.3. Quasi-static evolution in brittle fracture. Let Ω be an open bounded subset of $\mathbb{R}^N$ with Lipschitz boundary, and let $\partial_D\Omega$ be a subset of $\partial\Omega$ open in the relative topology. Let $g : [0,T] \to H^1(\Omega)$ be absolutely continuous (see [7] for a precise definition); we indicate the gradient of g at time t by $\nabla g(t)$, and the time derivative of g at time t by $\dot g(t)$. For $u \in SBV(\Omega)$, let $S^{g(t)}(u)$ be defined as in (2.3), and for every $A, B \subseteq \mathbb{R}^N$, let $A \tilde\subseteq B$ mean $A \subseteq B$ up to a set of $\mathcal{H}^{N-1}$-measure zero. The main result of [13] is the following theorem.

THEOREM 2.2. *There exists $\{(u(t), \Gamma(t)) : t \in [0,T]\}$ with $\Gamma(t) \tilde\subseteq \Omega \cup \partial_D\Omega$ rectifiable and $u(t) \in SBV(\Omega)$ such that*

(a) $\Gamma(s) \tilde\subseteq \Gamma(t)$ *for all* $0 \le s \le t \le T$;

(b) $u(0)$ *minimizes*

$$\|\nabla v\|^2 + \mathcal{H}^{N-1}(S^{g(0)}(v))$$

among all $v \in SBV(\Omega)$;

(c) *for $t \in]0,T]$, $u(t)$ minimizes*

$$\|\nabla v\|^2 + \mathcal{H}^{N-1}(S^{g(t)}(v) \setminus \Gamma(t))$$

among all $v \in SBV(\Omega)$;

(d) $S^{g(0)}(u(0)) = \Gamma(0)$, *and* $S^{g(t)}(u(t)) \tilde\subseteq \Gamma(t)$ *for all* $t \in]0,T]$.

Furthermore, the total energy

$$\mathcal{E}(t) := \|\nabla u(t)\|^2 + \mathcal{H}^{N-1}(\Gamma(t))$$

is absolutely continuous and satisfies

$$\mathcal{E}(t) = \mathcal{E}(0) + 2\int_0^t \int_\Omega \nabla u(\tau)\nabla \dot g(\tau)\, dx\, d\tau \tag{2.4}$$

for every $t \in [0,T]$.

Condition (a) stands for the *irreversibility* of the crack propagation, conditions (b) and (c) are *static equilibrium* conditions, while (2.4) stands for the *nondissipativity* of the process.

2.4. Γ-convergence. Let us recall the definition and some basic properties of De Giorgi's Γ-*convergence* in metric spaces. We refer the reader to [10] for an exhaustive treatment of this subject. Let (X,d) be a metric space. We say that a sequence $F_h : X \to [-\infty,+\infty]$ Γ-converges to $F : X \to [-\infty,+\infty]$ (as $h \to +\infty$) if for all $u \in X$ we have

(i) (Γ-*liminf inequality*) for every sequence $(u_h)_{h\in\mathbb{N}}$ converging to u in X,

$$\liminf_{h\to+\infty} F_h(u_h) \geq F(u);$$

(ii) (Γ-*limsup inequality*) there exists a sequence $(u_h)_{h\in\mathbb{N}}$ converging to u in X, such that

$$\limsup_{h\to+\infty} F_h(u_h) \leq F(u).$$

The function F is called the Γ-limit of $(F_h)_{h\in\mathbb{N}}$ (with respect to d). Γ-convergence is a convergence of variational type as explained in the following proposition.

PROPOSITION 2.3. *Assume that the sequence* $(F_h)_{h\in\mathbb{N}}$ Γ-*converges to* F *and that there exists a compact set* $K \subseteq X$ *such that for all* $h \in \mathbb{N}$

$$\inf_{u\in K} F_h(u) = \inf_{u\in X} F_h(u).$$

Then F *admits a minimum on* X, $\inf_X F_h \to \min_X F$, *and any limit point of any sequence* $(u_h)_{h\in\mathbb{N}}$ *such that*

$$\lim_{h\to+\infty} \Big(F_h(u_h) - \inf_{u\in X} F_h(u)\Big) = 0$$

is a minimizer of F.

Moreover, the following compactness result holds.

PROPOSITION 2.4. *If* (X,d) *is separable, and* $(F_h)_{h\in\mathbb{N}}$ *is a sequence of functionals on* X, *then there exist a subsequence* $(F_{h_k})_{k\in\mathbb{N}}$ *and a function* $F : X \to [-\infty;+\infty]$ *such that* $(F_{h_k})_{k\in\mathbb{N}}$ Γ-*converges to* F.

3. Discrete-in-time evolution of cracks in the cohesive case. In this section we are interested in generalized antiplanar shear of an elastic body Ω in the context of linearized elasticity and in presence of cohesive forces on the cracks.

The notion of *discrete-in-time evolution* for cracks relative to time-dependent boundary displacement $g(t)$ has been described in the introduction. It relies on the minimization of functionals of the form

$$\|\nabla u\|^2 + \int_{\Gamma\cup S^{g(t)}(u)} \varphi(|[u]| \vee \psi)\, d\mathcal{H}^{N-1}, \tag{3.1}$$

with ψ a positive function defined on Γ. We now define rigorously the functional space to which the displacements belong, and the properties of Ω, Γ, ψ, and $g(t)$ in order to prove an existence result for the discrete-in-time evolution of cracks.

Let Ω be an open bounded subset of $\mathbb{R}^N$ with Lipschitz boundary. Let $\partial_D\Omega \subseteq \partial\Omega$ be open in the relative topology, and let $\partial_N\Omega := \partial\Omega\setminus\partial_D\Omega$. Let $\varphi : [0,+\infty[\to [0,+\infty[$

be increasing and concave, $\varphi(0) = 0$ and such that $\lim_{s\to+\infty} \varphi(s) = 1$. If $a := \varphi'(0) < +\infty$, we have

$$\varphi(s) \le as \qquad \text{for all } s \in [0, +\infty[.$$

Let $T > 0$, and let us consider a boundary displacement $g \in AC([0,T]; H^1(\Omega))$ such that $\|g(t)\|_\infty \le C$ for all $t \in [0,T]$. We discretize g in the following way. Given $\delta > 0$, let I_δ be a subdivision of $[0,T]$ of the form $0 = t_0^\delta < t_1^\delta < \cdots < t_{N_\delta}^\delta = T$ such that $\max_i(t_i^\delta - t_{i-1}^\delta) < \delta$. For $0 \le i \le N_\delta$ we set $g_i^\delta := g(t_i^\delta)$.

As for the space of the displacements, it would be natural, following [13], to consider $u \in SBV(\Omega)$. Since $a = \varphi'(0) < +\infty$, we have unfortunately that the minimization of (3.1) is not well posed in $SBV(\Omega)$. Let us in fact consider $(u_n)_{n\in\mathbb{N}}$ minimizing sequence for (3.1): it turns out that we may assume $(u_n)_{n\in\mathbb{N}}$ bounded in $BV(\Omega)$. As a consequence $(u_n)_{n\in\mathbb{N}}$ admits a subsequence weakly* convergent in $BV(\Omega)$ to a function $u \in BV(\Omega)$. Then we have that minimizing sequences of (3.1) converge (up to a subsequence) to a minimizer of the relaxation of (3.1) with respect to the weak* topology of $BV(\Omega)$. By Proposition 9.1, the natural domain of this relaxed functional is $BV(\Omega)$, and that its form is

$$\int_\Omega f(\nabla u)\,dx + \int_{\Gamma\cup S^{g(t)}(u)} \varphi(|[u]| \vee \psi)\,d\mathcal{H}^{N-1} + a|D^c u|(\Omega), \tag{3.2}$$

where

$$f(\xi) := \begin{cases} |\xi|^2 & \text{if } |\xi| \le \frac{a}{2}, \\ \\ \frac{a^2}{4} + a(|\xi| - \frac{a}{2}) & \text{if } |\xi| \ge \frac{a}{2}. \end{cases} \tag{3.3}$$

In view of these remarks, we consider $BV(\Omega)$ as the space of displacements u of the body Ω, and a total energy of the form (3.2). The volume part in the energy (3.2) can be interpreted as the contribution of the elastic behavior of the body. The second term represents the work done to create the crack $\Gamma \cup S^{g(t)}(u)$ with opening given by $|[u]| \vee \psi$. The new term $a|D^c u|$ can be interpreted as the contribute of microcracks in the configuration which are considered as reversible.

Let us define the discrete evolution of the crack in this new setting. For $i = 0$, let $u_0^\delta \in BV(\Omega)$ be a minimum of

$$\min_{u\in BV(\Omega)} \left\{ \int_\Omega f(\nabla u)\,dx + \int_{S^{g_0^\delta}(u)} \varphi(|[u]|)\,d\mathcal{H}^{N-1} + a|D^c u| \right\}. \tag{3.4}$$

We set $\Gamma_0^\delta := S^{g_0^\delta}(u_0^\delta)$.

Supposing to have constructed u_j^δ and Γ_j^δ for all $j = 0, \dots, i-1$, let u_i^δ be a minimum of

$$\min_{u\in BV(\Omega)} \left\{ \int_\Omega f(\nabla u)\,dx + \int_{S^{g_i^\delta}(u)\cup\Gamma_{i-1}^\delta} \varphi(|[u]| \vee \psi_{i-1}^\delta)\,d\mathcal{H}^{N-1} + a|D^c u| \right\}, \tag{3.5}$$

where $\psi_{i-1}^\delta := |[u_0^\delta]| \vee \cdots \vee |[u_{i-1}^\delta]|$. We set $\Gamma_i^\delta := \Gamma_{i-1}^\delta \cup S^{g_i^\delta}(u_i^\delta)$.

In the following proposition we establish the existence of this discrete evolution.

PROPOSITION 3.1. *Let $I_\delta = \{0 = t_0^\delta < \cdots < t_{N_\delta}^\delta = T\}$ be a subdivision of $[0,T]$ such that $\max(t_i^\delta - t_{i-1}^\delta) < \delta$. Then for all $i = 0, \dots, N_\delta$ there exists $u_i^\delta \in BV(\Omega)$ such that setting $\Gamma_{-1}^\delta := \emptyset$, $\psi_{-1}^\delta := 0$,*

$$\Gamma_i^\delta := \bigcup_{j=0}^{i} S^{g_j^\delta}(u_j^\delta), \qquad \psi_i^\delta(x) := |[u_0^\delta]|(x) \vee \cdots \vee |[u_i^\delta]|(x),$$

the following holds:

(a) $\|u_i^\delta\|_\infty \le \|g_i^\delta\|_\infty \le C$;

(b) *for all $v \in BV(\Omega)$ we have*

$$\begin{aligned} \int_\Omega f(\nabla u_i^\delta)\,dx + \int_{\Gamma_i^\delta} \varphi(\psi_i^\delta)\,d\mathcal{H}^{N-1} + a|D^c u_i^\delta|(\Omega) \\ \le \int_\Omega f(\nabla v)\,dx + \int_{S^{g_i^\delta}(v)\cup\Gamma_{i-1}^\delta} \varphi(|[v]| \vee \psi_{i-1}^\delta)\,d\mathcal{H}^{N-1} + a|D^c v|(\Omega), \end{aligned} \tag{3.6}$$

where $a = \varphi'(0)$ and f is defined in (3.3);

(c) *we have that*

$$\begin{aligned} &\int_\Omega f(\nabla u_i^\delta)\,dx + \int_{\Gamma_i^\delta} \varphi(\psi_i^\delta)\,d\mathcal{H}^{N-1} + a|D^c u_i^\delta|(\Omega) \\ &\quad = \inf_{v \in SBV(\Omega)} \left\{ \|\nabla v\|^2 + \int_{S^{g_i^\delta}(v)\cup\Gamma_{i-1}^\delta} \varphi(|[v]| \vee \psi_{i-1}^\delta)\,d\mathcal{H}^{N-1} \right\}. \end{aligned}$$

Proof. We have to prove that problems (3.4) and (3.5) admit solutions. Let us consider, for example, problem (3.5), the other being similar. Let $(u_n)_{n\in\mathbb{N}}$ be a minimizing sequence for problem (3.5). By a truncation argument we may assume that $\|u_n\|_\infty \le \|g_i^\delta\|$. Comparing u_n with g_i^δ, we get for n large

$$\begin{aligned} \int_\Omega f(\nabla u_n)\,dx + \int_{S^{g_i^\delta}(u_n)\cup\Gamma_{i-1}^\delta} \varphi(|[u_n]| \vee \psi_{i-1}^\delta)\,d\mathcal{H}^{N-1} + a|D^c u_n|(\Omega) \\ \le \int_\Omega f(\nabla g_0^\delta)\,dx + \int_{\Gamma_{i-1}^\delta} \varphi(\psi_{i-1}^\delta)\,d\mathcal{H}^{N-1} + 1 \le C', \end{aligned}$$

with C' independent of n. Since there exists $d > 0$ such that $a|\xi| - d \le f(\xi)$ for all $\xi \in \mathbb{R}^N$, we deduce that $(\nabla u_n)_{n\in\mathbb{N}}$ is bounded in $L^1(\Omega;\mathbb{R}^N)$. Moreover, if $\bar s$ is such that $\varphi(\bar s) = \frac{1}{2}$ and $\bar a$ is such that $s \le \bar a \varphi(s)$ for all $s \in [0,\bar s]$, we have

$$\begin{aligned} \int_{S(u_n)} |[u_n]|\,d\mathcal{H}^{N-1} &= \int_{\{|[u_n]|\le\bar s\}} |[u_n]|\,d\mathcal{H}^{N-1} + \|g_i^\delta\|_\infty \mathcal{H}^{N-1}(\{|[u_n]| > \bar s\}) \\ &\le \bar a \int_{|[u_n]|\le\bar s} \varphi(|[u_n]|)\,d\mathcal{H}^{N-1} + 2\|g_i^\delta\|_\infty \int_{|[u_n]|>\bar s} \varphi(|[u_n]|)\,d\mathcal{H}^{N-1} \\ &\le (\bar a + 2\|g_i^\delta\|_\infty) C'. \end{aligned}$$

Finally for all n

$$|D^c u_n| \le \frac{C'}{a}.$$

We conclude that $(u_n)_{n\in\mathbb{N}}$ is bounded in $BV(\Omega)$. Then there exists $u \in BV(\Omega)$ such that up to a subsequence $u_n \overset{*}{\rightharpoonup} u$ weakly* in $BV(\Omega)$ and pointwise almost everywhere. Let us set $u_i^\delta := u$. By Lemma 9.2 we deduce that

$$\int_\Omega f(\nabla u)\,dx + \int_{S^{g_i^\delta}(u)\cup\Gamma_{i-1}^\delta} \varphi(|[u]|\vee\psi_{i-1}^\delta)\,d\mathcal{H}^{N-1} + a|D^c u|(\Omega)$$
$$\le \liminf_{n\to+\infty}\int_\Omega f(\nabla u_n)\,dx + \int_{S^{g_i^\delta}(u_n)\cup\Gamma_{i-1}^\delta} \varphi(|[u_n]|\vee\psi_{i-1}^\delta)\,d\mathcal{H}^{N-1} + a|D^c u_n|(\Omega).$$

Setting $\psi_i^\delta := \psi_{i-1}^\delta \vee |[u_i^\delta]|$, we have that point (b) holds. Moreover $\|u_i^\delta\|_\infty \le \|g_0^\delta\|_\infty \le C$, so that point (a) holds. Finally point (c) is a consequence of Proposition 9.1. ∎

Let us consider now the following piecewise constant interpolation in time:

$$u^\delta(t) := u_i^\delta, \qquad \Gamma^\delta(t) := \Gamma_i^\delta, \qquad \psi^\delta(t) := \psi_i^\delta, \qquad g^\delta(t) := g_i^\delta \tag{3.7}$$

for $t_i^\delta \le t < t_{i+1}^\delta$, with $u^\delta(T) := u_{N_\delta}^\delta$, $\Gamma^\delta(T) := \Gamma_{N_\delta}^\delta$, $\psi^\delta(T) := \psi_{N_\delta}^\delta$, and $g^\delta(T) := g(T)$.

For every $v \in BV(\Omega)$ and for every $t \in [0,T]$ let us set

$$\mathcal{E}^\delta(t,v) := \int_\Omega f(\nabla v)\,dx + \int_{S^{g^\delta(t)}(v)\cup\Gamma^\delta(t)} \varphi(|[v]|\vee\psi^\delta(t))\,d\mathcal{H}^{N-1} + a|D^c v|(\Omega). \tag{3.8}$$

Then the following estimate holds.

LEMMA 3.2. *There exists $e_a^\delta \to 0$ for $\delta \to 0$ and $a \to +\infty$ such that for all $t \in [0,T]$ we have*

$$\mathcal{E}^\delta(t,u^\delta(t)) \le \mathcal{E}^\delta(0,u^\delta(0)) + \int_0^{t_i^\delta}\int_\Omega f'(\nabla u^\delta(\tau))\nabla\dot g(\tau)\,dx\,d\tau + e_a^\delta,$$

where t_i^δ is the step discretization point such that $t_i^\delta \le t < t_{i+1}^\delta$.

Proof. Comparing u_i^δ with $u_{i-1}^\delta + g_i^\delta - g_{i-1}^\delta$ by means of (3.6) we obtain

$$\mathcal{E}^\delta(t_i^\delta,u_i^\delta) \le \int_\Omega f(\nabla u_{i-1}^\delta + \nabla g_i^\delta - \nabla g_{i-1}^\delta)\,dx + \int_{\Gamma_{i-1}^\delta}\varphi(\psi_{i-1}^\delta)\,d\mathcal{H}^{N-1} + a|D^c u_{i-1}^\delta|(\Omega).$$

Notice that by the very definition of f the following hold:

(1) if $|\nabla u_{i-1}^\delta + \nabla g_i^\delta - \nabla g_{i-1}^\delta| \ge \frac{a}{2}$ and $|\nabla u_{i-1}^\delta| \ge \frac{a}{2}$,

$$f'(\nabla u_{i-1}^\delta + \nabla g_i^\delta - \nabla g_{i-1}^\delta) = f'(\nabla u_{i-1}^\delta);$$

(2) if $|\nabla u_{i-1}^\delta + \nabla g_i^\delta - \nabla g_{i-1}^\delta| < \frac{a}{2}$ and $|\nabla u_{i-1}^\delta| \ge \frac{a}{2}$,

$$f(\nabla u_{i-1}^\delta + \nabla g_i^\delta - \nabla g_{i-1}^\delta) \le f(\nabla u_{i-1}^\delta);$$

(3) if $|\nabla u_{i-1}^\delta + \nabla g_i^\delta - \nabla g_{i-1}^\delta| \ge \frac{a}{2}$ and $|\nabla u_{i-1}^\delta| < \frac{a}{2}$,

$$f(\nabla u_{i-1}^\delta+\nabla g_i^\delta-\nabla g_{i-1}^\delta) \le f(\nabla u_{i-1}^\delta)+2(\nabla u_{i-1}^\delta, \nabla g_i^\delta-\nabla g_{i-1}^\delta)+|\nabla g_i^\delta-\nabla g_{i-1}^\delta|^2;$$

(4) if $|\nabla u_{i-1}^\delta + \nabla g_i^\delta - \nabla g_{i-1}^\delta| < \frac{a}{2}$ and $|\nabla u_{i-1}^\delta| < \frac{a}{2}$,

$$f(\nabla u_{i-1}^\delta+\nabla g_i^\delta-\nabla g_{i-1}^\delta) = f(\nabla u_{i-1}^\delta)+2(\nabla u_{i-1}^\delta, \nabla g_i^\delta-\nabla g_{i-1}^\delta)+|\nabla g_i^\delta-\nabla g_{i-1}^\delta|^2.$$

Then by convexity of f we deduce

$$\mathcal{E}^\delta(t_i^\delta, u_i^\delta) \le \mathcal{E}^\delta(t_{i-1}^\delta, u_{i-1}^\delta) + \int_\Omega f'(\nabla u_{i-1}^\delta)(\nabla g_i^\delta - \nabla g_{i-1}^\delta)\,dx + R_{i-1}^{\delta,a},$$

where

$$R_{i-1}^{\delta,a} := \int_\Omega |\nabla g_i^\delta - \nabla g_{i-1}^\delta|^2\,dx + \int_{\{|\nabla u_{i-1}^\delta| \ge \frac{a}{2}\}} |f'(\nabla u_{i-1}^\delta)|\,|\nabla g_i^\delta - \nabla g_{i-1}^\delta|\,dx.$$

Then summing up from t_i^δ to t_0^δ, and taking into account (3.7) we get

$$\mathcal{E}^\delta(t, u^\delta(t)) \le \mathcal{E}^\delta(0, u^\delta(0)) + \int_0^{t_i^\delta}\int_\Omega f'(\nabla u^\delta(\tau))\nabla \dot g(\tau)\,dx\,d\tau + \int_0^{t_i^\delta} R^{\delta,a}(\tau)\,d\tau,$$

where

$$R^{\delta,a}(\tau) := \sigma(\delta)\|\nabla \dot g(\tau)\| + \int_{\{|\nabla u^\delta(\tau)| \ge \frac{a}{2}\}} |f'(\nabla u^\delta(\tau))|\,|\nabla \dot g(\tau)|\,dx$$

and

$$\sigma(\delta) := \max_{i=1,\dots,N_\delta} \int_{t_{i-1}^\delta}^{t_i^\delta} \|\nabla \dot g\|\,d\tau.$$

In order to conclude the proof it is sufficient to see that

$$\int_0^T R^{\delta,a}(\tau)\,d\tau \to 0$$

as $\delta \to 0$ and $a \to +\infty$. Notice that $\sigma(\delta) \to 0$ as $\delta \to 0$ by the absolutely continuity of $\|\nabla \dot g\|$. Let us come to the second term. Note that $|f'(\nabla u^\delta(\tau))| = a$ on $\{|\nabla u^\delta(\tau)| \ge \frac{a}{2}\}$. Then we have to see

$$\int_0^T \int_\Omega a|\nabla \dot g(\tau)| 1_{\{|\nabla u^\delta(\tau)| \ge \frac{a}{2}\}}\,dx\,d\tau \to 0 \tag{3.9}$$

as $\delta \to 0$ and $a \to +\infty$. Setting $A_a^\delta(\tau) := \{x \in \Omega : |\nabla u^\delta(\tau)|(x) \ge \frac{a}{2}\}$ we have by Hölder inequality

$$\int_\Omega a|\nabla \dot g(\tau)| 1_{A_a^\delta(\tau)}\,dx \le a\sqrt{|A_a^\delta(\tau)|}\left(\int_{A_a^\delta(\tau)} |\nabla \dot g(\tau)|^2\,dx\right)^{\frac{1}{2}}.$$

Notice that

$$\frac{a^2}{2}|A_a^\delta(\tau)| \le a\int_{A_a^\delta(\tau)} |\nabla u^\delta(\tau)|\,dx \le 2\int_{A_a^\delta(\tau)} f(\nabla u^\delta(\tau))\,dx \le C', \tag{3.10}$$

where C' depends only on g and is obtained comparing $u^\delta(\tau)$ with $g^\delta(\tau)$ by means of (3.6). We deduce that

$$\int_\Omega a|\nabla \dot g(\tau)| 1_{A_a^\delta(\tau)}\,dx \le \sqrt{2C'}\left(\int_{A_a^\delta(\tau)} |\nabla \dot g(\tau)|^2\,dx\right)^{\frac{1}{2}} \le \sqrt{2C'}\|\nabla \dot g(\tau)\|.$$

As $\delta \to 0$ and $a \to +\infty$, by (3.10) we have that $|A_a^\delta(\tau)| \to 0$. Then by the equicontinuity of $|\nabla \dot g(\tau)|^2$ and by the Dominated Convergence theorem, we deduce that (3.9) holds, and the proof is finished. □

4. The main results. Let Ω be an open, connected, and bounded subset of $\mathbb{R}^N$ with Lipschitz boundary. Let $\partial_D\Omega \subseteq \partial\Omega$ be open in the relative topology, and let $\partial_N\Omega := \partial\Omega \setminus \partial_D\Omega$.

In this section we consider discrete-in-time evolution of cracks in a linearly elastic body whose reference configuration is given by $\Omega_h := h\Omega$, where h is positive and large. Let us assume that the cohesive forces on the cracks of Ω_h are given in the sense of section 3 by a function $\varphi : [0,+\infty[\to [0,1]$ which is increasing, concave, $\varphi(0) = 0$, $\varphi'(0) = a < +\infty$, and such that $\lim_{s\to+\infty} \varphi(s) = 1$. Let us moreover set

$$f(\xi) := \begin{cases} |\xi|^2 & \text{if } |\xi| \le \frac{a}{2}, \\ \\ \frac{a^2}{4} + a(|\xi| - \frac{a}{2}) & \text{if } |\xi| \ge \frac{a}{2}. \end{cases} \tag{4.1}$$

Let us consider on $\partial_D\Omega_h := h\partial_D\Omega$ boundary displacements of the following particular form:

$$g_h(t,x) := h^\alpha g\left(t, \frac{x}{h}\right), \tag{4.2}$$

with $g \in AC([0,T]; H^1(\Omega))$ such that $\|g(t)\|_\infty \le C$ for all $t \in [0,T]$. Given $\delta > 0$, let

$$I_\delta = \{0 = t_0^\delta < \cdots < t_{N_\delta}^\delta = T\}$$

be a subdivision of $[0,T]$ such that $\max(t_i^\delta - t_{i-1}^\delta) < \delta$, and let

$$\{t \to (u^{\delta,h}(t), \Gamma^{\delta,h}(t), \psi^{\delta,h}(t)) \,:\, t \in [0,T]\}$$

be the piecewise constant interpolation in the sense of (3.7) of a discrete-in-time evolution of cracks in Ω_h relative to the boundary datum g_h and the subdivision I_δ given by Proposition 3.1.

Our aim is to study the asymptotic behavior of $\{t \to (u^{\delta,h}(t), \Gamma^{\delta,h}(t), \psi^{\delta,h}(t)) : t \in [0,T]\}$ as $\delta \to 0$ and $h \to +\infty$. Let us consider $h \in \mathbb{N}$ (we can consider any sequence which diverges to $+\infty$), let us fix $\delta_h \to 0$, and let us set for all $t \in [0,T]$

$$u_h(t) := u^{\delta_h,h}(t), \qquad \Gamma_h(t) := \Gamma^{\delta_h,h}(t), \qquad \psi_h(t) := \psi^{\delta_h,h}(t), \tag{4.3}$$

and let $g_h^{\delta_h}(t) := g_h(t_i^{\delta_h})$ where $t_i^{\delta_h} \in I_{\delta_h}$ is such that $t_i^{\delta_h} \le t < t_{i+1}^{\delta_h}$. Let us moreover set for every $v \in BV(\Omega)$ and for every $t \in [0,T]$

$$\begin{aligned} \mathcal{E}_h(t,v) := &\int_{\Omega_h} f(\nabla v)\,dx + \int_{S^{g_h^{\delta_h}(t)}(v)\cup\Gamma_h(t)} \varphi(|[v]| \vee \psi_h(t))\,d\mathcal{H}^{N-1} \\ &+ a|D^c v|(\Omega_h). \end{aligned} \tag{4.4}$$

The asymptotic of (u_h, Γ_h, ψ_h) depends on α, and we have to distinguish three cases. The first case $\alpha = \frac{1}{2}$ was stated in the introduction and reveals the prevalence of brittle effects as the size of the body increases. We give here the precise statement we will prove.

THEOREM 4.1. *Let $g \in AC(0,T;H^1(\Omega))$ be such that $\|g(t)\|_\infty \le C$ for all $t \in [0,T]$. Let $\{t \to (u_h(t), \Gamma_h(t), \psi_h(t)) : t \in [0,T]\}$ be the piecewise constant interpolation given in* (4.3) *of a discrete-in-time evolution of cracks in Ω_h relative to the boundary data*

$$g_h(x,t) := \sqrt{h}\, g\left(\frac{x}{h}, t\right).$$

Then the following facts hold:

(a) *there exists a constant C' dependent only on g such that for all $t \in [0,T]$*

$$\frac{1}{h^{N-1}}\mathcal{E}_h(t,u_h(t)) \le C';$$

(b) *for all $t \in [0,T]$*

$$v_h(t,x) := \frac{1}{\sqrt{h}}u_h(t,hx) \qquad \text{is bounded in } BV(\Omega);$$

(c) *there exists a subsequence independent of t and there exists a quasi-static crack evolution $\{t \to (v(t),K(t)) : t \in [0,T]\}$ in Ω relative to boundary displacement g in the sense of Theorem 2.2 such that for all $t \in [0,T]$ we have*

$$\nabla v_h(t) \rightharpoonup \nabla v(t) \qquad \text{weakly in } L^1(\Omega;\mathbb{R}^N),$$

and every accumulation point v of $(v_h(t))_{h\in\mathbb{N}}$ in the weak topology of $BV(\Omega)$ is such that $v \in SBV(\Omega)$, $S^{g(t)}(v) \tilde{\subseteq} K(t)$, and $\nabla v = \nabla v(t)$. Moreover, for all $t \in [0,T]$ we have*

$$\frac{1}{h^{N-1}}\mathcal{E}_h(t,u_h(t)) \to \|\nabla v(t)\|^2 + \mathcal{H}^{N-1}(K(t)); \tag{4.5}$$

in particular $h^{-N+1}|D^c u_h(t)|(\Omega_h)| \to 0$,

$$\frac{1}{h^{N-1}}\int_{\Omega_h} f(\nabla u_h(t))\,dx \to \|\nabla v(t)\|^2, \tag{4.6}$$

and

$$\frac{1}{h^{N-1}}\int_{\Gamma_h(t)} \varphi(\psi_h(t))\,d\mathcal{H}^{N-1} \to \mathcal{H}^{N-1}(K(t)). \tag{4.7}$$

Notice that in point (c) we cannot assert that the sequence $(v_h(t))_{h\in\mathbb{N}}$ converges to $v(t)$ in the weak* topology of $BV(\Omega)$: this is due to the fact that $K(t)$ could disconnect Ω (in a weak sense), so that $v(t)$ is determined up to a constant on the components of $\Omega \setminus K$ which do not touch $\partial_D\Omega$.

The case $\alpha < \frac{1}{2}$ leads to a problem in elasticity in Ω_h in the sense of the following theorem.

THEOREM 4.2. *Let $g \in AC(0,T;H^1(\Omega))$ be such that $\|g(t)\|_\infty \le C$ for all $t \in [0,T]$. Let $\{t \to (u_h(t),\Gamma_h(t),\psi_h(t)) : t \in [0,T]\}$ be the piecewise constant interpolation given in* (4.3) *of a discrete-in-time evolution of cracks in Ω_h relative to the boundary data*

$$g_h(x,t) := h^\alpha g\left(t,\frac{x}{h}\right)$$

with $\alpha < \frac{1}{2}$. Then the following facts hold:

(a) *for all $t \in [0,T]$*

$$v_h(t,x) := \frac{1}{h^\alpha}u_h(t,hx) \qquad \text{is bounded in } BV(\Omega);$$

(b) *there exists a subsequence independent of* t *such that for all* $t \in [0,T]$ *we have* $v_h(t) \overset{*}{\rightharpoonup} v(t)$ *weakly* in* $BV(\Omega)$ *and*

$$\nabla v_h(t) \rightharpoonup \nabla v(t) \qquad \text{weakly in } L^1(\Omega;\mathbb{R}^N),$$

where $v(t)$ *is the minimizer of*

$$\min\{\|\nabla v\|^2 \,:\, v \in H^1(\Omega), v = g(t) \text{ on } \partial_D\Omega\};$$

moreover for all $t \in [0,T]$ *we have*

$$\frac{1}{h^{N+2\alpha-2}} \int_{\Omega_h} f(\nabla u_h(t))\,dx \to \|\nabla v(t)\|^2.$$

Finally for the case $\alpha > \frac{1}{2}$ the body goes to *rupture* at time $t = 0$, in the sense of the following theorem.

THEOREM 4.3. *Let* $g \in AC(0,T;H^1(\Omega))$ *be such that* $\|g(t)\|_\infty \le C$ *for all* $t \in [0,T]$. *Let* $\{t \to (u_h(t), \Gamma_h(t), \psi_h(t)) \,:\, t \in [0,T]\}$ *be the piecewise constant interpolation given in* (4.3) *of a discrete-in-time evolution of cracks in* Ω_h *relative to the boundary data*

$$g_h(x,t) := h^\alpha g\left(\frac{x}{h}, t\right)$$

with $\alpha > \frac{1}{2}$. *Let us set* $v_h(t,x) := \frac{1}{h^\alpha} u_h(t,hx)$ *for all* $x \in \Omega$ *and for all* $t \in [0,T]$.

Then $(v_h(0))_{h\in\mathbb{N}}$ *is bounded in* $BV(\Omega)$, *and every accumulation point* v *of* $(v_h(0))_{h\in\mathbb{N}}$ *in the weak* topology of* $BV(\Omega)$ *is piecewise constant in* Ω, *i.e.,* $v \in SBV(\Omega)$ *and* $\nabla v = 0$. *Moreover,*

$$\mathcal{H}^{N-1}(S^{g(0)}(v(0))) \le \mathcal{H}^{N-1}(S^{g(0)}(w)) \tag{4.8}$$

for all piecewise constant function $w \in SBV(\Omega)$.

Notice that the minimality property (4.8) can be restated saying that $v(0)$ determines a minimal partition of Ω (in the sense of the perimeter of $S^{g(0)}(w)$).

5. Some tools for the asymptotic analysis. In this section we prove some technical propositions which will be very useful in the proofs of the main results of the paper. More precisely, we will prove compactness results for the displacements and the cracks, and we will prove a generalization of the Transfer of Jump of [13] which will be employed in order to study what the minimality property of the discrete-in-time evolutions imply in the limit.

For all $h \in \mathbb{N}$ let $f_h : \mathbb{R}^N \to [0,+\infty[$ be such that for all $\xi \in \mathbb{R}^N$

$$f_h(\xi) \nearrow |\xi|^2, \qquad f_h(\xi) \ge \min\{|\xi|^2 - 1, b_h|\xi|\} \tag{5.1}$$

with $b_h \to +\infty$ as $h \to +\infty$, and let $\varphi_h : [0,+\infty[\to [0,1]$ be increasing and such that for all $s \in [0,+\infty[$

$$\varphi_h(s) \ge \min\{c_h s, d_h\} \tag{5.2}$$

with $c_h \to +\infty$ and $d_h \nearrow 1$ for $h \to +\infty$.

5.1. Compactness for the displacements. In this subsection we give a compactness and lower semicontinuity result for the displacements whose proof is inspired by the proof of Ambrosio's compactness theorem (see [3]).

PROPOSITION 5.1. *Let us consider the functionals*

$$F_h(u) := \sum_{i=1}^{m} \int_\Omega f_h(\nabla u_i)\,dx + \int_{S(u)} \varphi_h(|[u_1]| \vee \cdots \vee |[u_m]|)\,d\mathcal{H}^{N-1} + a_h|D^c u|(\Omega),$$

where $u = (u_1, \dots, u_m) \in BV(\Omega;\mathbb{R}^m)$ *(with* f_h *and* φ_h *defined in* (5.1) *and* (5.2)*). Let* $a_h \to +\infty$ *for* $h \to +\infty$*, and let* $(u_h)_{h\in\mathbb{N}}$ *be a sequence in* $BV(\Omega)$*. Then the following facts hold.*

(a) *If*

$$F_h(u^h) + \|u^h\|_{L^\infty(\Omega;\mathbb{R}^m)} \le C$$

for some $C \in [0, +\infty[$*, then up to a subsequence*

$$u^h \overset{*}{\rightharpoonup} u \qquad \text{weakly* in } BV(\Omega;\mathbb{R}^m).$$

(b) *If* $F_h(u^h) \le C$ *for some* $C \in [0, +\infty[$ *and* $u^h \overset{*}{\rightharpoonup} u$ *weakly* in* $BV(\Omega;\mathbb{R}^m)$*, then* $u \in SBV(\Omega;\mathbb{R}^m)$*,*

$$\nabla u^h \rightharpoonup \nabla u \qquad \text{weakly in } L^1(\Omega;\mathbb{R}^{m\times N}), \tag{5.3}$$

$$\|\nabla u_i\|^2 \le \liminf_{h\to+\infty} \int_\Omega f_h(\nabla u_i^h)\,dx, \qquad i = 1, \dots, m, \tag{5.4}$$

and

$$\mathcal{H}^{N-1}(S(u)) \le \liminf_{h\to+\infty} \int_{S(u^h)} \varphi_h(|[u_1^h]| \vee \cdots \vee |[u_m^h]|)\,d\mathcal{H}^{N-1}. \tag{5.5}$$

Proof. As for point (a), let us prove that there exists C' independent of h such that we have

$$|Du^h(t)|(\Omega) \le C'. \tag{5.6}$$

In fact, since for h large we have for all $\xi \in \mathbb{R}^N$

$$|\xi| - 1 \le f_h(\xi),$$

we deduce that for all $i = 1, \dots, m$

$$\int_\Omega |\nabla u_i^h|\,dx \le \int_\Omega [f_h(\nabla u_i^h) + 1]\,dx \le C + |\Omega|,$$

where $|\Omega|$ denotes the Lebesgue measure of Ω. Moreover, if h is so large that $s \le 2\varphi_h(s)$ for all $s \in [0, 1]$, we have for all $i = 1, \dots, m$

$$\begin{aligned}\int_{S(u_i^h)} |[u_i^h]|\,d\mathcal{H}^{N-1} &\le \int_{|[u_i^h]|<1} |[u_i^h]|\,d\mathcal{H}^{N-1} + \|u_i^h(t)\|_\infty \mathcal{H}^{N-1}\left(\left\{|[u_i^h]| \ge 1\right\}\right)\\ &\le 2\int_{|[u_i^h]|<1} \varphi_h(|[u_i^h]|)\,d\mathcal{H}^{N-1} + 2C\int_{|[u_i^h]|\ge1} \varphi_h(|[u_i^h]|)\,d\mathcal{H}^{N-1} \le 2(1+C)C.\end{aligned}$$

Finally for all h

$$|D^c u^h|(\Omega) \le \frac{C}{a_h}.$$

We deduce that (5.6) holds, and so up to a subsequence we may suppose that $u_h \overset{*}{\rightharpoonup} u$ weakly* in $BV(\Omega;\mathbb{R}^m)$.

Let us come to point (b). Let us consider $u^h \in BV(\Omega)$ such that $u^h \overset{*}{\rightharpoonup} u$ weakly* in $BV(\Omega;\mathbb{R}^m)$ and $F_h(u^h) \le C$. Notice that $(\nabla u^h)_{h\in\mathbb{N}}$ is equi-integrable. In fact if r_h is such that for all $|\xi| \le r_h$

$$|\xi|^2 - 1 \le b_h|\xi|,$$

we get for all $i = 1,\dots,m$ and for all $E \subseteq \Omega$

$$\begin{aligned}\int_E |\nabla u_i^h|\,dx &\le \int_{\{|\nabla u_i^h|\le r_h\}\cap E} |\nabla u_i^h|\,dx + \int_{\{|\nabla u_i^h|> r_h\}\cap E} |\nabla u_i^h|\,dx\\ &\le \left(\int_{\{|\nabla u_i^h|\le r_h\}\cap E} |\nabla u_i^h|^2\,dx\right)^{\frac12} |E|^{\frac12} + \int_{\{|\nabla u_i^h|> r_h\}\cap E} |\nabla u_i^h|\,dx\\ &\le \left(\int_\Omega (f_h(\nabla u_i^h)+1)\,dx\right)^{\frac12} |E|^{\frac12} + \frac{1}{b_h}\int_\Omega f_h(\nabla u_i^h)\,dx \le \sqrt{(C+|\Omega|)|E|} + \frac{C}{b_h}.\end{aligned}$$

This proves that ∇u_h is equi-integrable. Up to a subsequence we may suppose that for all $i = 1,\dots,m$ we have

$$\nabla u_i^h \rightharpoonup g_i \qquad \text{weakly in } L^1(\Omega;\mathbb{R}^N).$$

Since $a_h \to +\infty$, we get $D^c u^h \to 0$ strongly in the sense of measures.

Let $\psi : \mathbb{R} \to \mathbb{R}$ be bounded, Lipschitz, and C^1, and for all $i = 1,\dots,m$ let us consider the measures

$$\mu_i^h(B) := D\psi(u_i^h)(B) - \int_B \psi'(u_i^h)\nabla u_i^h\,dx, \qquad \lambda_i^h(B) := \int_{S(u_i^h)\cap B} \varphi_h(|[u_i^h]|)\,d\mathcal{H}^{N-1},$$

where B is a Borel set in Ω. Notice that $\psi(u_i^h) \in BV(\Omega)$, and that by chain rule in BV (see [4, Theorem 3.96]) we have

$$D\psi(u_i^h) = \psi'(u_i^h)\nabla u_i^h\,d\mathcal{L}^N + (\psi((u_i^h)^+) - \psi((u_i^h)^-))\nu\mathcal{H}^{N-1}\llcorner S(u_i^h) + \psi'(\tilde u_i^h)D^c u_i^h,$$

where $\tilde u_i^h(x)$ is the Lebesgue value of u_i^h at x. We deduce that

$$|D\psi(u_i^h) - \psi'(u_i^h)\nabla u_i^h\,d\mathcal{L}^N| \le ||\psi||_{\varphi_h}\lambda_i^h + ||\psi'||_\infty |D^c u_i^h|, \tag{5.7}$$

where

$$||\psi||_{\varphi_h} := \sup\left\{\frac{\psi(t)-\psi(s)}{\varphi_h(|t-s|)} \,:\, t \ne s\right\}.$$

Up to a subsequence we have

$$\mu_i^h \overset{*}{\rightharpoonup} D\psi(u_i) - \psi'(u)g_i\,d\mathcal{L}^N, \qquad \lambda_i^h \overset{*}{\rightharpoonup} \lambda_i$$

weakly* in the sense of measures, and so from (5.7), by lower semicontinuity for the variations of measures (see [4, Proposition 1.62]) and since $D^c u^h \to 0$ strongly in the sense of measures we get

$$|D\psi(u_i) - \psi'(u) g_i \, d\mathcal{L}^N| \le (\sup \psi - \inf \psi)\lambda_i.$$

As a consequence of SBV characterization (see [4, Proposition 4.12]), we get that $u_i \in SBV(\Omega)$, $\nabla u_i = g_i$, and $\mathcal{H}^{N-1} \llcorner S(u_i) \le \lambda_i$ for all $i = 1, \dots, m$. We deduce that (5.3) holds.

In order to prove (5.4), for every $M > 0$ let g_i^M be the weak limit in $L^1(\Omega)$ (up to a subsequence) of $|\nabla u_i^h| \wedge M$. Since $f_h(\xi) \to |\xi|^2$ uniformly on $[0, M]$, we have

$$\|g_i^M\|^2 \le \liminf_{h \to +\infty} \int_\Omega f_h(\nabla u_i^h) \, dx.$$

Then letting $M \to +\infty$ we obtain (5.4).

Let us come to (5.5). If λ is the weak limit in the sense of measures of

$$\lambda^h(A) := \int_{S(u^h) \cap A} \varphi_h(|[u_1^h]| \vee \cdots \vee |[u_m^h]|) \, d\mathcal{H}^{N-1},$$

we have that $\lambda_i \le \lambda$ for all $i = 1, \dots, m$. Since we have $\mathcal{H}^{N-1} \llcorner S(u_i) \le \lambda_i$ for all $i = 1, \dots, m$, we deduce that $\mathcal{H}^{N-1} \llcorner S(u) \le \lambda$, so that (5.5) is proved. □

5.2. Compactness for the cracks. This subsection is devoted to the proof of a compactness property for the cracks of the discrete-in-time evolutions, which is closely related to the notion of σ^p-convergence of sets defined in [11].

The convergence we propose is related to the energies which appear in the asymptotic study of the size effects, and so it depends on the cracks, but also on the bulk and surface energies, and on the rate at which the Cantor parts of the derivative of the displacements are disappearing.

Let $(K_h)_{h \in \mathbb{N}}$ be a sequence of rectifiable sets in $\Omega \cup \partial_D \Omega$, and let $f_h : \mathbb{R}^N \to [0, +\infty[$ and $\varphi_h : [0, +\infty[\to [0, 1]$ be such that (5.1) and (5.2) hold. Let γ_h be a positive function on K_h such that for all h

$$\int_{K_h} \varphi_h(\gamma_h) \, d\mathcal{H}^{N-1} < C \tag{5.8}$$

for some $C \in [0, +\infty[$, and let $a_h \to +\infty$. Let $g_h, g \in H^1(\Omega)$ be such that $g_h \to g$ strongly in $H^1(\Omega)$. Let us set for all $u \in BV(\Omega)$

$$\mathcal{E}_h(u) := \int_\Omega f_h(\nabla u) \, dx + \int_{S^{g_h}(u)} \varphi_h(|[u]|) \, d\mathcal{H}^{N-1} + a_h |D^c u|(\Omega).$$

Then the following compactness result holds.

PROPOSITION 5.2. *Up to a subsequence there exists a rectifiable set $K \tilde{\subseteq} \Omega \cup \partial_D \Omega$ such that the following facts hold:*

(a) *for all subsequences $(h_k)_{k \in \mathbb{N}}$ and for all $(u_k)_{k \in \mathbb{N}}$ such that $S^{g_{h_k}}(u_k) \tilde{\subseteq} K_{h_k}$, $|[u_k]| \le \gamma_{h_k}$, $\mathcal{E}_{h_k}(u_{h_k}) \le C'$ for some $C' \in [0, +\infty[$, and $u_k \overset{*}{\rightharpoonup} u$ weakly* in $BV(\Omega)$, we have $u \in SBV(\Omega)$, $\nabla u \in L^2(\Omega; \mathbb{R}^N)$, and $S^g(u) \tilde{\subseteq} K$;*

(b) *there exists a countable set D in $SBV(\Omega)$ such that*

$$K = \bigcup_{u\in D} S^g(u), \tag{5.9}$$

where for every $u \in D$ there exists a sequence $(u_h)_{h\in\mathbb{N}}$ in $BV(\Omega)$ with $S^{g_h}(u_h) \tilde{\subseteq} K_h$, $|[u_h]| \le \gamma_h$, $\mathcal{E}_h(u_h) \le C'$ for some $C' \in [0,+\infty[$, and $u_h \overset{*}{\rightharpoonup} u$ weakly in $BV(\Omega)$;*

(c) *we have*

$$\mathcal{H}^{N-1}(K) \le \liminf_{h\to+\infty} \int_{K_h} \varphi_h(\gamma_h)\, d\mathcal{H}^{N-1}. \tag{5.10}$$

Proof. Our approach is based on Γ-convergence (see section 2). In order to deal with $S^g(u)$ as an internal jump, let us consider $\tilde{\Omega} \subseteq \mathbb{R}^N$ open and bounded, such that $\overline{\Omega} \subseteq \tilde{\Omega}$, and let us set $\Omega' := \tilde{\Omega} \setminus \partial_N\Omega$.

Let us consider the functionals $\mathcal{E}'_h : BV(\Omega') \to [0,+\infty]$,

$$\mathcal{E}'_h(u) := \int_{\Omega'} f_h(\nabla u)\, dx + \int_{S(u)} \varphi_h(|[u]|)\, d\mathcal{H}^{N-1} + a_h|D^c u|(\Omega'),$$

if $u \in BV(\Omega')$, $u = g_h$ on $\Omega' \setminus \Omega$, $S(u) \tilde{\subseteq} K_h$, $|[u]| \le \gamma_h$ on K_h, and $\mathcal{E}'_h(u) = +\infty$ otherwise for $u \in BV(\Omega')$. Let us consider on $BV(\Omega')$ the strong topology of $L^1(\Omega')$. By Proposition 2.4, up to a subsequence, $(\mathcal{E}'_h)_{h\in\mathbb{N}}$ Γ-converges to a functional $\mathcal{E}'$. We denote this subsequence still by $(\mathcal{E}'_h)_{h\in\mathbb{N}}$, and we may suppose that the liminf in (5.10) and the liminf along this subsequence are equal.

Let us consider

$$\mathrm{epi}(\mathcal{E}') := \{(u,s) \in BV(\Omega') \times \mathbb{R} : \mathcal{E}'(u) \le s\},$$

and let $\mathcal{D} \subseteq \mathrm{epi}(\mathcal{E}'(t))$ be countable and dense. If $\pi : BV(\Omega') \times \mathbb{R} \to BV(\Omega')$ denotes the projection on the first factor, let $D := \pi(\mathcal{D})$ and let us set

$$K := \bigcup_{u\in D} S(u).$$

Notice that by a truncation argument we may suppose that each $u \in D$ is bounded in $L^\infty(\Omega)$, and moreover that there exists $u_h \in BV(\Omega')$ such that $u_h \overset{*}{\rightharpoonup} u$ weakly* in $BV(\Omega')$ and $\mathcal{E}'_h(u_h) \le C'$ with $C' \in [0,+\infty[$. By Proposition 5.1 and Γ-liminf inequality we deduce that $u \in SBV(\Omega')$ and

$$\|\nabla u\|^2 + \mathcal{H}^{N-1}(S(u)) \le \mathcal{E}'(u). \tag{5.11}$$

Then K is precisely of the form (5.9) once we consider the restriction of u to Ω and recall that $u = g$ on $\Omega' \setminus \Omega$. Thus point (b) is proved.

Let us prove that (5.10) holds. Let $u_1, \dots, u_k \in D$, and let $u_1^h, \dots, u_k^h \in BV(\Omega')$ be such that $u_i^h \overset{*}{\rightharpoonup} u_i$ weakly* in $BV(\Omega')$ and

$$\lim_{h\to+\infty} \mathcal{E}'_h(u_i^h) = \mathcal{E}'(u_i), \qquad i = 1,\dots,k. \tag{5.12}$$

Setting $u^h := (u_1^h, \dots, u_k^h)$, by (5.12) we have

$$\sum_{i=1}^k \int_{\Omega'} f_h(\nabla u_i^h)\, dx + \int_{S(u^h)} \varphi_h(|[u_1^h]| \vee \cdots \vee |[u_k^h]|)\, d\mathcal{H}^{N-1} + a_h|D^c u^h|(\Omega') \le \tilde{C}$$

with $\tilde{C}$ independent of h. By Proposition 5.1 we deduce

$$\mathcal{H}^{N-1}\left(\bigcup_{i=1}^{k} S(u_i)\right) \leq \liminf_{h\to+\infty} \int_{S(u^h)} \varphi_h(|[u_1^h]| \vee \cdots \vee |[u_k^h]|)\, d\mathcal{H}^{N-1}$$
$$\leq \liminf_{h\to+\infty} \int_{K_h} \varphi_h(\gamma_h)\, d\mathcal{H}^{N-1}.$$

Taking the sup over all possible $u_1, \ldots, u_k$ we get

$$\mathcal{H}^{N-1}(K) \leq \liminf_{h\to+\infty} \int_{K_h(t)} \varphi_h(\gamma_h(t))\, d\mathcal{H}^{N-1},$$

so that (5.10) is proved. In particular, by (5.8) we have that $\mathcal{H}^{N-1}(K) < +\infty$.

Let us come to point (a). Let us extend u_k and u to Ω' setting $u_k = g_{h_k}$ and $u = g$ on $\Omega' \setminus \overline{\Omega}$. By Proposition 5.1 we get $u \in SBV(\Omega')$ and $\nabla u \in L^2(\Omega'; \mathbb{R}^N)$. Let us see that $S(u) \tilde{\subseteq} K$. Notice that by Γ-liminf inequality we have

$$\mathcal{E}'(u) \leq \liminf_{k\to+\infty} \mathcal{E}'_{h_k}(u_k) < +\infty,$$

so that $(u, \mathcal{E}'(u)) \in \mathrm{epi}(\mathcal{E}')$. Let $(v_j, s_j) \in \mathcal{D}$ be such that $v_j \to u$ strongly in $L^1(\Omega')$ and $s_j \to \mathcal{E}'(u)$. By truncation, we may assume that u and v_j are uniformly bounded in L^∞. We know that $v_j \in SBV(\Omega')$ for all j, and that by (5.11)

$$\|\nabla v_j\|^2 + \mathcal{H}^{N-1}(S(v_j)) \leq \mathcal{E}'(v_j). \tag{5.13}$$

By lower semicontinuity of $\mathcal{E}'$ we have

$$\mathcal{E}'(u) \leq \liminf_{j\to+\infty} \mathcal{E}'(v_j).$$

Moreover, since $\mathcal{E}'(v_j) \leq s_j$, we deduce

$$\limsup_{j\to+\infty} \mathcal{E}'(v_j) \leq \lim_{j\to+\infty} s_j = \mathcal{E}'(u),$$

so that we have $\mathcal{E}'(v_j) \to \mathcal{E}'(u) < +\infty$. By (5.13) we conclude that $v_j \rightharpoonup u$ weakly in $SBV(\Omega')$: since $S(v_j) \tilde{\subseteq} K$ for all j, and $\mathcal{H}^{N-1}(K) < +\infty$, by Ambrosio's theorem we get $S(u) \tilde{\subseteq} K$. The proof is now complete. □

Remark 5.3. Notice that in the case $f_h(\xi) = |\xi|^p$ ($p \in]1, +\infty[$) and $\varphi_h = 1$, and no Cantor part is considered (i.e., $a_h = +\infty$), Proposition 5.2 gives an alternative proof of the compactness and lower semicontinuity properties of σ^p-convergence of sets formulated in [11].

Notice moreover that the limit set of Proposition 5.2 is contained (up to negligible $\mathcal{H}^{N-1}$ set) in $\Omega \cup \partial_D\Omega$, so that $\partial_N\Omega$ is not involved: this is done in view of the concrete application to quasi-static crack growth, where convergence for the surface energy holds, and so a crack would never approach $\partial_N\Omega$ otherwise but transversally. This can be seen also from an energetic point of view, since the displacement can choose the more convenient boundary datum on $\partial_N\Omega$ without creating a crack on this part of the boundary.

5.3. A generalization of the Transfer of Jump. In this subsection we prove a generalization of the Transfer of Jump theorem of Francfort–Larsen [13] which will be useful in the proof of Theorem 4.1.

Let $f_h : \mathbb{R}^N \to [0,+\infty[$ and $\varphi_h : [0,+\infty[\to [0,1]$ be such that (5.1) and (5.2) hold. Then the following proposition holds.

PROPOSITION 5.4. *Let $(u_h)_{h\in\mathbb{N}}$ be a sequence in $BV(\Omega)$ such that $(\nabla u_h)_{h\in\mathbb{N}}$ is equi-integrable,*

$$\sup_h \int_{S(u_h)} \varphi_h(|[u_h]|)\, d\mathcal{H}^{N-1} \le C, \quad \text{and} \qquad |D^c u_h|(\Omega) \to 0. \tag{5.14}$$

Let $u \in SBV(\Omega)$ be such that $u_h \overset{*}{\rightharpoonup} u$ weakly in $BV(\Omega)$, and let $g_h, g \in H^1(\Omega)$ be such that $g_h \to g$ strongly in $H^1(\Omega)$. Then for all $v \in SBV(\Omega)$ with $\nabla v \in L^2(\Omega;\mathbb{R}^N)$ there exists $v_h \in SBV(\Omega)$ such that $\nabla v_h \to \nabla v$ strongly in $L^2(\Omega;\mathbb{R}^N)$ and*

$$\limsup_{h\to+\infty} \left[\int_{S^{g_h}(v_h)\cup S^{g_h}(u_h)} \varphi_h(|[v_h]| \vee |[u_h]|)\, d\mathcal{H}^{N-1} - \int_{S^{g_h}(u_h)} \varphi_h(|[u_h]|)\, d\mathcal{H}^{N-1} \right] \le \mathcal{H}^{N-1}(S^g(v) \setminus S^g(u)).$$

Proof. In order to deal with $S^g(u)$ as an internal jump, let us consider $\tilde\Omega \subseteq \mathbb{R}^N$ open and bounded, and such that $\overline\Omega \subseteq \tilde\Omega$. Let us set $\Omega' := \tilde\Omega \setminus \partial_N\Omega$. Let $v \in SBV(\Omega)$ with $\nabla v \in L^2(\Omega;\mathbb{R}^N)$ and $\mathcal{H}^{N-1}(S^g(v)) < +\infty$. Let us consider

$$w := v - g, \qquad z := u - g, \qquad z_h := u_h - g_h,$$

and let us extend w, z, z_h to Ω' setting $w = z = z_h = 0$ on $\Omega' \setminus \Omega$. In this setting, we have to find $w_h \in SBV(\Omega')$ such that $w_h \equiv 0$ on $\Omega' \setminus \Omega$, $\nabla w_h \to \nabla w$ strongly in $L^2(\Omega';\mathbb{R}^N)$, and such that

$$\limsup_{h\to+\infty} \left[\int_{S(w_h)\cup S(z_h)} \varphi_h(|[w_h]| \vee |[z_h]|)\, d\mathcal{H}^{N-1} - \int_{S(z_h)} \varphi_h(|[z_h]|)\, d\mathcal{H}^{N-1} \right] \le \mathcal{H}^{N-1}(S(w) \setminus S(z)).$$

Then the result follows considering the restriction of w_h to Ω, and setting $v_h := w_h + g_h$.

The key point in the proof is the following: for all $\varepsilon > 0$ find $\delta > 0$ and $w_h \in SBV(\Omega')$ such that $w_h \equiv 0$ on $\Omega' \setminus \Omega$,

$$\limsup_{h\to+\infty} \|\nabla w_h - \nabla w\|_{L^2(\Omega';\mathbb{R}^N)} \le \varepsilon, \tag{5.15}$$

and

$$\limsup_{h\to+\infty} \mathcal{H}^{N-1}(S(w_h) \setminus K_h^\delta) \le \mathcal{H}^{N-1}(S(w) \setminus S(z)) + \varepsilon, \tag{5.16}$$

where $K_h^\delta := \{x \in S(z_h) : |[z_h]| \ge \delta\}$. In fact if (5.16) holds, noting that by (5.14) we get $\mathcal{H}^{N-1}(K_h^\delta) \le C + 1$ for h large enough, following the decomposition

$$\begin{aligned} S(w_h) \cup S(z_h) &= (S(w_h) \setminus K_h^\delta) \cup (S(w_h) \cap K_h^\delta) \\ &\qquad \cup (K_h^\delta \setminus S(w_h)) \cup [S(z_h) \setminus (S(w_h) \cup K_h^\delta)], \\ S(z_h) &= (K_h^\delta \cap S(w_h)) \cup (K_h^\delta \setminus S(w_h)) \cup (S(z_h) \setminus K_h^\delta), \end{aligned}$$

we have (d_h is defined in (5.2))

$$\begin{aligned}
&\limsup_{h\to+\infty}\left[\int_{S(w_h)\cup S(z_h)}\varphi_h(|[w_h]|\vee|[z_h]|)\,d\mathcal{H}^{N-1}-\int_{S(z_h)}\varphi_h(|[z_h]|)\,d\mathcal{H}^{N-1}\right]\\
&\qquad\le\mathcal{H}^{N-1}(S(w)\setminus S(z))+\varepsilon\\
+&\limsup_{h\to+\infty}\left[\int_{S(w_h)\cap K_h^\delta}\varphi_h(|[w_h]|\vee|[z_h]|)\,d\mathcal{H}^{N-1}-\int_{S(w_h)\cap K_h^\delta}\varphi_h(|[z_h]|)\,d\mathcal{H}^{N-1}\right]\\
&\qquad\le\mathcal{H}^{N-1}(S(w)\setminus S(z))+\varepsilon+\limsup_{h\to+\infty}(1-d_h)\mathcal{H}^{N-1}(S(w_h)\cap K_h^\delta)\\
&\qquad\le\mathcal{H}^{N-1}(S(w)\setminus S(z))+\varepsilon.
\end{aligned}$$

Letting now $\varepsilon\to 0$, and using a diagonal argument we obtain the result.

Let $\varepsilon>0$, and let us prove that we can find $\delta>0$ and $(w_h)_{h\in\mathbb{N}}$ such that (5.15) and (5.16) hold. Following the Transfer of Jump [13, Theorem 2.1], let us fix $G\subseteq\mathbb{R}$ countable and dense: we recall that we have up to a set of $\mathcal{H}^{N-1}$-measure zero

$$S(z)=\bigcup_{c_1,c_2\in G}\partial^*E_{c_1}(z)\cap\partial^*E_{c_2}(z),$$

where $E_c(z):=\{x\in\Omega'\,:\,x\text{ is a Lebesgue point for }z,\ z(x)>c\}$ and ∂^* denotes the essential boundary (see [4, Definition 3.60]). Let us orient ν_z in such a way that $z^-(x)<z^+(x)$ for all $x\in S(z)$, and let us consider

$$J_j:=\left\{x\in S(z)\,:\,z^+(x)-z^-(x)>\frac{1}{j}\right\},$$

with j so large that

$$\mathcal{H}^{N-1}\left(S(z)\setminus J_j\right)<\sigma,$$

where $\sigma>0$. Let U be a neighborhood of $S(z)$ such that

$$|U|<\frac{\sigma}{j^2},\qquad\int_U|\nabla w|^2\,dx<\sigma.$$

Following [13, Theorem 2.1], we can find a finite disjoint collection of closed cubes $\{Q_k\}_{k=1,\dots,n}$ with edge of length $2r_k$, with center $x_k\in S(z)$, and oriented as the normal $\nu(x_k)$ to $S(z)$ at x_k, such that $\bigcup_{k=1}^n Q_k\subseteq U$ and $\mathcal{H}^{N-1}(J_j\setminus\bigcup_{k=1}^n Q_k)\le\sigma$. Let H_k denote the intersection of Q_k with the hyperplane through x_k orthogonal to $\nu(x_k)$. Following [13] we can suppose that the following facts hold:

(a) if $x_k\in\Omega$, then $Q_k\subseteq\Omega$, and if $x_k\in\partial_D\Omega$, then $\partial\Omega\cap Q_k\subseteq\{y+s\nu(x_k)\,:\,y\in H_k,\ s\in[-\frac{\sigma r_k}{2},\frac{\sigma r_k}{2}]\}$;
(b) $\mathcal{H}^{N-1}(S(z)\cap\partial Q_k)=0$;
(c) $r_k^{N-1}<2\mathcal{H}^{N-1}(S(z)\cap Q_k)$;
(d) $z^-(x_k)<c_k^1<c_k^2<z^+(x_k)$ and $c_k^2-c_k^1>\frac{1}{2j}$;
(e) $\mathcal{H}^{N-1}([S(z)\setminus\partial^*E_{c_k^s}(z)]\cap Q_k)<\sigma r_k^{N-1}$ for $s=1,2$;
(f) $\mathcal{H}^{N-1}(\{y\in\partial^*E_{c_k^s}(z)\cap Q_k\,:\,\operatorname{dist}(y,H_k)\ge\frac{\sigma}{2}r_k\})<\sigma r_k^{N-1}$ for $s=1,2$;

(g) if $Q_k^+ := \{x \in Q_k \,|\, (x - x_k) \cdot \nu(x_k) > 0\}$ and $s = 1, 2$,

$$\|1_{E_{c_k^s}(z))\cap Q_k} - 1_{Q_k^+}\|_{L^1(\Omega')} < \sigma^2 r_k^N;$$

(h) $\mathcal{H}^{N-1}\left(\big(S(w) \setminus S(z)\big) \cap Q_k\right) < \sigma r_k^{N-1}$ and $\mathcal{H}^{N-1}(S(w) \cap \partial Q_k) = 0$.

Since $(\nabla z_h)_{h\in\mathbb{N}}$ is equi-integrable, we may assume that U is chosen so that for h large

$$\sum_{k=1}^n \int_{Q_k} |\nabla z_h|\, dx < \frac{\sigma}{j^2}. \tag{5.17}$$

Let $\eta \in]0, 1[$: we claim that there exists $\delta > 0$ such that for all $k = 1, \dots, n$

$$\limsup_{h\to+\infty} |Dz_h|(\{0 < |[z_h]| < \delta\} \cap Q_k) \le \eta |Q_k|. \tag{5.18}$$

Let $M > 0$: by (5.2) there exists $s_h \to 0$ with $\varphi_h(s_h) \to 1$ such that for h large enough

$$Ms \le \varphi_h(s) \qquad \text{for all } s \in [0, s_h].$$

Then we have

$$\begin{aligned} |Dz_h|\left(\{0 < |[z_h]| < s_h\} \cap Q_k\right) &= \int_{\{0<|[z_h]|<s_h\}\cap Q_k} |[z_h]|\, d\mathcal{H}^{N-1} \\ &\le \frac{1}{M} \int_{\{0<|[z_h]|<s_h\}\cap Q_k} \varphi_h(|[z_h]|)\, d\mathcal{H}^{N-1}, \end{aligned}$$

so that we conclude for h large

$$\begin{aligned} |Dz_h|\left(\{0 < |[z_h]| < \delta\} \cap Q_k\right) &\le \frac{1}{M} \int_{\{0<|[z_h]|<s_h\}\cap Q_k} \varphi_h(|[z_h]|)\, d\mathcal{H}^{N-1} \\ &+ \frac{\delta}{\varphi_h(s_h)} \int_{\{s_h\le|[z_h]|<\delta\}\cap Q_k} \varphi_h(|[z_h]|)\, d\mathcal{H}^{N-1} \le \left(\frac{1}{M} + \frac{\delta}{\varphi_h(s_h)}\right) C, \end{aligned}$$

where C is defined in (5.14). Taking the limsup in h and choosing δ small enough and M large enough, we have that (5.18) holds.

Let δ be as in (5.18), and let us set

$$K_h^\delta := \{x \in S(z_h) \,:\, |[z_h]|(x) \ge \delta\}.$$

Then in view of (5.17) and (5.18), since $|D^c z_h|(\Omega') \to 0$, by the Coarea formula for BV functions (see [4, Theorem 3.40]) we have for h large enough

$$\begin{aligned} \sum_{k=1}^n \int_{c_k^1}^{c_k^2} \mathcal{H}^{N-1}\left(\partial^* E_c(z_h) \cap (Q_k \setminus K_h^\delta)\right)\, dc &\le \sum_{k=1}^n |Dz_h|(Q_k \setminus K_h^\delta) \\ &= \sum_{k=1}^n \int_{Q_k} |\nabla z_h|\, dx + \sum_{k=1}^n |Dz_h|(Q_k \cap \{0 < |[z_h]| < \delta\}) + |D^c z_h|\left(\bigcup_{k=1}^n Q_k\right) \\ &\le (1+\eta)\frac{\sigma}{j^2}. \end{aligned}$$

By the Mean Value theorem and by property (d) we get that there exist $c_k^1 < c_k^h < c_k^2$, $k = 1, \dots, n$, such that

$$\sum_{k=1}^n \mathcal{H}^{N-1}\left(\partial^* E_{c_k^h}(z_h) \cap (Q_k \setminus K_h^\delta)\right) \le 2(1+\eta)\frac{\sigma}{j}. \tag{5.19}$$

Following [13], by property (g) we have that for h large

$$\|1_{E_{c_k^h}(z_h)\cap Q_k} - 1_{Q_k^+}\|_{L^1(\Omega')} \le \sigma^2 r_k^N.$$

Then by Fubini's theorem and by the Mean Value theorem, we can find $s_k^+ \in [\frac{\sigma r_k}{2}, \sigma r_k]$ and $s_k^- \in [-\sigma r_k, -\frac{\sigma r_k}{2}]$ such that setting $H_k^+ := \{x = y + s_k^+\nu(x_k),\, y \in H_k\}$ and $H_k^- := \{x = y + s_k^-\nu(x_k),\, y \in H_k\}$ we have

$$\mathcal{H}^{N-1}(H_k^+ \setminus (E_{c_k^h}(z_h)\cap Q_k)) + \mathcal{H}^{N-1}(H_k^- \cap (E_{c_k^h}(z_h)\cap Q_k)) \le 2\sigma r_k^{N-1}.$$

Let R_k be the region between H_k^- and H_k^+, i.e.,

$$R_k := \{x \in Q_k \,:\, x = y + s\nu(x_k),\, y \in H_k,\, s_k^2 \le s \le s_k^1\},$$

and let us indicate by $R_k^+ w$ the reflection in Q_k of $w_{|Q_k^+\setminus R_k}$ with respect to H_k^+, and by $R_k^- w$ the reflection in Q_k of $w_{|Q_k^-\setminus R_k}$ with respect to H_k^-. We can now consider w_h defined in the following way:

$$w_h := \begin{cases} w & \text{on } \Omega' \setminus \bigcup_{k=1}^n R_k, \\ R_k^+ w & \text{on } R_k \cap E_{c_k^h}(z_h), \\ R_k^- w & \text{on } R_k \setminus E_{c_k^h}(z_h). \end{cases}$$

w_h is well defined for σ small, and $w_h = 0$ on $\Omega' \setminus \Omega$. Notice that by construction we have that for h large

$$\|\nabla w_h - \nabla w\|_{L^2(\Omega';\mathbb{R}^N)} + \sum_{k=1}^n \mathcal{H}^{N-1}((S(w_h) \setminus K_h^\delta)\cap Q_k) \le e(\sigma),$$

where $e(\sigma) \to 0$ as $\sigma \to 0$: the proof follows analyzing the set $S(w_h)$ inside Q_k, and it is very similar to that contained in [13, Theorem 2.1]. Since

$$S(w_h) \setminus K_h^\delta \tilde{\subseteq} [S(w) \setminus S(z)] \cup \left[S(z) \setminus \bigcup_{k=1}^n Q_k\right] \cup \bigcup_{k=1}^n ((S(w_h) \setminus K_h^\delta)\cap Q_k),$$

we deduce

$$\limsup_{h\to+\infty} \mathcal{H}^{N-1}\left(S(w_h) \setminus K_h^\delta\right) \le \mathcal{H}^{N-1}\left(S(w) \setminus S(z)\right) + e(\sigma),$$

with $e(\sigma) \to 0$ as $\sigma \to 0$. Choosing σ small enough and using a diagonal argument, we obtain that (5.15) and (5.16) hold, and the proof is finished. □

The following proposition extends the Transfer of Jump to the case of cracks converging in the sense of Proposition 5.2.

PROPOSITION 5.5. *Let $(K_h, \gamma_h)_{h\in\mathbb{N}}$ and K be as in Proposition 5.2, and let $g_h, g \in H^1(\Omega)$ be such that $g_h \to g$ strongly in $H^1(\Omega)$. Then for all $v \in SBV(\Omega)$ with $\nabla v \in L^2(\Omega;\mathbb{R}^N)$ there exists $v_h \in SBV(\Omega)$ such that $\nabla v_h \to \nabla v$ strongly in $L^2(\Omega;\mathbb{R}^N)$ and*

$$\limsup_{h\to+\infty} \left[\int_{S^{g_h}(v_h)\cup K_h} \varphi_h(|[v_h]| \vee \gamma_h)\, d\mathcal{H}^{N-1} - \int_{K_h} \varphi_h(\gamma_h)\, d\mathcal{H}^{N-1}\right] \le \mathcal{H}^{N-1}(S^g(v) \setminus K).$$

Proof. We indicate how to modify the proof of Proposition 5.4 in order to get the result for $(K_h, \gamma_h)_{h\in\mathbb{N}}$ and K.

Notice that properties (5.15) and (5.16) can be extended to the case of a finite number of converging sequences: more precisely if $k = 1, \dots, m$, $(u_h^k)_{h\in\mathbb{N}}$ is a sequence in $BV(\Omega)$ such that $u_h^k \overset{*}{\rightharpoonup} u^k$ weakly* in $BV(\Omega)$, $(\nabla u_h^k)_{h\in\mathbb{N}}$ is equi-integrable,

$$\int_{S^{g_h}(u_h^k)} \varphi_h(|[u_h^k]|)\, d\mathcal{H}^{N-1} \le C, \qquad |D^c u_h^k|(\Omega) \to 0,$$

then for every $\varepsilon > 0$ and $v \in SBV(\Omega)$ with $\nabla v \in L^2(\Omega;\mathbb{R}^N)$ there exists $v_h \in SBV(\Omega)$ such that

$$\limsup_{h\to+\infty} \|\nabla v_h - \nabla v\|_{L^2(\Omega;\mathbb{R}^N)} \le \varepsilon \tag{5.20}$$

and

$$\limsup_{h\to+\infty} \mathcal{H}^{N-1}(S^{g_h}(v_h) \setminus \tilde{K}_h^\delta) \le \mathcal{H}^{N-1}\left(S^g(v) \setminus \bigcup_{k=1}^m S^g(u^k)\right) + \varepsilon, \tag{5.21}$$

where $\tilde{K}_h^\delta := \{x \in \bigcup_{k=1}^m S^{g_h}(u_h^k) : |[u_h^k]|(x) \ge \delta \text{ for some } k = 1, \dots, m\}$. This can be done using the localization on the squares already employed in [13, Theorem 2.3]: on each squares Q_j we have that $\bigcup_{k=1}^m S^g(u^k) \cap Q_j$ is essentially given by $S^g(u^{\tau(j)})$ for some $\tau(j) \in \{1, \dots, m\}$.

Let us come to the Transfer of Jump for K. We recall that

$$K = \bigcup_{u\in D} S^g(u)$$

for some countable set D, and that each $u \in D$ is limit in the weak* topology of $BV(\Omega)$ of a function u_h such that $S^{g_h}(u_h) \tilde{\subseteq} K_h$, $|[u_h]| \le \gamma_h$, $\nabla u_h \rightharpoonup \nabla u$ weakly in $L^1(\Omega;\mathbb{R}^N)$,

$$\sup_h \int_{S(u_h)} \varphi_h(|[u_h]|)\, d\mathcal{H}^{N-1} \le C, \quad \text{and} \qquad |D^c u_h|(\Omega) \to 0.$$

Let $\varepsilon > 0$ be fixed: since $\mathcal{H}^{N-1}(K) < +\infty$, we can find m such that

$$\mathcal{H}^{N-1}\left(K \setminus \bigcup_{k=1}^m S^g(u^k)\right) \le \varepsilon \tag{5.22}$$

for some $u^k \in D$, $k = 1, \dots, m$. Let u_h^k be the approximation of u^k for all $k = 1, \dots, m$, and let $v \in SBV(\Omega)$ with $\nabla v \in L^2(\Omega;\mathbb{R}^N)$. Then by (5.20) and (5.21) we can find $(v_h)_{h\in\mathbb{N}}$ such that

$$\limsup_{h\to+\infty} \|\nabla v_h - \nabla v\|_{L^2(\Omega;\mathbb{R}^N)} \le \varepsilon$$

and

$$\limsup_{h\to+\infty} \mathcal{H}^{N-1}(S^{g_h}(v_h) \setminus \tilde{K}_h^\delta) \le \mathcal{H}^{N-1}\left(S^g(v) \setminus \bigcup_k S^g(u^k)\right) + \varepsilon.$$

Setting $K_h^\delta := \{x \in K_h : \gamma_h(x) \geq \delta\}$, recalling that $\tilde{K}_h^\delta \tilde{\subseteq} K_h^\delta$ since $|[u_h^k]| \leq \gamma_h$, by (5.22) we deduce that

$$\limsup_{h\to+\infty} \mathcal{H}^{N-1}\left(S^{g_h}(v_h) \setminus K_h^\delta\right) \leq \mathcal{H}^{N-1}\left(S^g(v) \setminus K\right) + 2\varepsilon.$$

The proof now follows exactly as in Proposition 5.4. □

6. Proof of Theorem 4.1. In this section we will give the proof of Theorem 4.1. Let $\{t \to (u_h(t), \Gamma_h(t), \psi_h(t)) : t \in [0,T]\}$ be the piecewise constant interpolation of a discrete-in-time evolution of cracks in Ω_h relative to the subdivision $I_{\delta_h} := \{0 = t_0^{\delta_h} < \cdots < t_{N_{\delta_h}}^{\delta_h} = T\}$, and the boundary displacement $\sqrt{t}g(t, \frac{x}{h})$ given in (4.3). We divide the proof in several steps.

Step 1: Rescaling. For all $t \in [0,T]$ let $v_h(t) \in BV(\Omega)$ and $K_h(t) \tilde{\subseteq} \Omega \cup \partial_D\Omega$ be defined as

$$v_h(t,x) := \frac{1}{\sqrt{h}} u_h(t, hx), \qquad K_h(t) := \frac{1}{h}\Gamma_h(t). \tag{6.1}$$

Let us moreover set

$$\gamma_h(t,x) := \frac{1}{\sqrt{h}}\psi_h(t,hx) = \max_{0\leq s\leq t} |[v_h(s)](t,x)|, \qquad t \in [0,T], x \in \Omega. \tag{6.2}$$

We notice that $\{t \to (v_h(t), K_h(t), \gamma_h(t)) : t \in [0,T]\}$ is the piecewise constant interpolation of a discrete-in-time evolution of cracks in Ω relative to the subdivision I_{δ_h} and boundary displacement $g(t)$ with respect to the basic total energy

$$\int_\Omega f_h(\nabla v)\, dx + \int_{S^{g^{\delta_h}(t)}(v)\cup K_h(t)} \varphi_h(|[v]| \vee \gamma_h(t))\, d\mathcal{H}^{N-1} + a\sqrt{h}|D^c v|(\Omega),$$

where $g^{\delta_h}(t) := g(t_i^{\delta_h})$ for $t_i^{\delta_h} \leq t < t_{i+1}^{\delta_h}$, $a := \varphi'(0)$,

$$\varphi_h(s) := \varphi(\sqrt{h}s), \qquad s \in [0, +\infty[, \tag{6.3}$$

and

$$f_h(\xi) := \begin{cases} |\xi|^2 & \text{if } |\xi| \leq \frac{a\sqrt{h}}{2}, \\ \frac{a^2 h}{4} + a\sqrt{h}(|\xi| - \frac{a\sqrt{h}}{2}) & \text{if } |\xi| \geq \frac{a\sqrt{h}}{2}. \end{cases} \tag{6.4}$$

Let us recall some properties of the evolution $\{t \to (v_h(t), K_h(t), \gamma_h(t)) : t \in [0,T]\}$ which are derived from Proposition 3.1 and that will be employed in what follows:

(a) for all $t \in [0,T]$

$$\|v_h(t)\|_\infty \leq \|g^{\delta_h}(t)\|_\infty; \tag{6.5}$$

(b) $K_h(0) = S^{g^{\delta_h}(0)}(v_h(0))$ and $S^{g^{\delta_h}(t)}(v_h(t)) \tilde{\subseteq} K_h(t)$ for all $t \in]0,T]$;

(c) for all $w \in BV(\Omega)$ we have

$$\begin{aligned} &\int_\Omega f_h(\nabla v_h(0))\, dx + \int_{S^{g^{\delta_h}(0)}(v_h(0))} \varphi_h(|[v_h(0)]|)\, d\mathcal{H}^{N-1} + a\sqrt{h}|D^c v_h(0)|(\Omega) \\ &\leq \int_\Omega f_h(\nabla w)\, dx + \int_{S^{g^{\delta_h}(0)}(w)} \varphi_h(|[w]|)\, d\mathcal{H}^{N-1} + a\sqrt{h}|D^c w|(\Omega); \end{aligned} \tag{6.6}$$

(d) for all $w \in BV(\Omega)$ and for all $t \in]0,T]$ we have

$$\int_\Omega f_h(\nabla v_h(t))\,dx + \int_{K_h(t)} \varphi_h(\gamma_h(t))\,d\mathcal{H}^{N-1} + a\sqrt{h}|D^c v_h(t)|(\Omega) \leq \int_\Omega f_h(\nabla w)\,dx + \int_{S^{g^{\delta_h}(t)}(w)\cup K_h(t)} \varphi_h(|[w]| \vee \gamma_h(t))\,d\mathcal{H}^{N-1} + a\sqrt{h}|D^c w|(\Omega). \tag{6.7}$$

Let us set for all $w \in BV(\Omega)$ and for all $t \in [0,T]$

$$\mathcal{F}_h(t,w) := \int_\Omega f_h(\nabla w)\,dx + \int_{S^{g^{\delta_h}(t)}(w)\cup K_h(t)} \varphi_h(|[w]| \vee \gamma_h(t))\,d\mathcal{H}^{N-1} + a\sqrt{h}|D^c w|(\Omega). \tag{6.8}$$

Notice that for all $t \in [0,T]$

$$\mathcal{F}_h(t, v_h(t)) = \frac{1}{h^{N-1}}\mathcal{E}_h(t, u_h(t)), \tag{6.9}$$

where $\mathcal{E}_h(t,u)$ is defined in (4.4).

Recalling Lemma 3.2, for all $t \in [0,T]$ we have

$$\mathcal{F}_h(t, v_h(t)) \leq \mathcal{F}_h(0, v_h(0)) + \int_0^{t_h}\int_\Omega f_h'(\nabla v_h(\tau))\nabla \dot g(\tau)\,dx\,d\tau + e(h), \tag{6.10}$$

where $e(h) \to 0$ as $h \to +\infty$, and $t_h := t_{i_h}^{\delta_h}$ is the step discretization point of I_{δ_h} such that $t_{i_h}^{\delta_h} \leq t < t_{i_h+1}^{\delta_h}$.

Step 2: Uniform bound on the energy. There exists a constant C' independent of h such that for all $t \in [0,T]$ we have

$$\mathcal{F}_h(t, v_h(t)) + \|v_h(t)\|_\infty \leq C'. \tag{6.11}$$

In fact by (6.6) we have

$$\mathcal{F}_h(0, u_h(0)) \leq \|\nabla g(0)\|^2,$$

and by (6.7) for all $\tau \in [0,T]$,

$$\int_\Omega f_h(\nabla v_h(\tau))\,dx \leq \|\nabla g^{\delta_h}(\tau)\|^2.$$

Moreover, for all $\tau \in [0,T]$

$$\int_\Omega |f_h'(\nabla v_h(\tau))|^2\,dx \leq 4\int_\Omega f_h(\nabla v_h(\tau))\,dx.$$

Taking into account (6.10) and (6.5) we deduce that (6.11) holds.

Step 3: Compactness. In view of Step 2, by Propositions 5.1 and 5.2 we have that for all $t \in [0,T]$ the displacements $(v_h(t))_{h\in\mathbb{N}}$ are relatively compact with respect to the weak* topology of $BV(\Omega)$, while the cracks $(K_h(t), \gamma_h(t))_{h\in\mathbb{N}}$ are compact in a suitable energetic sense.

Let $B \subseteq [0,T]$ be countable and dense, and such that $0 \in B$. By Propositions 5.1 and 5.2 (with f_h and φ_h defined in (6.4) and (6.3), $a_h := a\sqrt{h}$, $\gamma_h := \gamma_h(t)$, and $g_h := g^{\delta_h}$) up to a subsequence (which we denote by the same symbol) for all $t \in B$ there exists $v(t) \in SBV(\Omega)$ and a rectifiable set $K(t) \tilde\subseteq \Omega \cup \partial_D\Omega$ such that the following facts hold:

(a) $v_h(t) \overset{*}{\rightharpoonup} v(t)$ in the weak* topology of $BV(\Omega)$, $\nabla v_h(t) \rightharpoonup \nabla v(t)$ weakly in $L^1(\Omega;\mathbb{R}^N)$, $\nabla v(t) \in L^2(\Omega;\mathbb{R}^N)$, and

$$S^{g(t)}(v(t)) \tilde{\subseteq} K(t);$$

(b) $K(s) \tilde{\subseteq} K(t)$ for all $s,t \in B$, $s \le t$;

(c) we have

$$\mathcal{H}^{N-1}(K(t)) \le \liminf_{h\to+\infty} \int_{K_h(t)} \varphi_h(\gamma_h(t))\, d\mathcal{H}^{N-1}; \tag{6.12}$$

(d) $K(0) = S^{g(0)}(v(0))$.

Points (a) and (c) comes directly from Propositions 5.1 and 5.2. Let us prove point (b). Let $s,t \in B$ with $s < t$. By Proposition 5.2 we know that there exists a countable set $D(s)$ in $SBV(\Omega)$ such that

$$K(s) = \bigcup_{u\in D} S^{g(s)}(u), \tag{6.13}$$

and such that for every $u \in D(s)$ there exists a sequence $(u_h)_{h\in\mathbb{N}}$ in $BV(\Omega)$ such that $u_h \overset{*}{\rightharpoonup} u$ weakly* in $BV(\Omega)$ with $S^{g^{\delta_h}(s)}(u_h) \tilde{\subseteq} K_h(s)$, $|[u_h]| \le \gamma_h(s)$, and $\mathcal{F}_h(s,u_h) \le C'$ for some $C' \in [0,+\infty[$. Let us set $v_h := u_h - g^{\delta_h}(s) + g^{\delta_h}(t)$. Since $K_h(s) \tilde{\subseteq} K_h(t)$ and $\gamma_h(s) \le \gamma_h(t)$, we have that $S^{g^{\delta_h}(t)}(v_h) \tilde{\subseteq} K_h(t)$, $|[v_h]| \le \gamma_h(t)$,

$$\int_\Omega f_h(\nabla v_h)\, dx + \int_{K_h(t)} \varphi_h(|[v_h]|)\, d\mathcal{H}^{N-1} + a\sqrt{h}|D^c v_h|(\Omega) \le \tilde{C}'$$

with $\tilde{C}'$ independent of h, and $v_h \overset{*}{\rightharpoonup} u - g(s) + g(t)$ weakly* in $BV(\Omega)$. We deduce that $S^{g(t)}(u - g(s) + g(t)) \tilde{\subseteq} K(t)$, that is, $S^{g(s)}(u) \tilde{\subseteq} K(t)$. Then by (6.13) we obtain $K(s) \tilde{\subseteq} K(t)$.

Let us come to point (d). Notice that

$$\begin{aligned} \|\nabla v(0)\|^2 + \mathcal{H}^{N-1}(K(0)) &\le \liminf_{h\to+\infty} \mathcal{F}_h(0, v_h(0)) \\ &\le \|\nabla v(0)\|^2 + \mathcal{H}^{N-1}(S^{g(0)}(v(0))), \end{aligned} \tag{6.14}$$

the first inequality coming from point (c) and Proposition 5.1, the last inequality coming from the minimality property (6.6). Since $S^{g(0)}(v(0)) \tilde{\subseteq} K(0)$, by (6.14) we get that $S^{g(0)}(v(0)) = K(0)$, so that point (d) is proved.

Step 4: Recovering the static equilibrium for $K(t)$, $t \in B$. Let B be the countable and dense set defined in Step 3, and let $K(t)$ be the limit crack associated with $(K_h(t), \gamma_h(t))_{h\in\mathbb{N}}$ for all $t \in B$. In order to prove that $K(t)$ is part of an evolution in the sense of [13] with respect to the boundary data $g(t)$, we have to prove that $K(t)$ satisfies the one-sided minimality property with respect to the Griffith's energy given by point (c) of Theorem 2.2. This is done in this step, where also some useful convergence results for the gradient of the displacements are obtained.

Let $t \in B$, and let us consider the subsequence of $(v_h(t), K_h(t), \gamma_h(t))_{h\in\mathbb{N}}$ (which we indicate with the same symbol), the displacement $v(t)$ and the rectifiable set $K(t)$ given by Step 3. Then for all $v \in SBV(\Omega)$ we have

$$\|\nabla v(0)\|^2 + \mathcal{H}^{N-1}(S^{g(0)}(v(0))) \le \|\nabla v\|^2 + \mathcal{H}^{N-1}(S^{g(0)}(v)), \tag{6.15}$$

and for all $t \in]0,T]$

$$\|\nabla v(t)\|^2 \le \|\nabla v\|^2 + \mathcal{H}^{N-1}(S^{g(t)}(v) \setminus K(t)). \tag{6.16}$$

Moreover,

$$\|\nabla v(0)\|^2 + \mathcal{H}^{N-1}(K(0)) = \lim_{h\to+\infty} \mathcal{F}_h(0, v_h(0)), \tag{6.17}$$

where $\mathcal{F}_h$ is defined in (6.8), and for all $t \in B$

$$\nabla v_h(t) 1_{E_h(t)} \to \nabla v(t) \qquad \text{strongly in } L^2(\Omega;\mathbb{R}^N), \tag{6.18}$$

where

$$E_h(t) := \left\{ x \in \Omega \,:\, |\nabla v_h(t)| \le \frac{a\sqrt{h}}{2} \right\}$$

and

$$\|\nabla v(t)\|^2 = \lim_{h\to+\infty} \int_\Omega f_h(\nabla v_h(t))\, dx. \tag{6.19}$$

In fact (6.15) and (6.17) come from point (d) of Step 3, from the minimality property (6.6), and from Proposition 5.1.

Let us come to (6.16). Let $t \in [0,T]$. By Proposition 5.5 we have that there exists $(v_h)_{h\in\mathbb{N}}$ sequence in $SBV(\Omega)$ such that $\nabla v_h \to \nabla v$ strongly in $L^2(\Omega;\mathbb{R}^N)$ and

$$\limsup_{h\to+\infty} \left[\int_{S^{g^{\delta_h(t)}}(v_h)\cup K_h(t)} \varphi_h(|[v_h]| \vee \gamma_h(t))\, d\mathcal{H}^{N-1} - \int_{K_h(t)} \varphi_h(\gamma_h(t))\, d\mathcal{H}^{N-1} \right] \le \mathcal{H}^{N-1}(S^{g(t)}(v) \setminus K(t)).$$

Then using the minimality property (6.7) we get

$$\limsup_{h\to+\infty} \int_\Omega f_h(\nabla v_h(t))\, dx \le \|\nabla v\|^2 + \mathcal{H}^{N-1}(S^{g(t)}(v) \setminus K(t)). \tag{6.20}$$

By Proposition 5.1 we have that

$$\|\nabla v(t)\|^2 \le \liminf_{h\to+\infty} \int_\Omega f_h(\nabla v_h(t))\, dx, \tag{6.21}$$

and so we obtain that (6.16) holds.

Let us now come to (6.18) and (6.19). Equation (6.19) is a direct consequence of (6.21) and (6.20) with $v = v(t)$. Finally, notice that $(\nabla v_h(t) 1_{E_h(t)})_{h\in\mathbb{N}}$ is bounded in $L^2(\Omega;\mathbb{R}^N)$. Since $\nabla v_h(t) \rightharpoonup \nabla v(t)$ weakly in $L^1(\Omega;\mathbb{R}^N)$ and $\nabla v(t) \in L^2(\Omega;\mathbb{R}^N)$, we get $\nabla v_h(t) 1_{E_h(t)} \rightharpoonup \nabla v(t)$ weakly in $L^2(\Omega;\mathbb{R}^N)$. By (6.20) with $v = v(t)$ we have

$$\limsup_{h\to+\infty} \|\nabla v_h(t) 1_{E_h(t)}\|^2 \le \limsup_{h\to+\infty} \int_\Omega f_h(\nabla v_h(t))\, dx \le \|\nabla v(t)\|^2,$$

so that (6.18) holds.

Step 5: Defining $K(t)$ for all $t \in [0,T]$. Since $\{t \to K(t) : t \in B\}$ is increasing by Step 3, setting

$$K^-(t) := \bigcup_{s\in B, s\le t} K(s), \qquad K^+(t) := \bigcap_{s\in B, s\ge t} K(s),$$

there exists a countable set $B' \subseteq [0,T]\setminus B$ such that we have $K^-(t) \tilde{=} K^+(t)$ for all $t \in [0,T]\setminus B'$. For all such t's let us set $K(t) := K^-(t) \tilde{=} K^+(t)$. Up to a further subsequence relative to the elements of B' (which we indicate still with the same symbol), we find $K(t)$ such that Steps 3 and 4 hold for every $t \in B'$. Notice that

$$\{t \to K(t) : t \in [0,T]\}$$

is increasing, and for all $t \in [0,T]$ we have $\mathcal{H}^{N-1}(K(t)) \le C'$, where C' is given by (6.11).

Let $v(t)$ be a minimum of the following problem:

$$\min\{\|\nabla v\|^2 : v \in SBV(\Omega),\ S^{g(t)}(v) \tilde{\subseteq} K(t)\}. \tag{6.22}$$

Notice that problem (6.22) is well posed since $K(t)$ has finite $\mathcal{H}^{N-1}$-measure, and $g(t)$ is bounded in $L^\infty(\Omega)$: moreover by strict convexity we have that $\nabla v(t)$ is uniquely determined.

Let us prove that $(v(t), K(t))$ satisfies Steps 3 and 4 for every $t \in [0,T]$. Moreover, let us see that

$$\nabla v_h(t) \rightharpoonup \nabla v(t) \qquad \text{weakly in } L^1(\Omega;\mathbb{R}^N), \tag{6.23}$$

and that every accumulation point v of $(v_h(t))_{h\in\mathbb{N}}$ in the weak* topology of $BV(\Omega)$ is such that $v \in SBV(\Omega)$, $S^{g(t)}(v) \tilde{\subseteq} K(t)$, and $\nabla v = \nabla v(t)$.

In fact let $t \notin B \cup B'$ (otherwise the result holds by construction), and let $v_{h_m}(t) \overset{*}{\rightharpoonup} v$ weakly* in $BV(\Omega)$ for some subsequence $(h_m)_{m\in\mathbb{N}}$. By Proposition 5.1 we get that $v \in SBV(\Omega)$, $\nabla v \in L^2(\Omega;\mathbb{R}^N)$, and $\nabla v_{h_m}(t) \rightharpoonup \nabla v$ weakly in $L^1(\Omega;\mathbb{R}^N)$.

Applying Steps 3 and 4 to $B \cup \{t\}$, we can find (up to a further subsequence) $\tilde{K}(t)$ such that $S^{g(t)}(v) \tilde{\subseteq} \tilde{K}(t)$, $\tilde{K}(t)$ satisfies static equilibrium, and

$$K(s_1) \tilde{\subseteq} \tilde{K}(t) \tilde{\subseteq} K(s_2)$$

for all $s_1, s_2 \in B$ with $s_1 < t < s_2$. Then we get $\tilde{K}(t) = K(t)$ up to a set of $\mathcal{H}^{N-1}$-measure zero.

Finally, in order to prove that (6.23) holds, notice that v is a minimum of problem (6.22): by uniqueness we obtain $\nabla v = \nabla v(t)$ so that (6.23) holds along the entire sequence.

Step 6: Recovering the nondissipativity condition. In order to prove that

$$\{t \to (v(t), K(t)),\ t \in [0,T]\}$$

is a quasi-static crack growth in the sense of [13], that is, in the sense of Theorem 2.2, we have just to prove the *nondissipativity* condition, that is,

$$\mathcal{E}(t) = \mathcal{E}(0) + 2\int_0^t (\nabla v(\tau), \nabla \dot{g}(\tau))_{L^2(\Omega;\mathbb{R}^N)}\, d\tau, \tag{6.24}$$

where $\mathcal{E}(t) := \|\nabla v(t)\|^2 + \mathcal{H}^{N-1}(K(t))$ for all $t \in [0,T]$. In fact *irreversibility* and *static equilibrium* are consequences of Steps 3, 4, and 5. First of all, for all $t \in [0,T]$ we have

$$\mathcal{E}(t) \geq \mathcal{E}(0) + 2\int_0^t (\nabla u(\tau), \nabla \dot{g}(\tau))_{L^2(\Omega;\mathbb{R}^N)}\, d\tau. \tag{6.25}$$

In fact as noticed in [15], using the minimality property (6.16), the map $\{t \to \nabla v(t)\}$ is continuous at all the continuity points of $\{t \to \mathcal{H}^{N-1}(K(t))\}$, in particular it is continuous up to a countable set in $[0,T]$. Given $t \in [0,T]$ and $k > 0$, let us set

$$s_i^k := \frac{i}{k}t, \qquad v^k(s) := v(s_{i+1}^k) \quad \text{for } s_i^k < s \leq s_{i+1}^k, \qquad i = 0, 1, \ldots, k.$$

By (6.16), comparing $v(s_i^k)$ with $v(s_{i+1}^k) - g(s_{i+1}^k) + g(s_i^k)$, it is easy to see that

$$\mathcal{E}(t) \geq \mathcal{E}(0) + 2\int_0^t (\nabla v^k(\tau), \nabla \dot{g}(\tau))_{L^2(\Omega;\mathbb{R}^N)}\, d\tau + e(k),$$

where $e(k) \to 0$ as $k \to +\infty$. By the continuity property of ∇v, passing to the limit for $k \to +\infty$ we deduce that (6.25) holds. On the other hand, for all $t \in [0,T]$ we have that

$$\mathcal{E}(t) \leq \mathcal{E}(0) + 2\int_0^t (\nabla u(\tau), \nabla \dot{g}(\tau))_{L^2(\Omega;\mathbb{R}^N)}\, d\tau. \tag{6.26}$$

In fact by Step 4 we have that for all $t \in [0,T]$

$$\nabla v_h(t) 1_{E_h(t)} \to \nabla v(t) \qquad \text{strongly in } L^2(\Omega;\mathbb{R}^N), \tag{6.27}$$

where

$$E_h(t) := \left\{ x \in \Omega \,:\, |\nabla v_h(t)| \leq \frac{a\sqrt{h}}{2} \right\}.$$

By (6.10) and by the very definition of f_h we deduce

$$\begin{aligned} \mathcal{F}_h(t, v_h(t)) \leq \mathcal{F}_h(0, v_h(0)) &+ 2\int_0^t (\nabla v_h(\tau) 1_{E_h(\tau)}, \nabla \dot{g}(\tau))_{L^2(\Omega;\mathbb{R}^N)}\, d\tau \\ &+ a\sqrt{h}\int_0^t \int_{\Omega \setminus E_h(\tau)} |\nabla \dot{g}(\tau)|\, dx\, d\tau + e(h), \end{aligned} \tag{6.28}$$

where $e(h) \to 0$ as $h \to +\infty$. Notice that by (6.11) we have

$$\frac{a}{2} h |\Omega \setminus E_h(\tau)| \leq \sqrt{h} \int_{\Omega \setminus E_h(\tau)} |\nabla v_h(\tau)|\, dx \leq \frac{2}{a} \int_{\Omega \setminus E_h(\tau)} f_h(\nabla v_h(\tau))\, dx \leq \frac{2}{a} C'.$$

We deduce that

$$\begin{aligned} \sqrt{h} \int_{\Omega \setminus E_h(\tau)} |\nabla \dot{g}(\tau)|\, dx &\leq \left(\int_{\Omega \setminus E_h(\tau)} |\nabla \dot{g}(\tau)|^2\, dx \right)^{\frac{1}{2}} \sqrt{h |\Omega \setminus E_h(\tau)|} \\ &\leq \frac{2\sqrt{C'}}{a} \left(\int_{\Omega \setminus E_h(\tau)} |\nabla \dot{g}(\tau)|^2\, dx \right)^{\frac{1}{2}} \to 0 \end{aligned} \tag{6.29}$$

as $h \to +\infty$ by equicontinuity of $\nabla \dot g(\tau)$. Then passing to the limit for $h \to +\infty$ in (6.28), in view of (6.19), (6.12), (6.17), (6.27), and (6.29) we deduce that (6.26) holds. This proves that (6.24) holds, and so $\{t \to (v(t), K(t)) : t \in [0,T]\}$ is a quasi-static crack growth in the sense of [13].

Step 7: Convergence of bulk and surface energies. In order to conclude the proof, let us see that (4.5), (4.6), and (4.7) hold. By (6.28) we deduce that for all $t \in [0,T]$

$$\mathcal{F}_h(t, v_h(t)) \to \mathcal{E}(t),$$

so that by (6.19) and (6.12) we deduce that

$$\mathcal{H}^{N-1}(K(t)) = \lim_{h\to+\infty} \int_{K_h(t)} \varphi_h(\gamma_h(t))\, d\mathcal{H}^{N-1}, \qquad a\sqrt{h}|D^c v_h(t)|(\Omega) \to 0.$$

Theorem 4.1 is now completely proved in view of the rescaling (6.1), of (6.2), (6.3), and (6.9).

7. Proof of Theorem 4.2. In this section we will give the proof of Theorem 4.2. Let $\{t \to (u_h(t), \Gamma_h(t), \psi_h(t)) : t \in [0,T]\}$ be the piecewise constant interpolation given in (4.3) of a discrete-in-time evolution of crack in Ω_h relative to the subdivision $I_{\delta_h} := \{0 = t_0^{\delta_h} < \cdots < t_{N_{\delta_h}}^{\delta_h} = T\}$, and the boundary displacement $h^\alpha g(t, \frac{x}{h})$ with $\alpha \in \,]0, \frac{1}{2}[$. We divide the proof in several steps.

Step 1: Rescaling. We rescale u_h and Γ_h in the following way: for all $t \in [0,T]$ let $v_h(t) \in BV(\Omega)$ and $K_h(t) \tilde\subseteq \Omega \cup \partial_D\Omega$ be given by

$$v_h(t,x) := \frac{1}{h^\alpha} u_h\,(t, hx)\,, \qquad K_h(t) := \frac{1}{h}\Gamma_h(t), \qquad t \in [0,T], x \in \Omega. \tag{7.1}$$

Let us moreover set

$$\gamma_h(t,x) := \frac{1}{h^\alpha}\psi_h(t,hx) = \max_{0\le s\le t} |[v_h(s)](t,x)|, \qquad t \in [0,T], x \in \Omega.$$

It turns out that $\{t \to (v_h(t), K_h(t), \gamma_h(t)) : t \in [0,T]\}$ is the piecewise constant interpolation of a discrete-in-time evolution of cracks in Ω relative to the subdivision I_{δ_h} and boundary displacement $g(t)$ with respect to the basic total energy

$$\int_\Omega f_h(\nabla v)\, dx + h^{1-2\alpha} \int_{S^{g^{\delta_h}(t)}(v)\cup K_h(t)} \varphi_h(|[v]| \vee \gamma_h(t))\, d\mathcal{H}^{N-1} + ah^{1-\alpha}|D^c v|(\Omega),$$

where $g^{\delta_h}(t) := g(t_i^{\delta_h})$ for $t_i^{\delta_h} \le t < t_{i+1}^{\delta_h}$, $a := \varphi'(0)$,

$$\varphi_h(s) := \varphi(h^\alpha s), \qquad s \in [0, +\infty[,$$

and

$$f_h(\xi) := \begin{cases} |\xi|^2 & \text{if } |\xi| \le \frac{ah^{1-\alpha}}{2}, \\ \frac{a^2h^{2(1-\alpha)}}{4} + ah^{1-\alpha}(|\xi| - \frac{ah^{1-\alpha}}{2}) & \text{if } |\xi| \ge \frac{ah^{1-\alpha}}{2}. \end{cases}$$

We have that the following facts hold:

(a) for all $t \in [0,T]$

$$\|v_h(t)\|_\infty \le \|g^{\delta_h}(t)\|_\infty \le C; \tag{7.2}$$

(b) $K_h(0) = S^{g^{\delta_h}(0)}(v_h(0))$, and $S^{g^{\delta_h}(t)}(v_h(t)) \tilde{\subseteq} K_h(t)$ for all $t \in]0,T]$;
(c) for all $w \in BV(\Omega)$ we have

$$\begin{aligned}\int_\Omega f_h(\nabla v_h(0))\,dx + h^{1-2\alpha}\int_{S^{g^{\delta_h}(0)}(v_h(0))}\varphi_h(|[v_h(0)]|)\,d\mathcal{H}^{N-1}\\ + ah^{1-\alpha}|D^c v_h(0)|(\Omega)\\ \le \int_\Omega f_h(\nabla w)\,dx + h^{1-2\alpha}\int_{S^{g^{\delta_h}(0)}(w)}\varphi_h(|[w]|)\,d\mathcal{H}^{N-1} + ah^{1-\alpha}|D^c w|(\Omega);\end{aligned}$$

(d) for all $w \in BV(\Omega)$ and $t \in]0,T]$ we have

$$\begin{aligned}\int_\Omega f_h(\nabla v_h(t))\,dx + h^{1-2\alpha}\int_{K_h(t)}\varphi_h(\gamma_h(t))\,d\mathcal{H}^{N-1} + ah^{1-\alpha}|D^c v_h(t)|(\Omega)\\ \le \int_\Omega f_h(\nabla w)\,dx + h^{1-2\alpha}\int_{S^{g^{\delta_h}(t)}(w)\cup K_h(t)}\varphi_h(|[w]|\vee\gamma_h(t))\,d\mathcal{H}^{N-1}\\ + ah^{1-\alpha}|D^c w|(\Omega).\end{aligned}$$

Let us set for all $v \in BV(\Omega)$ and for all $t \in [0,T]$

$$\begin{aligned}\mathcal{F}_h(t,w) := \int_\Omega f_h(\nabla w)\,dx + h^{1-2\alpha}\int_{S^{g^{\delta_h}(t)}(w)\cup K_h(t)}\varphi_h(|[w]|\vee\gamma_h(t))\,d\mathcal{H}^{N-1}\\ + ah^{1-\alpha}|D^c w|(\Omega).\end{aligned}$$

Notice that

$$\mathcal{F}_h(t,v_h(t)) = \frac{1}{h^{N+2\alpha-2}}\mathcal{E}(t,u_h(t)),$$

where $\mathcal{E}(t,u_h(t))$ is defined in (4.4).

By Lemma 3.2 we obtain for all $t \in [0,T]$

$$\mathcal{F}_h(t,v_h(t)) \le \mathcal{F}_h(0,v_h(0)) + \int_0^{t_h}\int_\Omega f_h'(\nabla v_h(\tau))\nabla\dot{g}(\tau)\,dx\,d\tau + e(h), \tag{7.3}$$

where $e(h) \to 0$ as $h \to +\infty$, and $t_h := t_{i_h}^{\delta_h}$ is the step discretization point of I_{δ_h} such that $t_{i_h}^{\delta_h} \le t < t_{i_h+1}^{\delta_h}$.

Step 2: Uniform bound on the energy. By point (c) comparing $v_h(0)$ and $g(0)$ we have

$$\int_\Omega f_h(\nabla v_h(0))\,dx + h^{1-2\alpha}\int_{S^{g(0)}(v_h(0))}\varphi_h(|[v_h(0)]|) + ah^{1-\alpha}|D^c v_h(0)|(\Omega) \le \|\nabla g(0)\|^2.$$

By point (d) comparing $v_h(t)$ and $g^{\delta_h}(t)$ we obtain

$$\int_\Omega f_h(\nabla v_h(t))\,dx \le \|\nabla g^{\delta_h}(t)\|^2,$$

and since we have

$$\int_\Omega |f_h'(\nabla v_h(\tau))|^2\,dx \le 4\int_\Omega f_h(\nabla v_h(\tau))\,dx,$$

by (7.3) we deduce that

$$\int_\Omega f_h(\nabla v_h(t))\,dx + h^{1-2\alpha}\int_{K_h(t)}\varphi_h(\gamma_h(t))\,d\mathcal{H}^{N-1} + ah^{1-\alpha}|D^c v|(\Omega) \le C'$$

with C' independent of h and of t. By Proposition 5.1 and by (7.2) we deduce that $(v_h(t))_{h\in\mathbb{N}}$ is bounded in $BV(\Omega)$, and this proves point (a).

Step 3: Convergence to the elastic solution. Let $v(t)$ be an accumulation point for $(v_h(t))_{h\in\mathbb{N}}$ in the weak* topology of $BV(\Omega)$, and let us consider $\tilde\Omega \subseteq \mathbb{R}^N$ open and bounded, and such that $\overline\Omega \subseteq \tilde\Omega$. Let us set $\Omega' := \tilde\Omega \setminus \partial_N\Omega$. Then we can extend $v_h(t)$ and $v(t)$ to Ω' setting $v_h(t) = g^{\delta_h}(t)$ and $v(t) = g(t)$ on $\Omega'\setminus\Omega$, respectively. We have $v_{h_j}(t) \overset{*}{\rightharpoonup} v(t)$ weakly* in $BV(\Omega')$ for a suitable $h_j \nearrow +\infty$, and

$$\begin{aligned}(7.4)\qquad &\int_{\Omega'} f_{h_j}(\nabla v_{h_j}(t))\,dx + h_j^{1-2\alpha}\int_{S(v_{h_j}(t))}\varphi_{h_j}(|[v_{h_j}(t)]|)\,d\mathcal{H}^{N-1}\\ &\qquad + ah_j^{1-\alpha}|D^c v_{h_j}(t)|(\Omega') \le \tilde C\end{aligned}$$

with $\tilde C$ independent of j. In particular, we have

$$\int_{\Omega'} f_{h_j}(\nabla v_{h_j}(t))\,dx + \int_{S(v_{h_j}(t))}\varphi_{h_j}(|[v_{h_j}(t)]|)\,d\mathcal{H}^{N-1} + ah_j^{1-\alpha}|D^c v_{h_j}(t)|(\Omega') \le \tilde C'$$

with $\tilde C'$ independent of j. Then by Proposition 5.1 we have that $v(t)\in SBV(\Omega)$,

$$\nabla v_{h_j}(t) \rightharpoonup \nabla v(t) \qquad \text{weakly in } L^1(\Omega;\mathbb{R}^N),$$

and

$$(7.5)\qquad \|\nabla v(t)\|^2 \le \liminf_{j\to+\infty}\int_\Omega f_{h_j}(\nabla v_{h_j}(t))\,dx.$$

Finally, if we consider for all Borel sets $B\subseteq\Omega'$

$$\lambda_j(B) := \int_{B\cap S(v_{h_j}(t))}\varphi_{h_j}(|[v_{h_j}(t)]|)\,d\mathcal{H}^{N-1}$$

and if (up to a subsequence) $\lambda_j \overset{*}{\rightharpoonup} \lambda$ weakly* in the sense of measures, we deduce following Proposition 5.1 that

$$\mathcal{H}^{N-1}\llcorner S(v(t)) \le \lambda \qquad \text{as measures.}$$

Since by (7.4) we have $\lambda = 0$, then we have $S(v(t)) = \emptyset$, that is, $v(t)\in H^1(\Omega)$ and $v(t) = g(t)$ on $\partial_D\Omega$.

Let us consider $v\in H^1(\Omega)$ with $v = g(t)$ on $\partial_D\Omega$. Comparing $v_h(t)$ with $v - g(t) + g^{\delta_h}(t)$ by minimality property of point (d) we obtain

$$\begin{aligned}(7.6)\qquad \int_\Omega f_h(\nabla v_h(t))\,dx + ah^{1-\alpha}|D^c v_h(t)|(\Omega) &\le \int_\Omega f_h(\nabla v - \nabla g(t) + \nabla g^{\delta_h}(t))\,dx\\ &\le \|\nabla v - \nabla g(t) + \nabla g^{\delta_h}(t)\|^2.\end{aligned}$$

In view of (7.5) we deduce that

$$\|\nabla v(t)\|^2 \le \|\nabla v\|^2,$$

so that $v(t)$ is a minimizer of

$$\min\{\|\nabla v\|^2 \,:\, v \in H^1(\Omega),\, v = g(t) \text{ on } \partial_D\Omega\}.$$

By strict convexity and since Ω is connected, we have that $v(t)$ is uniquely determined, and so we deduce that $v_h(t) \overset{*}{\rightharpoonup} v(t)$ weakly* in $BV(\Omega)$ and $\nabla v_h(t) \rightharpoonup \nabla v(t)$ weakly in $L^1(\Omega;\mathbb{R}^N)$.

Choosing $v = v(t)$ in (7.6) and taking the limsup in h we have

$$\limsup_{h\to+\infty} \int_\Omega f_h(\nabla v_h(t))\,dx \le \|\nabla u(t)\|^2,$$

so that

$$\lim_{h\to+\infty} \int_\Omega f_h(\nabla v_h(t))\,dx = \|\nabla u(t)\|^2.$$

The proof of point (b) is now concluded thanks to the rescaling (7.1).

8. Proof of Theorem 4.3. In this section we will give the proof of Theorem 4.3. Let $\{t \to (u_h(t), \Gamma_h(t), \psi_h(t)) \,:\, t \in [0,T]\}$ be the piecewise constant interpolation given in (4.3) of a discrete-in-time evolution of crack in Ω_h relative to the subdivision $I_{\delta_h} := \{0 = t_0^{\delta_h} < \cdots < t_{N_{\delta_h}}^{\delta_h} = T\}$, and the boundary displacement $h^\alpha g(t, \frac{x}{h})$ with $\alpha > \frac{1}{2}$.

We rescale u_h and Γ_h in the following way: for all $t \in [0,T]$ let $v_h(t) \in BV(\Omega)$ and $K_h(t) \tilde{\subseteq} \Omega \cup \partial_D\Omega$ be given by

$$v_h(t,x) := \frac{1}{h^\alpha} u_h\,(t, hx)\,, \qquad K_h(t) := \frac{1}{h}\Gamma_h(t), \qquad t \in [0,T], x \in \Omega.$$

Let us moreover set

$$\gamma_h(t,x) := \frac{1}{h^\alpha}\psi_h(t,hx) = \max_{0\le s\le t} |[v_h(s)](t,x)|, \qquad t \in [0,T], x \in \Omega.$$

It turns out that $\{t \to (v_h(t), K_h(t), \gamma_h(t)) \,:\, t \in [0,T]\}$ is the piecewise constant interpolation of a discrete-in-time evolution of cracks in Ω relative to the subdivision I_{δ_h} and boundary displacement $g(t)$ with respect to the basic total energy

$$h^{2\alpha-1}\int_\Omega f_h(\nabla v)\,dx + \int_{S^{g^{\delta_h}(t)}(v)} \varphi_h(|[v]| \vee \gamma_h(t))\,d\mathcal{H}^{N-1} + ah^\alpha |D^c v|(\Omega),$$

where $g^{\delta_h}(t) := g(t_i^{\delta_h})$ for $t_i^{\delta_h} \le t < t_{i+1}^{\delta_h}$, $a := \varphi'(0)$,

$$\varphi_h(s) := \varphi(h^\alpha s), \qquad s \in [0, +\infty[,$$

and

$$f_h(\xi) := \begin{cases} |\xi|^2 & \text{if } |\xi| \le \frac{ah^{1-\alpha}}{2}, \\ \frac{a^2h^{2(1-\alpha)}}{4} + ah^{1-\alpha}(|\xi| - \frac{ah^{1-\alpha}}{2}) & \text{if } |\xi| \ge \frac{ah^{1-\alpha}}{2}. \end{cases}$$

Notice that by Proposition 3.1 we have

$$\|v_h(0)\|_\infty \le \|g(0)\|_\infty \le C, \tag{8.1}$$

and for all $w \in BV(\Omega)$ we have

$$h^{2\alpha-1}\int_\Omega f_h(\nabla v_h(0))\,dx + \int_{S^{g^{\delta_h(0)}(v_h(0))}} \varphi_h(|[v_h(0)]|)\,d\mathcal{H}^{N-1} + ah^\alpha |D^c v_h(0)|(\Omega)$$
$$\le h^{2\alpha-1}\int_\Omega f_h(\nabla w)\,dx + \int_{S^{g^{\delta_h(0)}(w)}} \varphi_h(|[w]|)\,d\mathcal{H}^{N-1} + ah^\alpha |D^c w|(\Omega). \tag{8.2}$$

Comparing $v_h(0)$ and $w = -C$ by means of (8.2) we have

$$h^{2\alpha-1}\int_\Omega f_h(\nabla v_h(0))\,dx + \int_{S^{g^{\delta_h(0)}(v_h(0))}} \varphi_h(|[v_h(0)]|)\,d\mathcal{H}^{N-1} + ah^\alpha |D^c v_h(0)|(\Omega)$$
$$\le \mathcal{H}^{N-1}(\partial_D\Omega). \tag{8.3}$$

As a consequence, since $\|v_h(0)\|_\infty \le C$ by (8.1), following Proposition 5.1, we deduce that $(v_h(0))_{h\in\mathbb{N}}$ is bounded in $BV(\Omega)$. Let v be an accumulation point for $(v_h(0))_{h\in\mathbb{N}}$ in the weak* topology of $BV(\Omega)$. Let us prove that $v \in SBV(\Omega)$ and that $\nabla v = 0$: in fact we have that for all $\xi \in \mathbb{R}^N$

$$\tilde{f}_h(\xi) \le h^{2\alpha-1} f_h(\xi),$$

where

$$\tilde{f}_h(\xi) := \begin{cases} |\xi|^2 & \text{if } |\xi| \le \frac{ah^\alpha}{2}, \\ \frac{a^2h^{2\alpha}}{4} + ah^\alpha(|\xi| - \frac{ah^\alpha}{2}) & \text{if } |\xi| \ge \frac{ah^\alpha}{2}. \end{cases}$$

We deduce that there exists C'' independent of h such that for all h

$$\int_\Omega \tilde{f}_h(\nabla v_h(0))\,dx + \int_{S(v_h(0))} \varphi_h(|[v_h(0)]|)\,d\mathcal{H}^{N-1} + ch^\alpha |D^c v_h(0)|(\Omega) \le C''.$$

By Proposition 5.1, we obtain that $v \in SBV(\Omega)$ and that $\nabla v_h(0) \rightharpoonup \nabla v$ weakly in $L^1(\Omega;\mathbb{R}^N)$. By (8.3) we obtain that

$$\|\nabla v_h(0)\|_{L^1(\Omega;\mathbb{R}^N)} \le \frac{\mathcal{H}^{N-1}(\partial_D\Omega) + 1}{ah^\alpha},$$

so that we deduce $\nabla v = 0$; that is, v is piecewise constant in Ω. Finally taking the limit in (8.2) with w piecewise constant, by Proposition 5.1 we get exactly (4.8), so that the proof of Theorem 4.3 is concluded.

9. Appendix. In this section, we prove a relaxation result we used in order to study the discrete-in-time evolution of cracks in the cohesive case. It consists of a variant of a result by Bouchitté, Braides, and Buttazzo [6]: the difference here is that we have to take into account the presence of a preexisting crack with a given opening which enters in the surface part of the energy.

Let $f : \mathbb{R} \to [0, +\infty[$ be convex, $f(0) = 0$, and with superlinear growth, i.e.,

$$\limsup_{|\xi|\to+\infty} \frac{f(\xi)}{|\xi|} = +\infty.$$

Let $\varphi : [0,+\infty[\to [0,+\infty[$ be increasing, concave, and such that $\varphi(0) = 0$. Notice that if $a := \varphi'(0) < +\infty$, we have

$$\varphi(s) \le as \qquad \text{for all } s \in [0,+\infty[.$$

Let Ω be a Lipschitz bounded open set in $\mathbb{R}^N$, and let $\partial_D\Omega \subseteq \partial\Omega$ be open in the relative topology. Let Γ be a rectifiable set in $\Omega \cup \partial_D\Omega$, and let ψ be a positive function defined on Γ. Let us extend ψ to $\Omega \cup \partial_D\Omega$ setting $\psi = 0$ outside Γ. Let $g \in W^{1,1}(\Omega)$: we may assume that g is extended to the whole $\mathbb{R}^N$, and we indicate this extension still by g.

We will study the following functional:

$$F(u) := \begin{cases} \int_\Omega f(|\nabla u|)\,dx + \int_{S^g(u)\cup\Gamma} \varphi(|[u]| \vee \psi)\,d\mathcal{H}^{N-1} & \text{if } u \in SBV(\Omega), \\ +\infty & \text{otherwise in } BV(\Omega), \end{cases}$$

where $S^g(u)$ is defined in (2.3), and $a \vee b := \max\{a,b\}$ for all $a,b \in \mathbb{R}$. The functional F naturally appears (see section 3) when dealing with quasi-static growth of cracks in the cohesive case, where one is required to look for its minima. We are led to compute the relaxation of F with respect to the strong topology of $L^1(\Omega)$. The relaxation in the case $\Gamma = \emptyset$ (without boundary conditions but without superlinear growth on f) has been proved in [6]. Let

$$f_1(\xi) := \inf\{f(\xi_1) + a|\xi_2| \,:\, \xi_1 + \xi_2 = \xi\}, \tag{9.1}$$

where $a := \varphi'(0)$. We have that the following result holds.

PROPOSITION 9.1. *The relaxation of the functional F with respect to the weak* topology of $BV(\Omega)$ is given by $\overline{F} : BV(\Omega) \to [0,+\infty]$ defined as*

$$\overline{F}(u) := \int_\Omega f_1(|\nabla u|)\,dx + \int_{S^g(u)\cup\Gamma} \varphi(|[u]| \vee \psi)\,d\mathcal{H}^{N-1} + a|D^c u|,$$

where $a = \varphi'(0)$ and f_1 is defined in (9.1).

In order to prove Proposition 9.1, the first step is the following lemma.

LEMMA 9.2. *Let $\bar{F} : BV(\Omega) \to [0,+\infty]$ be defined by*

$$\overline{F}(u) := \int_\Omega f_1(|\nabla u|)\,dx + \int_{S^g(u)\cup\Gamma} \varphi(|[u]| \vee \psi)\,d\mathcal{H}^{N-1} + a|D^c u|,$$

with $a = \varphi'(0)$ and f_1 as in (9.1). *Then $\bar{F}$ is lower semicontinuous with respect to the weak* topology of $BV(\Omega)$.*

The proof of Lemma 9.2 is obtained by a standard slicing argument (see, for example, [4, Theorem 5.4]) based on the lower semicontinuity result in dimension one. We establish this last one.

Let $I \subseteq \mathbb{R}$ be a finite union of disjoint intervals, and let $J \subseteq I$ be a countable set. Let us consider the functional

$$\begin{aligned} \mathcal{F}(\mu) := \int_I f_1(|\phi_\mu|)\,dx + \sum_{t\in S_\mu\setminus J} \varphi(|\mu(\{t\})|) \\ + \sum_{t\in J} \varphi(|\mu(\{t\})| \vee \psi(t)) + a|\mu^c|(I) \end{aligned} \tag{9.2}$$

defined for all $\mu \in \mathcal{M}_b(I;\mathbb{R}^k)$, i.e., μ is a bounded $\mathbb{R}^k$-valued Radon measure on I. Here ϕ_μ is the density of the absolutely continuous part μ^a of μ, S_μ is the set of atoms

of μ, $\mu^c := \mu - \mu^a - \mu \llcorner S_\mu$, ψ is a strictly positive function defined on J, $a = \varphi'(0)$, and f_1 is defined in (9.1).

LEMMA 9.3. *The functional $\mathcal{F}$ defined in* (9.2) *is lower semicontinuous with respect to the weak* convergence in the sense of measures.*

Proof. Since $\mathcal{F}$ can be obtained as the sup of functionals of the form (9.2) with J finite, we may assume that $J = \{x_1, \dots, x_m\}$. Let $\mu_n \overset{*}{\rightharpoonup} \mu$ weakly* in the sense of measures, and let λ be the weak* limit (up to a subsequence) of $|\mu_n \llcorner J|$. Let $J := J_1 \cup J_2$, with

$$J_1 := \{t \in J \,:\, |\mu(\{t\})| \geq \psi(t)\}, \qquad J_2 := J \setminus J_1.$$

Let $\varepsilon > 0$ be such that

$$\bigcup_{x_i \in J_2} \bar{B}_\varepsilon(x_i) \subseteq I$$

and such that for all n

$$|\mu_n| \left(\bigcup_{x_i \in J_2} \partial \bar{B}_\varepsilon(x_i) \right) = |\mu| \left(\bigcup_{x_i \in J_2} \partial \bar{B}_\varepsilon(x_i) \right) = 0.$$

Let us set

$$I_1 := I \setminus \bigcup_{x_i \in J_2} \bar{B}_\varepsilon(x_i), \qquad I_2 := \bigcup_{x_i \in J_2} B_\varepsilon(x_i).$$

Let $\mathcal{F}_1$ and $\mathcal{F}_2$ denote the restriction of $\mathcal{F}$ to $\mathcal{M}_b(I_1; \mathbb{R}^k)$ and $\mathcal{M}_b(I_2; \mathbb{R}^k)$, respectively. We have

$$\liminf_{n \to +\infty} \mathcal{F}(\mu_n) \geq \liminf_{n \to +\infty} \mathcal{F}_1(\mu_n \llcorner I_1) + \liminf_{n \to +\infty} \mathcal{F}_2(\mu_n \llcorner I_2).$$

We notice that

$$\mathcal{F}_1(\mu_n \llcorner I_1) \geq \mathcal{G}_1(\mu_n \llcorner I_1),$$

where

$$\mathcal{G}_1(\eta) := \int_{I_1} f_1(|\phi_\eta|)\, dx + \sum_{t \in S_\eta} \varphi(|\eta(\{t\})|) + a|\eta^c|(I_1)$$

for all $\eta \in \mathcal{M}_b(I_1; \mathbb{R}^k)$. By [4, Thorem 5.2] we have that

$$\mathcal{G}_1(\mu \llcorner I_1) \leq \liminf_{n \to +\infty} \mathcal{G}_1(\mu_n \llcorner I_1),$$

so that

$$\mathcal{F}_1(\mu \llcorner I_1) = \mathcal{G}_1(\mu \llcorner I_1) \leq \liminf_{n \to +\infty} \mathcal{F}_1(\mu_n \llcorner I_1).$$

On the other hand, we have

$$\mathcal{F}_2(\mu_n \llcorner I_2) = \mathcal{G}_2(\mu_n \llcorner I_2 \setminus J_2) + \sum_{t \in J_2} \varphi(|\mu_n(\{t\})| \vee \psi(t)),$$

where

$$\mathcal{G}_2(\eta) := \int_{I_2} f_1(|\phi_\eta|)\,dx + \sum_{t\in S_\eta} \varphi(|\eta(\{t\})|) + a|\eta^c|(I_2)$$

for all $\eta \in \mathcal{M}_b(I_2;\mathbb{R}^k)$. We have

$$\begin{aligned}\liminf_{n\to+\infty} \mathcal{F}_2(\mu_n \llcorner I_2) &\geq \mathcal{G}_2(\mu \llcorner I_2 \setminus J_2) + \sum_{t\in J_2} \varphi(\lambda(\{t\}) \vee \psi(t)) \\ &\geq \mathcal{G}_2(\mu \llcorner I_2 \setminus J_2) + \sum_{t\in J_2} \varphi(\psi(t)).\end{aligned}$$

We deduce

$$\mathcal{F}_2(\mu \llcorner I_2) = \mathcal{G}_2(\mu \llcorner I_2 \setminus J_2) + \sum_{t\in J_2} \varphi(\psi(t)) \leq \liminf_{n\to+\infty} \mathcal{F}_2(\mu_n \llcorner I_2),$$

and so we get

$$\mathcal{F}(\mu) = \mathcal{F}_1(\mu \llcorner I_1) + \mathcal{F}_2(\mu \llcorner I_2) \leq \liminf_{n\to+\infty} \mathcal{F}(\mu_n).$$

The proof is now concluded. □

Let us now come to the proof of Proposition 9.1.

Proof of Proposition 9.1. We can assume without loss of generality that

$$\int_\Gamma \varphi(\psi)\,d\mathcal{H}^{N-1} < +\infty.$$

Following Lemma 9.2, let us consider $\tilde{\Omega}$ open and bounded in $\mathbb{R}^N$ such that $\overline{\Omega} \subset \tilde{\Omega}$, and let us set $\Omega' := \tilde{\Omega} \setminus \partial_N\Omega$. Let us consider the functional

$$F'(u) := \begin{cases} \int_\Omega f(|\nabla u|)\,dx + \int_{S(u)\cup\Gamma} \varphi(|[u]| \vee \psi)\,d\mathcal{H}^{N-1} & \text{if } u \in SBV(\Omega'), \\ & u = g \text{ on } \Omega' \setminus \Omega, \\ +\infty & \text{otherwise in } BV(\Omega'). \end{cases}$$

The relaxation result of Proposition 9.1 is equivalent to proving that the relaxation of F' under the weak* topology of $BV(\Omega')$ is

$$\overline{F'}(u) := \int_\Omega f_1(|\nabla u|)\,dx + \int_{S(u)\cup\Gamma} \varphi(|[u]| \vee \psi)\,d\mathcal{H}^{N-1} + a|D^c u|(\Omega')$$

if $u \in BV(\Omega')$, $u = g$ on $\Omega' \setminus \Omega$, and $\overline{F'}(u) = +\infty$ otherwise in $BV(\Omega')$.

Following [6], it is useful to introduce the localized version of F'; namely, for all open set $A \subseteq \Omega'$ let us set

$$F'(u,A) := \int_{A\cap\Omega} f(|\nabla u|)\,dx + \int_{A\cap(S(u)\cup\Gamma)} \varphi(|[u]| \vee \psi)\,d\mathcal{H}^{N-1} \tag{9.3}$$

if $u \in SBV(\Omega')$, $u = g$ on $\Omega' \setminus \Omega$, and $F'(u,A) = +\infty$ otherwise in $BV(\Omega')$. Let us indicate by $\overline{F'}(u,A)$ the relaxation of (9.3) under the weak* topology of $BV(\Omega')$.

Arguing as in [6, Proposition 3.3], we have that for every $u \in BV(\Omega')$, $\overline{F'}(u,\cdot)$ is the restriction to the family $\mathcal{A}(\Omega')$ of all open subsets of Ω' of a regular Borel measure. Since for all $u \in SBV(\Omega')$ with $u = g$ on $\Omega' \setminus \Omega$ and for all $A \in \mathcal{A}(\Omega')$ we have

$$\begin{aligned}\int_{A\cap\Omega} f(|\nabla u|)\,dx + \int_{A\cap S(u)} \varphi(|[u]|)\,d\mathcal{H}^{N-1} &\le F'(u,A)\\ &\le \int_{A\cap\Omega} f(|\nabla u|)\,dx + \int_{A\cap S(u)} \varphi(|[u]|)\,d\mathcal{H}^{N-1} + \int_{A\cap\Gamma} \varphi(\psi)\,d\mathcal{H}^{N-1},\end{aligned}$$

by [6, Theorem 3.1] we obtain that for all $u \in BV(\Omega')$ with $u = g$ on $\Omega' \setminus \Omega$ and for all $A \in \mathcal{A}(\Omega')$ with $A \cap \partial_D\Omega = \emptyset$

$$\begin{aligned}\int_{A\cap\Omega} f_1(|\nabla u|)\,dx + \int_{A\cap S(u)} \varphi(|[u]|)\,d\mathcal{H}^{N-1} + a|D^c u|(A) &\le F'(u,A)\\ &\le \int_{A\cap\Omega} f_1(|\nabla u|)\,dx + \int_{A\cap S(u)} \varphi(|[u]|)\,d\mathcal{H}^{N-1} + a|D^c u|(A) + \int_{A\cap\Gamma} \varphi(\psi)\,d\mathcal{H}^{N-1}.\end{aligned} \tag{9.4}$$

As a consequence of (9.4), we deduce that

$$\overline{F'}(u,\cdot)\llcorner(\Omega' \setminus (S(u)\cup\Gamma\cup\partial_D\Omega)) = f_1(|\nabla u|)\,d\mathcal{L}^N\llcorner\Omega + a|D^c u|.$$

In order to evaluate $\overline{F'}(u,\cdot)\llcorner(S(u)\cup\Gamma\cup\partial_D\Omega)$, we notice that for all $A \in \mathcal{A}(\Omega')$ and for all $u \in SBV(\Omega')$ with $u = g$ on $\Omega' \setminus \Omega$

$$\int_{A\cap\Omega} f_1(|\nabla u|)\,dx + \int_{A\cap(S(u)\cup\Gamma)} \varphi(|[u]|\vee\psi)\,d\mathcal{H}^{N-1} + a|D^c u|(A) \le F'(u,A),$$

and since the left-hand side is lower semicontinuous by Lemma 9.2, we get that for all $u \in BV(\Omega')$ with $u = g$ on $\Omega' \setminus \Omega$

$$\int_{A\cap\Omega} f_1(|\nabla u|)\,dx + \int_{A\cap(S(u)\cup\Gamma)} \varphi(|[u]|\vee\psi)\,d\mathcal{H}^{N-1} + a|D^c u|(A) \le \overline{F'}(u,A).$$

By outer regularity of $\overline{F'}(u,\cdot)$ we conclude that

$$\overline{F'}(u,E) \ge \int_E \varphi(|[u]|\vee\psi)\,d\mathcal{H}^{N-1}$$

for all Borel sets $E \subseteq S(u)\cup\Gamma\cup\partial_D\Omega$. We have to prove the opposite inequality, namely,

$$\overline{F'}(u,E) \le \int_E \varphi(|[u]|\vee\psi)\,d\mathcal{H}^{N-1}$$

for all Borel sets $E \subseteq S(u)\cup\Gamma\cup\partial_D\Omega$. Without loss of generality, we may assume that

$$\int_{S(u)} \varphi(|[u]|)\,d\mathcal{H}^{N-1} < +\infty,$$

and by a truncation argument, we can suppose that $u_{|\Omega} \in L^\infty(\Omega)$. Let K be a compact subset of $S(u)\cup\Gamma\cup\partial_D\Omega$, $\varepsilon > 0$, and let A_ε be open with $K \subseteq A_\varepsilon$ and

$$|Du|(A_\varepsilon \setminus K) < \varepsilon, \qquad \int_{(A_\varepsilon\setminus K)\cap\Gamma} \varphi(\psi)\,d\mathcal{H}^{N-1} < \varepsilon.$$

We can find $u_h \in BV(\Omega')$ with $u_h = g$ on $\Omega' \setminus \Omega$ and such that u_h is piecewise constant in Ω (that is $(u_h)_{|\Omega} \in SBV(\Omega)$ with $\nabla u_h = 0$ in Ω), $u_h \to u$ strongly in $L^\infty(\Omega)$, and $|Du_h|(A_\varepsilon \setminus K) < \varepsilon$. Since u_h is piecewise constant in Ω we have for all h

$$\overline{F'}(u_h, A_\varepsilon) \le \int_{A_\varepsilon \cap (S(u_h) \cup \Gamma)} \varphi(|[u_h]| \vee \psi)\, d\mathcal{H}^{N-1}. \tag{9.5}$$

We conclude

$$\begin{aligned}
\overline{F'}(u, A_\varepsilon) &\le \liminf_{h \to +\infty} \overline{F'}(u_h, A_\varepsilon) \le \liminf_{h \to +\infty} \int_{A_\varepsilon \cap (S(u_h) \cup \Gamma)} \varphi(|[u_h]| \vee \psi)\, d\mathcal{H}^{N-1} \\
&\le \int_{K \cap (S(u) \cup \Gamma)} \varphi(|[u]| \vee \psi)\, d\mathcal{H}^{N-1} + a|Du_h|(A_\varepsilon \setminus K) + \int_{(A_\varepsilon \setminus K) \cap \Gamma} \varphi(\psi)\, d\mathcal{H}^{N-1} \\
&\le \int_{K \cap (S(u) \cup \Gamma)} \varphi(|[u]| \vee \psi)\, d\mathcal{H}^{N-1} + (a+1)\varepsilon
\end{aligned}$$

so that, letting $\varepsilon \to 0$ we obtain

$$\overline{F'}(u, K) \le \int_{K \cap (S(u) \cup \Gamma)} \varphi(|[u]| \vee \psi)\, d\mathcal{H}^{N-1}.$$

Since K is arbitrary in $S(u) \cup \Gamma \cup \partial_D \Omega$, the proof is concluded. □

Acknowledgments. The author wishes to thank Gianni Dal Maso for many helpful and interesting discussions, and Gilles Francfort for useful comments on a preliminary version of the paper.

REFERENCES

[1] L. Ambrosio, *A compactness theorem for a new class of functions of bounded variations*, Boll. Un. Mat. Ital. (B) 7, 3 (1989), pp. 857–881.
[2] L. Ambrosio, *Existence theory for a new class of variational problems*, Arch. Ration. Mech. Anal., 111 (1990), pp. 291–322.
[3] L. Ambrosio, *A new proof of the SBV compactness theorem*, Calc. Var. Partial Differential Equations, 3 (1995), pp. 127–137.
[4] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of Bounded Variations and Free Discontinuity Problems*, Clarendon Press, Oxford, 2000.
[5] G. I. Barenblatt, *The mathematical theory of equilibrium cracks in brittle fracture*, Adv. Appl. Mech., 7 (1962), pp. 55–129.
[6] G. Bouchitté, A. Braides, and G. Buttazzo, *Relaxation results for some free discontinuity problems*, J. Reine Angew. Math., 458 (1995), pp. 1–18.
[7] H. Brezis, *Opérateurs Maximaux Monotones et Semi-groupes de Contractions Dans Les Espaces de Hilbert*, North–Holland, Amsterdam, 1973.
[8] A. Carpinteri, *Size effects on strength, toughness, and ductility*, J. Engrg. Mech., 115 (1989), pp. 1375–1392.
[9] A. Chambolle, *A density result in two-dimensional linearized elasticity, and applications*, Arch. Ration. Mech. Anal., 167 (2003), pp. 211–233.
[10] G. Dal Maso, *An Introduction to Γ-Convergence*, Birkhäuser, Boston, 1993.
[11] G. Dal Maso, G. A. Francfort, and R. Toader, *Quasi-static crack growth in nonlinear elasticity*, Arch. Ration. Mech. Anal., to appear.
[12] G. Dal Maso and R. Toader, *A model for the quasistatic growth of brittle fractures: Existence and approximation results*, Arch. Ration. Mech. Anal., 162 (2002), pp. 101–135.
[13] G. A. Francfort and C. J. Larsen, *Existence and convergence for quasistatic evolution in brittle fracture*, Comm. Pure Appl. Math., 56 (2003), pp. 1465–1500.
[14] G. A. Francfort and J.-J. Marigo, *Revisiting brittle fractures as an energy minimization problem*, J. Mech. Phys. Solids, 46 (1998), pp. 1319–1342.

[15] A. Giacomini and M. Ponsiglione, *A discontinuous finite element approximation of quasistatic growth of brittle fractures*, Numer. Funct. Anal. Optim., 24 (2003), pp. 813–850.
[16] J. N. Goodier, *Mathematical theory of equilibrium cracks*, in Fracture: An Advanced Treatise, Vol. II, Mathematical Fundamentals, H. Liebowitz, ed., Academic Press, New York, 1968, pp. 1–66.
[17] J. R. Rice, *Mathematical analysis in the mechanics of fracture*, in Fracture: An Advanced Treatise, Vol. II, Mathematical Fundamentals, H. Liebowitz, ed., Academic Press, New York, 1968, pp. 191–311.
[18] G. C. Sih and H. Liebowitz, *Mathematical theories of brittle fracture*, in Fracture: An Advanced Treatise, Vol. II, Mathematical Fundamentals, H. Liebowitz, ed., Academic Press, New York, 1968, pp. 67–190.

SIAM J. MATH. ANAL.
Vol. 36, No. 6, pp. 1929–1942

REGULARITY OF SOLUTIONS TO A CLASS OF CROSS DIFFUSION SYSTEMS*

DUNG LE†

Abstract. Regularity of bounded solutions to a class of strongly coupled parabolic systems is investigated. Conditions on the structure of the systems are found to assure that bounded solutions are Hölder continuous. The theory is then applied to the general Shigesada–Kawasaki–Teramoto model in population dynamics.

Key words. cross diffusion systems, Hölder regularity

AMS subject classifications. 35K57, 35B65

DOI. 10.1137/S0036141003428354

1. Introduction. In the last twenty years, cross diffusion systems have attracted great attention in both mathematical analysis and real life modeling. The introduction of cross diffusion terms into regular diffusion systems allows the mathematical models to capture much more important features of phenomena in physics, biology, ecology, and engineering sciences. At the same time, the presence of these terms caused enormous difficulties in the mathematical treatment due to the strong coupling in the diffusion terms. Among unanswered fundamental questions, we face the regularity of bounded weak solutions.

In this work, we study Hölder continuity of a bounded weak solution $\vec{u} = (u_1, \dots, u_m)$ to the following parabolic system:

$$\vec{u}_t = \nabla \cdot (\mathcal{A}(\vec{u})\nabla\vec{u}) + \mathcal{F}(\vec{u}). \tag{1.1}$$

Here, $\mathcal{A}(\vec{u}) = (P_{ij}(\vec{u}))$ is an $m \times m$ matrix, $\mathcal{F}(\vec{u}) = (F_1(\vec{u}), \dots, F_m(\vec{u}))$, and $\nabla\vec{u} = (\nabla_x u_1, \dots, \nabla_x u_m)$.

This system is coupled with certain boundary conditions on the boundary of a bounded domain Ω in $\mathbb{R}^n$. For the sake of simplicity, we will study only interior regularity of solutions although our method can be easily modified to cover the boundary case, given suitable boundary conditions.

Fundamental theory for strongly coupled systems like (1.1) was presented in [1]. The question on the global existence of solutions was also discussed there. For regular reaction diffusion systems, when $\mathcal{A}$ is a diagonal matrix, it is well known that bounded weak solutions are Hölder continuous (see [6]). Moreover, $\vec{u}$ exists globally if its supremum norm $\|\vec{u}(\bullet, t)\|_{\infty,\Omega}$ does not blow up in time. In contrast, things are more complicated if $\mathcal{A}$ is a full matrix. Counterexamples in [4] showed that, in general, a bounded solution to (1.1) can disappear in finite time while staying bounded. Thus, the boundedness of the L^∞ norm of a solution $\vec{u}$ is not sufficient to guarantee its global existence. Fortunately, there is an elaborated theory developed in [1] showing that a solution $\vec{u}$ to (1.1) exists globally (and is classical) if one has controls on both of its L^∞ and Hölder norms. Thus, it is necessary to study the Hölder continuity

*Received by the editors May 21, 2003; accepted for publication (in revised form) October 15, 2004; published electronically June 22, 2005.
http://www.siam.org/journals/sima/36-6/42835.html
†Department of Applied Mathematics, University of Texas at San Antonio, 6900 North Loop 1604 West, San Antonio, TX 78249 (dle@math.utsa.edu).

of the solutions. Furthermore, estimates on Hölder norms of solutions also provide valuable information on the compactness of the trajectories of solutions if one wishes to study their long time dynamics, such as the existence of global attractors.

Partial regularity results were obtained by Giaquinta and Struwe in [3] for a general class of systems. Everywhere regularity results for bounded solutions were proven only in very few situations assuming restrictive structure conditions. Among these are triangular systems (see [1, 9]) or strongly coupled systems of special form (see [9, 11]). In [9], we had to assume certain structural conditions that prevent the application of our results to many important models. In fact, the strongly coupled parabolic system

$$\begin{cases} \dfrac{\partial u}{\partial t} = \Delta[(d_1 + a_{11}u + a_{12}v)u] + F(u,v), \\ \dfrac{\partial v}{\partial t} = \Delta[(d_2 + a_{21}u + a_{22}v)v] + G(u,v), \end{cases} \tag{1.2}$$

does not satisfy the structures studied in [9, 11]. This system was proposed by Shigesada, Kawasaki and Teramoto in [10] to study spatial segregation of interacting species. Global existence and long time dynamics of solutions were investigated in either triangular cases or under the assumption that the dimension n of the domain Ω is two (see [5, 7] and the reference therein). For $n > 2$, to our best knowledge, the question of whether bounded positive solutions to this model are Hölder continuous (everywhere) remains open.

The aim of this paper is to present certain sufficient conditions on the structure of (1.1) for its bounded weak solutions to be Hölder continuous everywhere (and therefore classical). The dimension n of the domain Ω and the number m of equations in (1.1) can be arbitrary. This will be done in section 2. Our key assumption (see (H.1)) is the existence of a function $H(\vec{u})$, being defined on the range of a solution $\vec{u}$, which links the structures of the equations in a way that we can derive certain regularity information of $H(\vec{u}(x,t))$ in (x,t). To this end, we follow logarithmic function techniques developed in [8] dealing with scalar equations. Such regularity of H will be exploited later to study that of $\vec{u}$.

The condition (H.1) was motivated by a biological model of two equations studied in [10], where the following form of $P_{i,j}$ was considered:

$$P_{ij}(\vec{u}) = a\nabla u_i + c_i \nabla H(x,t,\vec{u}), \text{ with } i,j \in \{1,2\}. \tag{1.3}$$

The function H was a given affine function in $\vec{u}$ and represented the environmental influences on the species $\vec{u}$. Naturally, this influence might depend nonlinearly on $\vec{u}$ itself. For simplicity, let us assume that H depends only on $\vec{u}$, then (H.1) can be fulfilled (see Remark 2.2). Our condition (H.1) generalizes this situation in hope that it can cover more general structure and, in particular, (1.2).

Unless H is described explicitly as in (1.3), the existence of such function H required in (H.1) seems to be unclear and of little use in practice. In particular, we may ask if (H.1) is ever satisfied for (1.2). We take a closer look at these hypotheses in section 3, and try to find sufficient conditions on the parameters defining (1.1). This is not an easy task for systems of more than two equations. We are forced to study the case $m = 2$ (but n is still arbitrary). It turns out that H must be a solution to a first order PDE dictated by the coefficients of the system. This PDE can be solved by elementary methods of characteristics. Sufficient conditions for H to exist

will be formulated in Theorem 3. We will describe a way of constructing such H in Lemma 3.3 of section 3.

The system (1.2) is a test case of the general result in section 2. Since u, v are population densities, only positive solutions are of interest. We then study these solutions in section 4 where we will give the proof of the following theorem.

THEOREM 1. *Assume that $d_i, a_{ij} > 0$, $i, j = 1, 2$. If $(a_{22} - a_{12})(a_{11} - a_{21}) > 0$, then bounded positive weak solutions to* (1.2) *are Hölder continuous everywhere.*

That is, in population dynamics terms, we assumed that self diffusion rates are either stronger or weaker than cross diffusion ones. In fact, our method also works for a more general setting than that of (1.2) as we briefly point out in Theorem 4.

2. The general case. Let $\vec{u} = (u_1, \dots, u_m)$ be a bounded solution to (1.1) that exists on some interval $(0, T)$. We consider the range of $\vec{u}$

$$\Gamma = \{\vec{u}(x,t) \,:\, (x,t) \in \Omega_T = \Omega \times (0,T)\} \subset \mathbb{R}^m. \tag{2.1}$$

Hereafter, we use the summation convention, where the i, j indices run from 1 to m. We set $\mathcal{P}_i = P_{i,j}\nabla u_j = \sum_{j=1}^m P_{i,j}\nabla u_j$.

Our main assumptions are the followings.

(P.1) The functions $P_{i,j}, F_i$ are continuous functions in $\vec{u} \in \mathbb{R}^m$.

(H.1) There exist a C^2 function $H : \mathbb{R}^m \to \mathbb{R}$ defined on a bounded neighborhood Γ_0 of Γ, and positive numbers $\lambda_1, \lambda_2, \lambda_3$ such that

$$\Delta_1 = \nabla H \cdot (H_{u_j}\mathcal{P}_j) = \nabla H \cdot \sum_{j=1}^m H_{u_j}\mathcal{P}_j \geq \lambda_1 |\nabla H|^2, \tag{2.2}$$

$$\Delta_2 = \mathcal{P}_j \cdot \nabla H_{u_j} = \sum_{j=1}^m \mathcal{P}_j \cdot \nabla H_{u_j} \geq \lambda_2 |\nabla \vec{u}|^2, \tag{2.3}$$

$$\Delta_3 = |H_{u_j}\mathcal{P}_j| = |\sum_{j=1}^m H_{u_j}\mathcal{P}_j| \leq \lambda_3 |\nabla H| \tag{2.4}$$

for every C^1 function $\vec{u} : \Omega_T \to \mathbb{R}^m$ whose range is contained in Γ_0.

Here, with a slight abuse of notation, we will write $H_t = \frac{\partial}{\partial t} H(\vec{u}(x,t))$, $H_{u_j} = \frac{\partial}{\partial u_j} H(\vec{u})$, $H_{u_i u_j} = \frac{\partial^2}{\partial u_i \partial u_j} H(\vec{u})$, $\nabla H = \nabla_x H(\vec{u}(x,t))$, and so on.

Our main result of this section is the following theorem.

THEOREM 2. *Assume that Γ is bounded and* (P.1) *and* (H.1) *hold. We assert that $\vec{u}$ belongs to the Hölder class $C^{\alpha,\alpha/2}(\Omega_T)$ for any $\alpha \in (0,1)$. Moreover, the $C^{\alpha,\alpha/2}$ norm of $\vec{u}$ can be estimated in terms of the data of the equation and $\|\vec{u}\|_{\infty,\Omega_T}$.*

Let us fix a point $(x_0, t_0) \in \Omega_T$. For $R, r > 0$, we consider the cylinder $Q(R,r) := Q(x_0, t_0, R, r) := B_{x_0}(R) \times [t_0 - r, t_0]$, and always assume that R, r are sufficiently small such that $Q(R,r) \subset \Omega_T$. For $i = 1, 2, \dots$, we denote $Q_{iR} = Q(iR, iR^2)$.

The key ingredient of our proof is to show that

$$\liminf_{R\to 0} \frac{1}{R^n} \iint_{Q_R} |\nabla \vec{u}|^2 \, dxdt < \varepsilon, \quad Q_R := B(x_0, R) \times [t_0 - R^2, t_0] \quad \forall \varepsilon > 0. \tag{2.5}$$

Once this is proven, thanks to the Poincaré type inequality (see [3, Prop. 3.1])

$$\iint_{Q_R} |\vec{u} - \vec{u}_R|^2 \, dxdt \leq cR^2 \iint_{Q_R} |\nabla \vec{u}|^2 \, dxdt, \tag{2.6}$$

we see that (2.5) implies $\liminf_{R\to 0} \fint\!\!\!\fint_{Q_R} |\vec{u}-\vec{u}_R|^2\, dxdt < \varepsilon$. Since $\varepsilon > 0$ can be arbitrarily small, the Hölder continuity of $\vec{u}$ follows from [3, Theorem 3.1]. Moreover, provided ε is taken sufficiently small, the proof (see [3, pages 445–446]) also shows that $\fint\!\!\!\fint_{Q_R} |\vec{u}-\vec{u}_R|^2\, dxdt \le CR^\alpha$, for *any* $\alpha \in (0,1)$ and $R > 0$ with the constant C depends uniformly on α, ε the data and the supremum norm of $\vec{u}$. By the Campanato imbedding theorem (see [3]), the desired estimate for the $C^{\alpha,\alpha/2}$ norm of $\vec{u}$ is implied and we conclude the proof of Theorem 2.

For $R > 0$, we denote

$$M_i = \sup_{Q_{iR}} H(\vec{u}(x,t)), \quad m_i = \inf_{Q_{iR}} H(\vec{u}(x,t)), \text{ and } \omega_i = M_i - m_i,$$

and, for some positive θ, α to be determined later, define the following function:

$$w(x,t) := \log\left(\frac{\omega_4 + R^\alpha}{N(\vec{u}(x,t))}\right), \text{ with } N(\vec{u}) = \theta(M_4 - H(\vec{u})) + R^\alpha.$$

Let $Q^0 = \{(x,t) \in Q_{2R} : w(x,t)_+ = 0\}$. It is easy to see that Q^0 is the set where $H \le (1-1/\theta)M_4 + m_4$, or $H(\vec{u}(x,t))$ is away from M_4.

We consider the following two alternatives discussing the largeness of the measure $|Q^0|$ of this set Q^0.

(A) There is $R_0 > 0$ such that

$$|Q^0| > \frac{1}{\theta} R^{n+2} \quad \forall R \in (0, R_0). \tag{2.7}$$

(B) There is a sequence $\{R_k\}$, $R_k \to 0$, such that

$$|Q^0| < \frac{1}{\theta} R^{n+2}, \quad \text{for } R = R_k. \tag{2.8}$$

For any given $\varepsilon > 0$, we will try to determine $\theta = \theta(\varepsilon) > 0$ such that (2.5) holds. If (A) holds, the role of θ is not relevant in this case as any positive θ will do. Here, the values of $H(\vec{u}(x,t))$ in Q_{2R} are "evenly" between m_4, M_4, and we will follow the argument in [8] to show that $H(\vec{u}(x,t))$ is Hölder continuous by proving that w is uniformly bounded. On the other hand, if (B) holds for large θ, then H is close to M_4 in a large part of Q_{2R}. We will show that $\vec{u}$ will not oscillate to much by proving that its gradients is averagely small if θ is large.

First, we test the jth equation by $H_{u_j}\eta$, sum the results, and use (2.3) to get

$$\int_\Omega \frac{\partial H}{\partial t}\eta\, dx + \int_\Omega \left[(H_{u_j}\mathcal{P}_j)\nabla\eta + \lambda_2|\nabla\vec{u}|^2\eta\right] dx \le \int_\Omega \Psi\eta\, dx. \tag{2.9}$$

Here we denoted $\Psi = \sum F_j H_{u_j}$.

Proof of (2.5) *given (A).*

Let η be a function with compact support in Q_{2R}. We test the jth equation by $\theta H_{u_j}\eta/N$, and add the results to get

$$\begin{aligned} &\int_\Omega \frac{\partial w}{\partial t}\eta\, dx \qquad + \int_\Omega \theta\left[\frac{(H_{u_j}\mathcal{P}_j)}{N}\nabla\eta + \frac{\mathcal{P}_j\nabla H_{u_j}}{N}\eta\right] dx \\ &\qquad + \int_\Omega \theta^2 \frac{(H_{u_j}\mathcal{P}_j)\nabla H}{N^2}\eta\, dx = \int_\Omega \frac{\theta\Psi}{N}\eta\, dx. \end{aligned} \tag{2.10}$$

If $\eta \geq 0$, then (2.2), (2.3) and the above imply

$$\int_\Omega \frac{\partial w}{\partial t}\eta \, dx + \theta \int_\Omega \frac{(H_{u_j}\mathcal{P}_j)}{N}\nabla\eta \, dx \leq \int_\Omega \frac{\theta\Psi}{N}\eta \, dx. \tag{2.11}$$

We first show that $\|w\|_{\infty,Q_R}$ can be estimated in terms of $\|w\|_{2,Q_{2R}}$. We replace η in (2.11) by $(w-k)_+\eta^2$, with η being a cut-off function in Q_{2R}. Note that $\nabla w = \theta\nabla H/N$. We derive

$$\begin{aligned}\int_\Omega \frac{\partial (w-k)_+^2}{\partial t}\eta^2 \, dx \quad &+ \quad \int_{w\geq k} \frac{\theta^2(H_{u_j}\mathcal{P}_j)\nabla H}{N^2}\eta^2 dx \\ &\leq \int_\Omega \frac{\theta(H_{u_j}\mathcal{P}_j)(w-k)_+}{N}\eta|\nabla\eta| \, dx + \int_\Omega \frac{\theta\Psi(w-k)_+}{N}\eta^2 \, dx.\end{aligned}$$

Integrating with respect to $t \in I := [t_0 - 4R^2, t_0]$ and using the Young inequality and (2.4) to the first term on the right-hand side, we obtain

$$\begin{aligned}&\sup_{t\in[t_0-R^2,t_0]}\int_{B(R)} \frac{\partial (w-k)_+^2}{\partial t}\eta^2 \, dx + \iint_{Q_{2R}} |\nabla(w-k)_+|^2\eta^2 \, dxdt \\ &\leq \theta^2 \iint_{Q_{2R}} (w-k)_+^2 R^{-2} \, dxdt + \theta^2 \int_I \int_{w\geq k} \frac{R^2\Psi^2}{N^2}\eta^2 dxdt.\end{aligned} \tag{2.12}$$

Let $A_k = \{(x,t) \in Q_{2R} : w(x,t) > k\}$. Because Ψ is bounded on Γ_0 and $N \geq R^\alpha \geq C|A_k|^{\frac{\alpha}{n+2}}$, we bound the last term in (2.12) by

$$\int_I \int_{w\geq k} \frac{R^2\Psi^2}{N^2}\eta^2 dxdt \leq C|A_k|^{1-\frac{2\alpha}{n+2}}. \tag{2.13}$$

We now choose $\alpha < 1$ and see that $1 - \frac{2\alpha}{n+2} > \frac{n}{n+2}$. This fact, (2.12), (2.13), and the well known DiGiorgi's method in [6] will give

$$\sup_{B_{x_0}(R)\times[t_0-R^2,t_0]} w \leq \text{Const}\left(1 + \frac{1}{R^{n+2}}\iint_{B_{x_0}(2R)\times[t_0-2R^2,t_0]} (w^+)^2 \, dxdt\right). \tag{2.14}$$

Next, we replace η by η^2 in (2.10) and use (2.2), (2.3), and (2.4) to get

$$\int_\Omega \frac{\partial w}{\partial t}\eta^2 \, dx + \int_\Omega |\nabla w|^2\eta^2 \, dx \leq C\int_\Omega \left(|\nabla w|\eta|\nabla\eta| + \frac{\theta\Psi}{N}\eta^2\right) \, dx.$$

This and the condition on the measure of Q^0 in (2.7) allow us to follow the proof of [8, Lemma 2.4] to show that $\|w\|_{2,Q_{2R}}$, and therefore $\|w\|_{\infty,Q_R}$ (see (2.14)), can be bounded by a constant independent of R. The argument in [8, page 924] then gives a decay estimate for the oscillation of H and then its Hölder continuity.

Moreover, let η be a cutoff function for Q_R and satisfy: $\eta \equiv 1$ in Q_R, $\eta(x,t) \equiv 0$ outside the cylinder Q' given by $B_{x_0}(2R)\times[t_0-2R^2, t_0+R^2]$. In addition, $|D\eta| \leq 1/R$ and $\left|\frac{\partial\eta}{\partial t}\right| \leq 1/R^2$. Replacing η in (2.9) by $(H(\vec{u}) - \inf_{Q_R} H(\vec{u}))\eta^2$, with η is the above cutoff function, and using the fact that for some positive ν, $|H(\vec{u}) - \inf_{Q_R} H(\vec{u})| \leq CR^\nu$ we easily show that $|DH(\vec{u})|^2 \in L^{1,\mu}_{loc}$ for $\mu = n + 2\nu > n$.

Let $\bar{H}(\vec{u}) = H(\vec{u}) - H(\vec{u})_R$, where $H(\vec{u})_R$ denotes the mean value of $H(\vec{u}(x,t))$ over Q_R. We can replace H in (2.9) by $\bar{H}$. We use (2.4) to get

$$\int_\Omega \frac{\partial(\bar{H}\eta)}{\partial t}\, dx + \lambda_2 \int_\Omega |\nabla \vec{u}|^2 \eta\, dx \le C \int_\Omega \left[|\nabla H||\nabla \eta| + |\bar{H}| \frac{\partial \eta}{\partial t} + \Psi \eta \right] dx.$$

We integrate the above inequality with respect to t over the interval $[t_0 - 2R^2, t_0 + R^2]$ and obtain

$$\lambda_2 \iint_{Q'} |\nabla \vec{u}|^2 \eta\, dxdt \le C \iint_{Q'} \left(|\nabla H||\nabla \eta| + |\bar{H}||\frac{\partial \eta}{\partial t}| + |\Psi \eta| \right) dxdt.$$

Using the definition of η and the facts that $|DH(\vec{u})|^2 \in L^{1,\mu}_{loc}$, and that $|\bar{H}| \le CR^\nu$ (since $H(\vec{u}(x,t))$ is Hölder continuous), we can majorize the terms on the right as follows:

$$\iint_{Q'} |\bar{H}| \left| \frac{\partial \eta}{\partial t} \right| dxdt \le CR^{n+\nu},$$

$$\iint_{Q'} |\nabla H||\nabla \eta|\, dxdt \le \frac{|Q'|^{1/2}}{R} \left(\iint_{Q'} |\nabla H|^2\, dxdt \right)^{1/2} \le CR^{(n+\mu)/2}.$$

Since $\mu > n$, we have $(n+\mu)/2 > n$. From these estimates, for some $\gamma > 0$, we obtain

$$\iint_{Q_R} |\nabla \vec{u}|^2\, dxdt \le \iint_{Q'} |\nabla \vec{u}|^2 \eta\, dxdt \le CR^{n+\gamma}.$$

This gives (2.5) and completes this alternative.

Proof of (2.5) *given (B).* Since we can replace H by $H + k$ with k being any constant, we will hereafter assume that $m_4 = 0$ in the definition of w. It is clear that (B) implies

$$|\{(x,t) \in Q_R | H \le (1-\rho) M_4\}| < \rho R^{n+2}, \quad \text{with } \rho = 1/\theta. \tag{2.15}$$

We have the following estimate for the integral of $|\nabla \vec{u}|$.

LEMMA 2.1. *Let $q = 2n/(n+2)$. There exists a constant $C(M_4)$ such that*

$$\left(\fint\!\!\!\fint_{Q_R} |\nabla \vec{u}|^q\, dxdt \right)^{\frac{2}{q}} \le C(M_4) \rho^{\frac{2}{n}} R^{-2} \quad \forall R < \rho M_4. \tag{2.16}$$

Proof. We substitute $(H-k)^+ \eta^2$, with $k \in \mathbb{R}$ and η being a cut-off function on Q_{2R} with respect to Q_R, into the places of η in (2.9). Since we can choose η such that $|D\eta|^2 + |\eta_t| \le cR^{-2}$, standard estimates give

$$\lambda_2 \int \int_{Q_R} |\nabla \vec{u}|^2 (H-k)_+ dxdt + \frac{\lambda_1}{2} \int \int_{Q_R} |\nabla H|^2 \le \int \int_{Q_{2R}} (H-k)_+^2 |\nabla \eta|^2 + CR^{n+2}.$$

By the choice of $k := (1-2\rho)M_4$, $(H-k)^+ \le 2\rho M_4$ on Q_{2R} so that

$$\iint_{Q_R} |\nabla \vec{u}|^2 (H-k)^+\, dxdt \le C(\rho M_4)^2 R^n + CR^{n+2}.$$

Let $A_0 := \{(x,t) \in Q_R | H \geq (1-\rho)M_4\}$. Then $(H-k)^+ \geq \rho M_4$ on A_0. So,

$$\iint_{A_0} |\nabla \vec{u}|^2 \, dxdt \leq CR^n \left(\rho M_4 + \frac{R^2}{\rho M_4}\right) \leq 2\rho M_4 R^n \quad \forall R < \rho M_4. \tag{2.17}$$

Also, by substituting $e^{sH}\eta^2$ into places of η in (2.9) with $s > 0$ sufficiently large, it is standard to show

$$\iint_{Q_R} |\nabla \vec{u}|^2 \, dxdt \leq CR^n. \tag{2.18}$$

For any subset A of Q_R, Hölder's inequality gives

$$\iint_A |\nabla \vec{u}|^q \, dxdt \leq \left(\iint_A |\nabla \vec{u}|^2 \, dxdt\right)^{\frac{n}{n+2}} |A|^{\frac{2}{n+2}}. \tag{2.19}$$

Taking $A = A_0$ and using (2.17), we obtain

$$\iint_{A_0} |\nabla \vec{u}|^q \, dxdt \leq (2\rho M_4 R^n)^{\frac{n}{n+2}} R^2 = (2\rho M_4)^{\frac{n}{n+2}} R^{n+\frac{4}{n+2}} \quad \forall R < \rho M_4.$$

Similarly, we take $A = Q_R \setminus A_0$ in (2.19). Using (2.18) and also the fact that $|A| \leq \rho R^{n+2}$ by (2.15), we have

$$\iint_{Q_R \setminus A_0} |\nabla \vec{u}|^q \, dxdt \leq (CR^n)^{\frac{n}{n+2}} (\rho R^{n+2})^{\frac{2}{n+2}} R^2 = C\rho^{\frac{2}{n+2}} R^{n+\frac{4}{n+2}}.$$

Since $(n + \frac{4}{n+2} - n - 2)\frac{2}{q} = -2$. The above estimates give the lemma. □

From [3, page 443], we have

$$⨍\!\!⨍_{Q_R} |\nabla \vec{u}|^2 \, dxdt \leq C(\varepsilon) \left(⨍\!\!⨍_{Q_{4R}} |\nabla \vec{u}|^q \, dxdt\right)^{\frac{2}{q}} + \varepsilon ⨍\!\!⨍_{Q_{4R}} |\nabla \vec{u}|^2 \, dxdt \quad \forall \varepsilon > 0. \tag{2.20}$$

Using (2.18), (2.16) in (2.20) to estimate the integrals on the right-hand side and multiplying through by R^2, we obtain

$$\frac{1}{R^n} \iint_{Q_{R/4}} |\nabla \vec{u}|^2 \, dxdt \leq C(\varepsilon) C(M_4) \rho^{\frac{2}{n}} + C\varepsilon \quad \forall R < \rho M_4.$$

Obviously, we can make the right-hand side arbitrarily small by choosing ε and then $\rho = \rho(\varepsilon)$ sufficiently small. We have shown (2.5), given (B) with θ being sufficiently large. Our proof is now complete.

Remark 2.2. If (1.3) is considered, then one can easily check that (H.1) is verified if a, c_i are positive constants, and the matrix $(H_{u_i u_j})$ is positive definite (that is, $H_{u_i u_j}\xi_i\xi_j \geq \gamma|\xi|^2$ for some $\gamma > 0$ and $\xi \in \mathbb{R}^{m\times n}$). The conditions (2.2)–(2.4) seem to be technical. They are discovered in the search for an equation satisfied by the influence H. It turns out that these conditions make H a subsolution to some scalar parabolic equation. This information was used in the proof of the alternative (A).

3. Sufficient conditions for the existence of H. In this section we will find sufficient conditions that guarantee the existence of the function H satisfying our main assumptions (2.2)–(2.4). Unless H is explicitly present in systems like (1.3), the question of the existence of H verifying (H.1) is not easy to answer for general $m \geq 2$. Here, we restrict ourself to the case of two equations ($m = 2$). Parts of our calculation can be useful to study the case $m > 2$, but the possibility of formulating a similar sufficient condition, like (H.2) below, is questionable.

From now on, as we will deal with two equations and to simplify our notation, we denote $\vec{u} = (u, v)$, $P_1 = P_{11}$, $P_2 = P_{12}$, $Q_1 = P_{21}$, and $Q_2 = P_{22}$.

We then assume further that

(P.2) P_1, P_2, Q_1, Q_2 are C^1 functions on Γ_0 and are positive on Γ. Moreover, P_1, Q_2 are greater than a positive constant on $\overline{\Gamma}$.

We consider the following equation:

$$-P_2 f^2 + (P_1 - Q_2) f + Q_1 = 0, \tag{3.1}$$

because $P_2 Q_1 > 0$, (3.1) has two solutions $f = f_1, f_2$ with $f_1 f_2 < 0$. In what follows, we denote by $f(u, v)$ the positive solution of (3.1). By (P.2), f is differentiable so that there exists a C^2 solution g (see [2, pages 106–109]) to the following first order equation:

$$g_u - f(u, v) g_v = 0. \tag{3.2}$$

We will impose the following main assumption of this section.

(H.2) Let Γ_0 be connected. Assume that there exists a solution g to (3.2) such that g is defined on Γ_0, and

$$g_v \neq 0, \text{ and } 4(Q_1 P_2 - P_1 Q_2)(f_u - f_v f) \neq 0 \quad \forall (u, v) \in \Gamma_0. \tag{3.3}$$

The existence of H is then given by the following theorem.

THEOREM 3. *Assume* (H.2) *and let* $H(u, v) = K \exp(k g(u, v))$. *There exist* $K > 0$ *and* $k \in \mathbb{R}$ *such that* (H.1) *holds.*

Note also that if $G : \mathbb{R} \to \mathbb{R}$ is a C^2 function, then $H(u, v) = G(g(u, v))$ is also a C^2 solution to (3.2). To prove this theorem, we need only to verify the conditions (2.2)–(2.4) that amount to the positivity of the following quadratics for every $(u, v) \in \overline{\Gamma}$ and $U, V \in \mathbb{R}^n$.

$$\begin{aligned} A_1 \;=\; & \left((P_1 H_u + Q_1 H_v) H_u - \lambda_1 {H_u}^2\right) U^2 + \left((P_2 H_u + Q_2 H_v) H_v - \lambda_1 {H_v}^2\right) V^2 \\ & + \left((P_2 H_u + Q_2 H_v) H_u + (P_1 H_u + Q_1 H_v) H_v - 2\lambda_1 H_v H_u\right) VU, \end{aligned} \tag{3.4}$$

$$\begin{aligned} A_2 \;=\; & (Q_1 H_{uv} + P_1 H_{uu} - \lambda_2) U^2 + (P_2 H_{uv} + Q_2 H_{vv} - \lambda_2) V^2 \\ & + (P_2 H_{uu} + P_1 H_{uv} + Q_2 H_{uv} + Q_1 H_{vv}) VU, \end{aligned} \tag{3.5}$$

and

$$\begin{aligned} A_3 \;=\; & \left(\lambda_3 {H_u}^2 - (P_1 H_u + Q_1 H_v)^2\right) U^2 + \left(\lambda_3 {H_v}^2 - (P_2 H_u + Q_2 H_v)^2\right) V^2 \\ & + \left(2\lambda_3 H_v H_u - 2 (P_2 H_u + Q_2 H_v) (P_1 H_u + Q_1 H_v)\right) VU. \end{aligned} \tag{3.6}$$

To study these quadratics, we will need some lemmas and calculations.

LEMMA 3.1. *If* H *is a solution to* (3.2) *and* (P.2) *holds, then there exist* $\lambda_1, \lambda_3 > 0$ *such that* A_1, A_3 *are positive definite.*

Proof. First, we will prove that the discriminants Θ_1, Θ_3 of A_1, A_3 are nonpositive. However, simple calculation shows that

$$\begin{aligned}\Theta_1 &= \left(\left(P_2H_u + Q_2H_v\right)H_u + \left(P_1H_u + Q_1H_v\right)H_v - 2\lambda_1 H_v H_u\right)^2 \\ &\quad -4\left(\left(P_1H_u + Q_1H_v\right)H_u - \lambda_1 {H_u}^2\right)\left(\left(P_2H_u + Q_2H_v\right)H_v - \lambda_1 {H_v}^2\right) \\ &= \left(H_v P_1 H_u - P_2 {H_u}^2 - H_u Q_2 H_v + Q_1 {H_v}^2\right)^2,\end{aligned}$$

and

$$\Theta_3 = 4\lambda_3\left(P_1 H_u H_v - {H_u}^2 P_2 - H_v H_u Q_2 + Q_1 {H_v}^2\right)^2.$$

Because H satisfies (3.2), it is easy to see that $\Theta_1 = \Theta_3 = 0$. We need only to show that the coefficients of U^2, V^2 in A_1, A_3 are positive. By (3.1), they can be written as

$$H_v^2(P_1 f^2 + Q_1 f - \lambda_1 f^2) = H_v^2 f^2(P_2 f + Q_2 - \lambda_1), \quad H_v^2(P_2 f + Q_2 - \lambda_1),$$

$$H_u^2(\lambda_3 - (P_1 + Q_1/f)^2) = H_u^2(\lambda_3 - (P_2 f + Q_2)^2), \quad H_v^2(\lambda_3 - (P_2 f + Q_2)^2).$$

Since $f > 0$ and because of (P.2), we can take

$$\lambda_1 = \frac{1}{2}\inf_{\Gamma}(P_2 f + Q_2), \quad \lambda_3 = 2\sup_{\Gamma}(P_2 f + Q_2)^2,$$

which are positive thanks to (P.2). The lemma is proven. □

To verify the positivity of A_2 in (2.3), we consider its discriminant Θ_2. Easy computation shows that we can write

$$\Theta_2 = -4\lambda_2^2 + \Theta_{11}\lambda_2 + \Theta_{12},$$

with

$$\Theta_{11} := 4\left(Q_1 H_{uv} + P_1 H_{uu} + P_2 H_{uv} + Q_2 H_{vv}\right),$$

$$\Theta_{12} := \left(P_2 H_{uu} + P_1 H_{uv} + Q_2 H_{uv} + Q_1 H_{vv}\right)^2 - 4\left(Q_1 H_{uv} + P_1 H_{uu}\right)\left(P_2 H_{uv} + Q_2 H_{vv}\right).$$

Differentiating $H_u = f H_v$, we get $H_{uu} = f_u H_v + f H_{uv}$ and $H_{uv} = f_v H_v + f H_{vv}$. Substitute these into Θ_{12} and simplify to obtain

$$\Theta_{12} := \alpha_1 H_{vv}^2 + \alpha_2 H_{vv} H_v + \alpha_3 H_v^2,$$

with

$$\begin{aligned}\alpha_1 &= \left(P_2 f^2 + P_1 f + Q_2 f + Q_1\right)^2 - 4\left(Q_1 f + P_1 f^2\right)\left(P_2 f + Q_2\right), \\ \alpha_2 &= 2\left(P_2\left(f_u + f f_v\right) + P_1 f_v + Q_2 f_v\right)\left(P_2 f^2 + P_1 f + Q_2 f + Q_1\right) \\ &\quad -4\left(Q_1 f_v + P_1\left(f_u + f f_v\right)\right)\left(P_2 f + Q_2\right) - 4\left(Q_1 f + P_1 f^2\right) P_2 f_v, \\ \alpha_3 &= \left(P_2(f_u + f f_v) + P_1 f_v + Q_2 f_v\right)^2 - 4(Q_1 f_v + P_1(f_u + f f_v)) P_2 f_v.\end{aligned}$$

From (3.1), we have $P_2 f^2 + Q_2 f = P_1 f + Q_1$. Hence,

$$\alpha_1 = 4(P_1 f + Q_1)^2 - 4(Q_1 + P_1 f)(P_2 f^2 + Q_2 f) = 0,$$

$$\begin{aligned}\alpha_2 &= 4[(P_2(f_u+ff_v)+P_1f_v+Q_2f_v)(P_1f+Q_1)\\ &\quad -(Q_1f_v+P_1(f_u+ff_v))(P_2f+Q_2)-(Q_1f+P_1f^2)P_2f_v]\\ &= 4[P_2Q_1(f_u-ff_v)-P_1f_v(P_2f^2-P_1f-Q_1)-P_1f_uQ_2]\\ &= 4(Q_1P_2-P_1Q_2)(f_u-f_vf),\end{aligned}$$

$$\begin{aligned}\alpha_3 &= (f_u+ff_v)^2{P_2}^2+(P_1+Q_2)^2f_v^2+2P_2f_v[(f_u+ff_v)(Q_2-P_1)-2Q_1]\\ &= (f_u+ff_v)^2{P_2}^2+(P_1-Q_2)^2f_v^2+2P_2f_v(f_u+ff_v)(Q_2-P_1)\\ &\quad +4(P_1Q_2-Q_1P_2)f_v^2\\ &= [(f_u+ff_v)P_2+f_v(Q_2-P_1)]^2+4(P_1Q_2-Q_1P_2)f_v^2.\end{aligned}$$

Similarly, we also have $\Theta_{11}=\beta_2H_{vv}+\beta_3H_v$, with

$$\beta_2=4(Q_1f+P_1f^2+P_2f+Q_2)=4(f^2+1)(P_2f+Q_2),$$

$$\beta_3=4(Q_1f_v+P_1(f_u+ff_v)+P_2f_v).$$

Proof of Theorem 3. Thanks to Lemma 3.1 and the choice of g, we need only to check the positivity of A_2. We first show that $\Theta_2<0$ on Γ for suitable choice of k. Let $G(x)=\exp(kx)$ and $\widetilde{H}(u,v)=G(g(u,v))$. As $\widetilde{H}_v=G'g_v$, $\widetilde{H}_{vv}=(G''g_v^2+G'g_{vv})$, and $G''/G'=k$, we have

$$\widetilde{\Theta}_{12}=\widetilde{H}_{vv}\widetilde{H}_v\alpha_2+\widetilde{H}_v^2\alpha_3=(G')^2\left[kg_v^3\alpha_2+(g_{vv}g_v\alpha_2+g_v^2\alpha_3)\right].$$

Thanks to our assumption (3.3) and the fact that Γ_0 is connected, the coefficient of k never vanishes on Γ_0. That is, either $\sup_{\overline{\Gamma}}g_v^3\alpha_2<0$ or $\inf_{\overline{\Gamma}}g_v^3\alpha_2>0$. Moreover, $(g_{vv}g_v\alpha_2+g_v^2\alpha_3)$ is bounded on $\overline{\Gamma}$ and $g_v,G'\neq 0$. The above shows that $\widetilde{\Theta}_{12}<0$ on $\overline{\Gamma}$ for suitable choice of k with $|k|$ sufficiently large.

We then choose K sufficiently large in

$$\Theta_2=-4\lambda_2^2+\Theta_{11}\lambda_2+\Theta_{12}=-4\lambda_2^2+K(\beta_2\widetilde{H}_{vv}+\beta_3\widetilde{H}_v)\lambda_2+K^2\widetilde{\Theta}_{12}$$

to see that $\Theta_2<0$, because $\widetilde{\Theta}_{12}<0$ on $\bar{\Gamma}$.

Finally, we need only to check that the coefficients of U^2,V^2 in A_2 are positive. This amounts to show that the quantities

$$\delta_1=(Q_1H_{uv}+P_1H_{uu})\ \text{ and }\ \delta_2=(P_2H_{uv}+Q_2H_{vv})$$

are strictly positive on $\overline{\Gamma}$, and then choose λ_2 sufficiently small. Similar calculation as before yields

$$\delta_1=\left(Q_1f+P_1f^2\right)H_{vv}+Q_1f_vH_v+P_1\left(f_uH_v+ff_vH_v\right)=K\exp(kg)\left[k\delta_{12}+f\delta_{11}k^2\right],$$

$$\delta_2=(P_2f+Q_2)H_{vv}+P_2f_vH_v=K\exp(kg)\left[k\delta_{21}+\frac{\delta_{22}}{f}k^2\right],$$

where $\delta_{12}=\left(Q_1f+P_1f^2\right)g_{vv}+Q_1f_vg_v+P_1\left(f_ug_v+ff_vg_v\right)$, $\delta_{21}=(P_2f+Q_2)g_{vv}+P_2f_vg_v$ and $\delta_{11}=\delta_{22}={g_v}^2(Q_1+P_1f)={g_v}^2(P_2f^2+Q_2f)$.

Since δ_{12},δ_{21} are bounded on $\overline{\Gamma}$, and $\delta_{11}=\delta_{22}>0$, we can choose $|k|$ large to have that $\delta_1,\delta_2>0$ on $\overline{\Gamma}$. □

Remark 3.2. Differentiate (3.1) with respect to u, v, and then solve for f_u, f_v to see that

$$f_u - f_v f = \frac{f\mathcal{F}}{P_2 f^2 + Q_1}, \tag{3.7}$$

with $\mathcal{F} = \partial_v P_v f^3 + (\partial_v Q_2 - \partial_u P_2 - \partial_v P_1) f^2 + (\partial_u P_1 - \partial_u Q_2 - \partial_v Q_1) f + \partial_u Q_1$. Therefore, (3.3) of (H.2) requires that $\mathcal{F} \neq 0$ on $\overline{\Gamma}$.

We conclude this section by discussing the existence of a solution g of (3.2) and the condition $g_v \neq 0$ in (3.3). In applications, see section 4, the function $f(u, v)$ may be singular in u, v so that we will rewrite (3.2) in the form

$$M(u) g_u - N(u, v) g_v = 0 \tag{3.8}$$

for some function M, N such that $f = N/M$, $M \neq 0$. The above equation can be solved by characteristic methods. If g is such a solution, then we know that (see [2, pp 97–99]) $\vec{x}(t) = (u(t), v(t))$, $z(t) = g(\vec{x}(t))$ and $\vec{p}(t) = (p_u(t), p_v(t)) = \nabla g(\vec{x}(t))$ solve the following system:

$$\begin{aligned} \vec{x}'(t) &= (M, -N), \\ \vec{p}'(t) &= (-M_u p_u + N_u p_v, N_v p_v), \\ z'(t) &= M p_u - N p_v. \end{aligned} \tag{3.9}$$

We first observe that the above system is decoupled. In fact, one can solve for $u(t)$ from the first equation $u'(t) = M(u(t))$, and then substitute the result into the second equation $v'(t) = -N(u(t), v(t))$ to get $v(t)$. Once $u(t), v(t)$ are known, we can solve the (linear) equation for $\vec{p}$. Equation (3.8) simply says that g is constant along the characteristic curve $\vec{x}(t)$, which is an integral curve of the planar vector field $(M, -N)$.

Let γ be a smooth curve in $\Gamma_0 \subset \mathbb{R}^2$, which is a diffeomorphic image of an open interval, and assume that γ is noncharacteristic with respect to (3.9). For each $x_0 \in \gamma$, we assume that the first equation of (3.9) has a solution $\vec{x}(t)$ with $\vec{x}(0) = x_0$. This solution is locally defined on certain open interval I_{x_0} containing 0. We then define

$$\Xi = \{(x_0, s) \in \gamma \times I_{x_0} \,:\, \vec{x}(t) \text{ exists on } I_{x_0} \text{ with } \vec{x}(0) = x_0\},$$

and the map $X : \Xi \to \mathbb{R}^2$ by $X(\vec{x}(0), t) = \vec{x}(t)$. The system (3.9) is also supplied with its initial data

$$z(0) = \widetilde{g}(\vec{x}(0)) \quad \text{and} \quad \vec{p}(0) = \vec{p}_0(\vec{x}(0)) \tag{3.10}$$

along γ. Here, the function $\widetilde{g}$ is given in a neighborhood of γ and $\vec{p}_0(\vec{x}_0) = (p_u^0, p_v^0)$ is a vector field defined along γ. The initial data must satisfy the compatibility conditions on γ,

$$\vec{p}_0 = (p_u^0, p_v^0) = \nabla \widetilde{g}, \quad M p_u^0 = N p_v^0. \tag{3.11}$$

Let $\vec{T} = (T_u, T_v)$ be the unit tangent vector along γ. Since γ is a diffeomorphic image of an open interval, we can pick $\widetilde{g}$ such that $\nabla \widetilde{g} \cdot \vec{T} \neq 0$ (see also the proof of Theorem 1 in section 4).

The following lemma concerns the first condition in (3.3) of (H.2).

LEMMA 3.3. *Assume that* $\bar{\Gamma} \subset X(\Xi)$ *and* (3.11). *Then there exists a solution to* (3.2) *with* $g_v \neq 0$ *on* $\bar{\Gamma}$.

Proof. We define g on $\bar{\Gamma}$ by $g(\vec{x}(t)) = \widetilde{g}(\vec{x}(0))$. From the condition (3.11), we have $\vec{p_0} \cdot \vec{T} = \nabla \widetilde{g} \cdot \vec{T}$. Therefore, $p_v^0(NT_u + MT_v) = M \nabla \widetilde{g} \cdot \vec{T}$. Since γ is noncharacteristic, we have $(NT_u + MT_v) \neq 0$. Because $\nabla \widetilde{g} \cdot \vec{T} \neq 0$, $p_v^0 \neq 0$ along γ. From the second equation of (3.9) we have $p_v' = N_v p_v$. This gives

$$p_v(t) = p_v^0(\vec{x}(0)) \exp\left(\int_0^t N_v(\vec{x}(s)) ds \right).$$

Because $g_v(\vec{x}(t)) = p_v(t)$ and $p_v^0(\vec{x}(0)) \neq 0$, we see that $g_v(\vec{x}(t)) \neq 0$ on $I_{\vec{x}(0)}$. Thus, $g_v \neq 0$ on $\bar{\Gamma}$. ∎

In a more special situation, we can explicitly find g.

LEMMA 3.4. *If* $f(u,v) = \hat{f}(v/u)$ *for some function* $\hat{f}$, *then there exists a solution to* (3.2) *defined on the set* $u, v > 0$ *with* $g_v > 0$.

Proof. Set $v = Yu$, we derive

$$\frac{dv}{du} = u\frac{dY}{du} + Y = -\hat{f}(Y) \Rightarrow \frac{dY}{\hat{f}(Y) + Y} = -\frac{du}{u},$$

which has the general solution $g(u,v) = C$ with

$$g(u,v) = \int \frac{dY}{\hat{f}(Y) + Y} + \log(u), \quad Y = \frac{v}{u}.$$

Hence, $g_v = 1/[\hat{f}(\frac{v}{u}) + \frac{v}{u})u] = 1/(f(u,v)u + v)$, which is positive if u, v are positive. ∎

4. Applications. We conclude our paper by giving the proof of Theorem 1. We consider the system (1.2) and let

$$P = d_1 u + a_{12} uv + a_{11} u^2, \quad Q = d_2 v + a_{21} uv + a_{22} v^2.$$

Obviously, (1.2) is a special case of (1.1) with P_1, P_2, Q_1, Q_2 simply the partial derivatives of P, Q with respect to u, v. That is,

$$P_1 = d_1 + a_{12} v + 2a_{11} u, \quad P_2 = a_{12} u, \quad Q_1 = a_{21} v, \quad Q_2 = d_2 + a_{21} u + 2a_{22} v. \tag{4.1}$$

The proof is to verify the assumption (H.2) and that of Lemma 3.3 in the previous section. As we consider bounded positive solutions, the set Γ in (2.1) is a bounded subset of $\mathbb{R}^2_+$. One can see that (P.2) is satisfied.

Proof of Theorem 1. We first look at the condition $\alpha_2 \neq 0$ in (H.2). From Remark 3.2 and (4.1), we have $\mathcal{F} = 2(a_{22} - a_{12}) f^2 + 2(a_{11} - a_{21}) f \neq 0$. This is the case because $f > 0$ and $(a_{22} - a_{12})(a_{11} - a_{21}) > 0$. Calculation also shows that $Q_1 P_2 - P_1 Q_2$ is

$$-(2a_{11}u^2 a_{21} + 4a_{11} u a_{22} v + (d_1 a_{21} + 2a_{11} d_2)\, u + 2a_{12} v^2 a_{22} + (2d_1 a_{22} + a_{12} d_2)\, v + d_1 d_2),$$

which is negative on $\mathbb{R}^2_+$. Therefore, the second condition in (3.3) of (H.2). is verified.

Finally, we show the existence of g solving (3.8) with $g_v \neq 0$. Since Γ is bounded, there exists $L > 0$ such that $\Gamma \subset \Gamma_0 = (0, L) \times (0, L)$. On the other hand, we see that the positive solution to (3.1) is given by

$$f(u,v) = \frac{(P_1 - Q_2) + \sqrt{(P_1 - Q_2)^2 + 4P_2 Q_1}}{2P_2}, \quad u, v > 0.$$

Because $P_2 = a_{12}u$, $f(u,v)$ is singular in u. Therefore, we will take in the equation (3.8),

$$M(u) = u, \quad N(u,v) = ((P_1 - Q_2) + \sqrt{(P_1 - Q_2)^2 + 4P_2Q_1})/(2a_{12}). \tag{4.2}$$

Let $(u_0, v_0) \in \Gamma_0$ be given. The characteristic curves of (3.8), with $u(0) = u_0$ and $v(0) = v_0$, are given by $u(t) = u_0 \exp(t)$ and $v(t)$ being the solution to $\frac{d}{dt}v(t) = -N(u(t), v(t))$. Let $I_0 = (a_0, b_0)$ be the interval on which $u(t), v(t)$ are positive.

From (4.2), the equation for v is sublinear. That is, there are functions $\alpha(t), \beta(t)$ such that $|N(u(t), v(t))| \le \alpha(t)v(t) + \beta(t)$. Using the Gronwall inequality, it is easy to show that $v(t)$ exists for all t in some interval I as long as $N(u,v)$ is defined. In particular, this is the case when $u(t), v(t)$ stay positive. Since $N(u,v) > 0$, $v(t)$ is decreasing so that $v(t) > v_0 > 0$ for all $t < 0$. Of course, $u(t) > 0$, for all t and $\lim_{t\to-\infty} u(t) = 0$. Hence, we can take $I_0 = (-\infty, b_0)$.

We argue that $\lim_{t\to b_0^-} v(t) = 0$. If not, let $v_1 > 0$ be the limit. Continuation arguments allow us to extend u, v, being positive, beyond b_0 if b_0 is finite. Thus $b_0 = \infty$. From the definitions of P, Q, N, we easily see that $N(u(t), v(t)) \ge c_0$ for some $c_0 > 0$ and t sufficiently large. But this implies $v(t) \le -c_0 e^t + v_0$ for some constant v_0, and v can be negative on I_0 as $t \to \infty$, a contradiction. This shows that $\lim_{t\to b_0^-} v(t) = 0$.

The function $u(t) - v(t)$ is increasing on I_0. It is negative when $t \to -\infty$, and positive when $t \to b_0^-$. Thus, there is t_0 such that $u(t_0) = v(t_0) \le \max\{u_0, v_0\}$. So, if we take the curve γ in Lemma 3.3 to be the diagonal $u = v$ of Γ_0, then for every $(u_0, v_0) \in \Gamma_0$ the characteristic curve intersects γ. In other words, Γ is contained in $X(\Xi)$. Moreover, as M, N are positive, we see that γ is noncharacteristic. Finally, we need to define the initial condition $\widetilde{g}(u,v) = \phi(u) + \psi(v)$ on γ to fulfill the assumption (3.11) of Lemma 3.3. For $\gamma = \{u = v\}$, (3.11) reads $M(u)\widetilde{g}_u = N(u,u)\widetilde{g}_v$ Clearly, this holds if we choose $\widetilde{g}$ such that $\widetilde{g}_v = p_v^0 = M(u)$ and $\widetilde{g}_u = p_u^0 = N(u,u)$. That is,

$$\phi(u) = \int_0^u N(s,s)ds, \quad \psi(v) = \int_0^v M(s)ds. \tag{4.3}$$

Since $M, N > 0$, we see that $\nabla\widetilde{g} \cdot \vec{T} \neq 0$. Lemma 3.3 is applicable and our proof is complete. □

Remark 4.1. If $d_1 = d_2$, then $f(u,v) = (Au + Bv + \sqrt{a_1u^2 + a_2v^2 + a_3uv})/(a_{21}u)$ for some constants A, B, a_1, a_2, a_3. If $u, v \ge 0$, then $f(u,v) > 0$. We also see that f can be written as $\hat{f}(v/u)$ for some function $\hat{f}$. Therefore, Lemma 3.4 applies here to give an explicit formula for g.

We conclude this paper by considering a generalized version of (1.2)

$$\begin{aligned} u_t &= \nabla((d_1 + a_{12}v + a_{11}u)\nabla u + (c_1 + b_{12}u)\nabla v) + F(u,v), \\ v_t &= \nabla((c_2 + b_{21}v)\nabla u + (d_2 + a_{21}u + a_{22}v)\nabla v) + G(u,v). \end{aligned} \tag{4.4}$$

The constants d_i, c_i, a_{ij}, b_{ij} are assumed to be positive. With minor modifications in the calculation of the proof of Theorem 1, we can prove the following.

THEOREM 4. *Bounded positive solutions to* (4.4) *are Hölder continuous if*

$$a_{11} > a_{21} + b_{21}, \quad a_{22} > a_{12} + b_{12}, \tag{4.5}$$

$$d_1d_2 > c_1c_2, \quad d_1a_{22} + d_2a_{12} > c_1b_{21}, \quad d_1a_{21} + d_2a_{11} > c_2b_{12}. \tag{4.6}$$

Again, these conditions simply say that the self diffusion dominates the cross diffusion. We only remark here that (4.5) gives $f_u - f_v f > 0$, and (4.6) provides $Q_1P_2 - P_1Q_2 < 0$ so that $\alpha_2 < 0$ as required by (H.2). The proof of the existence of g can be repeated here with $M(u) = c_1 + b_{12}u$ (so that the interval $I_0 = (a_0, b_0)$, on which $u(t), v(t)$ are positive, may not be unbounded as before).

Acknowledgment. We would like to thank the anonymous referees for their invaluable comments that lead to many improvements of this version of the paper.

REFERENCES

[1] H. Amann, *Dynamic theory of quasilinear parabolic systems* III. *Global existence*, Math. Z., 202 (1989), pp. 219–250.
[2] L. C. Evans, *Partial Differential Equations*, AMS Graduate Studies in Math. 19, AMS, Providence, RI, 1998.
[3] M. Giaquinta and M. Struwe, *On the partial regularity of weak solutions of nonlinear parabolic systems*, Math. Z., 179 (1982), pp. 437–451.
[4] O. John and J. Stara, *Some (new) counterexamples of parabolic systems*, Comment. Math. Univ. Carolin., 36 (1995), pp. 503–510.
[5] O. John and J. Stara, *On the regularity of weak solutions to parabolic systems in two spatial dimensions*, Comm. Partial Differential Equations, 26 (1998), pp. 1159–1170.
[6] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Ural'tseva, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monogr. 23, AMS, Providence, RI, 1968.
[7] D. Le, *Cross diffusion systems on n spatial dimensional domains*, Indiana Univ. Math. J., 51 (2002), pp. 625–643.
[8] D. Le, *Remark on Hölder continuity for parabolic equations and the convergence to global attractors*, Nonlinear Anal. T.M.A., 41 (2000), pp. 921–941.
[9] D. Le, *Hölder regularity for certain strongly coupled parabolic systems*, J. Differential Equations, 151 (1999), pp. 313–344.
[10] N. Shigesada, K. Kawasaki, and E. Teramoto, *Spatial segregation of interacting species*, J. Theoret. Biol., 79 (1979), pp. 83–99.
[11] M. Wiegner, *Global solutions to a class of strongly coupled parabolic systems*, Math. Ann., 292 (1992), pp. 711–727.

SIAM J. MATH. ANAL.
Vol. 36, No. 6, pp. 1943–1964

Γ-LIMIT OF A PHASE-FIELD MODEL OF DISLOCATIONS*

A. GARRONI† AND S. MÜLLER‡

Abstract. We study, by means of Γ-convergence, the asymptotic behavior of a variational problem modeling a dislocation ensemble moving on a slip plane through a discrete array of obstacles. The variational problem is a two-dimensional phase transition-type energy given by a nonlocal term and a nonlinear potential which penalizes noninteger values. In this paper we consider a regime corresponding to a diluted distribution of obstacles. In this case the leading term of the energy can be described by means of a cell problem formula defining an appropriate notion of capacity (that we call dislocation capacity).

Key words. dislocations, phase transitions, capacity, Γ-convergence

AMS subject classifications. 82B26, 31C15, 49J45

DOI. 10.1137/S003614100343768X

1. Introduction.

1.1. Formulation of the problem. In this paper we begin the study of the large body limit of a phase-field model for dislocations, recently introduced by Koslowski, Cuitino, and Ortiz [10]. This model studies a dislocation ensemble moving within a single slip plane through an array of discrete obstacles (e.g., forest dislocations) under the action of an applied shear stress. In this theory, after a suitable rescaling (see the end of this subsection for the details), the slip (measured in units of the Burgers vectors) on the slip plane is represented by a scalar phase field u, which prefers to take integer values. We will consider a periodic setting; i.e., u will be a periodic scalar-valued function defined on the slip plane which is chosen as $T^2 \times \{0\}$. The first contribution to the energy, the so-called Peierls potential, penalizes noninteger values of the slip distribution u and is given by

$$\frac{1}{2\varepsilon} \int_{T^2} \operatorname{dist}^2(u, \mathbf{Z})\, dx\,. \tag{1.1}$$

Here $T^2 = \mathbf{R}^2/\mathbf{Z}^2$ denotes the standard torus; i.e., functions on T^2 are periodic with period one. The small parameter ε is proportional to the ratio between the Burgers vector (or, equivalently, the lattice spacing) and the physical size of the (periodic) domain under consideration. In particular the large body limit is characterized by the limit $\varepsilon \to 0$. The arguments in this paper do not require the special form of (1.1). Instead of the special integrand $\operatorname{dist}^2(u, \mathbf{Z})$ we could consider a general integrand $W(u)$, where W is a $\mathbf{Z}$ periodic C^1 function satisfying $W(u) \geq c \operatorname{dist}^2(u, \mathbf{Z})$, $c > 0$.

The second term in the energy represents the long-range elastic interaction induced by the slip. This can be obtained by considering a field $\widetilde{u}$ on $T^2 \times \mathbf{R}$ which has

*Received by the editors November 13, 2003; accepted for publication (in revised form) July 16, 2004; published electronically June 22, 2005.
http://www.siam.org/journals/sima/36-6/43768.html
†Dipartimento di Matematica, Università di Roma "La Sapienza", P.le Aldo Moro 3, 00185 Roma, Italy (garroni@mat.uniroma1.it).
‡Max-Planck Institut für Mathematik in den Naturwissenschaften, Inselstr. 22-26, D–04103 Leipzig, Germany (Stefan.Mueller@mis.mpg.de).

a jump of size u across $\{x_3 = 0\}$ and considering its elastic energy

$$\frac{1}{2}\int_{T^2\times\mathbf{R}} |\nabla \widetilde{u}|^2\, dx\,.$$

One can easily verify that the optimal $\widetilde{u}$ (for a given jump u) is antisymmetric in x_3 (up to an irrelevant constant) and the elastic energy is given by the minimizer of the expression

$$\int_{T^2\times(0,+\infty)} |\nabla \widetilde{u}|^2\, dx$$

subject to the boundary condition $\widetilde{u}(x',0) = \frac{1}{2}u(x')$. This energy is nothing but the square of the $H^{\frac{1}{2}}$ seminorm of $\frac{1}{2}u$ which in the Fourier representation is given by

$$\frac{1}{4}[u]^2_{H^{\frac{1}{2}}(T^2)} = \frac{1}{4}\sum_{k\in(2\pi\mathbf{Z})^2} |k||\widehat{u}(k)|^2\,, \tag{1.2}$$

where

$$\widehat{u}(k) = \int_{T^2} \mathrm{e}^{-ikx}u(x)\, dx\,.$$

In real space the energy can be written as

$$\frac{1}{4}[u]^2_{H^{\frac{1}{2}}(T^2)} = \frac{1}{2}\int\!\!\int_{T^2\times T^2} K(x-y)|u(x)-u(y)|^2\, dx\, dy\,, \tag{1.3}$$

where the kernel $K(t)$ has the following properties:

(i) $K(t) = O(|t|^{-3})$ as $|t| \to 0$.

(ii) $K(t)$ is periodic, i.e., is defined in T^2.

In fact, the Fourier coefficients of K are given by $\widehat{K}(k) = -\frac{1}{4}|k|$, so that $K(t) \sim \frac{1}{8\pi}t^{-3}$ as $t \to 0$.

The energy we are really looking at is the isotropic elastic bulk energy given in terms of the symmetrized displacement gradient $e(U) = \frac{1}{2}(\nabla U + \nabla U^T)$ as

$$\int_{T^2\times\mathbf{R}} \mu\,|e(U)|^2 + \frac{\lambda}{2}\,|\mathrm{tr}\, e(U)|^2\; dx'\,,$$

where U is the vector displacement and the jump of U across $\{x_3 = 0\}$ is given by $u\mathbf{e_1}$. Using Fourier variables this leads to the following $H^{\frac{1}{2}}$-like energy:

$$\frac{\mu}{4}\sum_{k\in(2\pi\mathbf{Z})^2} m_\nu(k)|\widehat{u}(k)|^2. \tag{1.4}$$

Here the weight $m_\nu(k)$ is homogeneous of degree 1 and is explicitly given by

$$m_\nu(k) = \frac{k_2^2}{\sqrt{k_1^2+k_2^2}} + \frac{1}{1-\nu}\frac{k_1^2}{\sqrt{k_1^2+k_2^2}}\,,$$

where $\nu < \frac{1}{2}$ is the Poisson ratio (see [10] for a detailed derivation of the above formula). If $\nu = 0$, then $m_\nu(k) = |k|$, and (1.4) reduces to (1.2); we call this the *isotropic case*.

One can also compute the real space version of the energy in (1.4), and this gives the following representation of the elastic energy:

$$\frac{\mu}{2}\iint_{T^2\times T^2} K_\nu(x-y)|u(x)-u(y)|^2\,dx\,dy\,, \tag{1.5}$$

where the kernel $K_\nu(t)$ satisfies conditions (i) and (ii). In fact, K_ν is the Fourier series of $-\frac{1}{4}m_\nu(k)$, and a more explicit formula is given in (2.4) below. It is also clear that this nonlocal energy is controlled from above and below by the $H^{\frac{1}{2}}$-periodic seminorm introduced above; more precisely,

$$[u]^2_{H^{\frac{1}{2}}(T^2)} \le \frac{1}{4}\sum_{k\in(2\pi\mathbf{Z})^2} m_\nu(k)|\widehat{u}(k)|^2 \le \frac{1}{1-\nu}[u]^2_{H^{\frac{1}{2}}(T^2)}. \tag{1.6}$$

We now take $\mu = 1$. Then the total energy is thus given by

$$\frac{1}{2\varepsilon}\int_{T^2}\operatorname{dist}^2(u,\mathbf{Z})\,dx + \frac{1}{2}\iint_{T^2\times T^2} K_\nu(x-y)|u(x)-u(y)|^2\,dx\,dy - \int_{T^2} S^\varepsilon u\,dx. \tag{1.7}$$

The last term of the energy takes into account the interaction with the (nondimensionalized) resolved shear stress S^ε. In this paper we will consider the case that S^ε is of order 1, more precisely that it converges weakly in L^2 as $\varepsilon \to 0$. To study the Γ-limit of the above energy it thus suffices to consider only the first two terms and to regard the third term as a continuous perturbation (see Remark 15). We shall study this energy subject to a pinning condition in order to model, e.g., a forest hardening mechanism by secondary dislocation. For definiteness we focus on the idealization of obstacles with infinitely strong pinning; i.e., we require that u vanishes on the union of discs $B(x_i^\varepsilon, R\varepsilon) = B^i_{R\varepsilon}$ of radius $R\varepsilon$ and centers x_i^ε, $i = 1,\dots,N_\varepsilon$. (The effects of a finite pinning strength are discussed in the appendix.)

To summarize, we will study the asymptotic behavior, in terms of Γ-convergence, of the following functional:

$$E_\varepsilon(u) = \begin{cases} \dfrac{1}{\varepsilon}\displaystyle\int_{T^2}\operatorname{dist}^2(u,\mathbf{Z})\,dx + \iint_{T^2\times T^2} K_\nu(x-y)|u(x)-u(y)|^2\,dx\,dy & \\ \qquad\qquad \text{if } u\in H^{\frac{1}{2}}(T^2)\,, u=0, \text{ on } \bigcup_i B^i_{R\varepsilon}, & \\ +\infty \qquad\qquad \text{otherwise}\,. & \end{cases} \tag{1.8}$$

Before discussing in more detail the relevant limits let us briefly indicate how the above nondimensionalized functional is related to the energy considered in [10]. Passing from the Fourier representation to the real space formulation and directly taking into account the periodicity of the phase field we can write the energy in [10] as follows:

$$\frac{b^2}{2d}\int_{Q_L}\operatorname{dist}^2(v,\mathbf{Z})\,dx' + b^2\iint_{Q_L\times Q_L} L^{-3}K_\nu\left(\frac{x'-y'}{L}\right)|v(x')-v(y')|^2\,dx'\,dy',$$

where Q_L is the square in $\mathbf{R}^2$ of side L, where b is the length of the Burgers vector, and where d is the interplanar spacing (which is of the same order of b). Scaling with $x = \frac{x'}{L}$ and $y = \frac{y'}{L}$ and dividing the energy by b^2L we get (1.8) with ε of order $\frac{b}{L}$.

1.2. Mathematical context and scaling regimes for N_ε. From a mathematical point of view the functional E_ε combines two features: a singular perturbation of Modica–Mortola type and a boundary condition on perforated domains, i.e., domains with many small holes.

If we ignore the boundary condition and also replace the $H^{\frac{1}{2}}$ seminorm by the Dirichlet integral we obtain exactly a version of the Modica–Mortola problem for a potential with infinitely many wells [13, 14]. In this case the typical scaling of the energy is proportional to $1/\sqrt{\varepsilon}$. The situation for the $H^{\frac{1}{2}}$ seminorm is more delicate, and Alberti, Bouchitté, and Seppecher [2] showed (for the case of two wells) that the natural scaling for the energy is $\ln 1/\varepsilon$ and that after rescaling by $1/\ln(1/\varepsilon)$ the Γ-limit of the energy is proportional to the BV-norm $\int |\nabla u|$, i.e., to the length of the jump set of u. (The limiting energy is finite only on functions which take values in the wells.)

If we ignore instead the singular Peierls energy, then our functional falls in the class of variational problems in perforated domains. Again for the Dirichlet integral (and many other local functionals) a large amount of literature is available (see, e.g., [12] and [4], or [6], for a more general approach). The general idea is that in the limit a violation of the boundary condition no longer carries an infinite cost but only a finite cost computed by the integration of a suitable function of u against a suitable measure which captures the local density of holes (in the sense of capacity). At least in terms of scaling, our problem (without the Peierls term) can be reduced to that setting by working with the harmonic extension $\widetilde{u}$ of u. This shows that without Peierls energy, E_ε should scale like the "capacity density" $\varepsilon N_\varepsilon$. Combining these two results we expect two standard scaling regimes.

1. $\varepsilon N_\varepsilon \to 1$. In this case we expect that E_ε is of order 1 and that the limiting energy can be obtained by solving a suitable cell problem which involves one obstacle in an infinite medium with boundary conditions at the obstacle and at infinity. In view of the results of [2] we expect also that the limit energy functional is finite only on constant, integer-valued functions, since a jump would result in an energy cost of order $\ln 1/\varepsilon$. A typical minimizer u_ε of (1.8) looks almost constant with small perturbations (on a length scale ε) near the holes. The shape of these perturbations is essentially determined by the cell problem.
2. $\varepsilon N_\varepsilon/\ln(1/\varepsilon) \to 1$. In this case the contributions of the pinning energy discussed above and the Modica–Mortola-like energy are of the same order. After rescaling by $1/\ln(1/\varepsilon)$ we expect a limiting energy of the form $\int |\nabla u| + \int D(u)$ (subject to the constraint $u(x) \in \mathbf{Z}$ almost everywhere) where, as before, the function $D(a)$ is computed from a cell problem with boundary condition a at infinity. In the physics literature this functional is referred to as a line-tension model because the first term penalizes the length of the jump set of u. In fact, for the anisotropic kernel K_ν above, the term $|\nabla u|$ has to be replaced by an anisotropic line energy of the type $\gamma(\frac{\nabla u}{|\nabla u|})|\nabla u|$.

In this paper we investigate the first regime $\varepsilon N_\varepsilon \sim 1$ (see Theorem 14 below for a precise statement). In fact, it turns out that the regime $\varepsilon N_\varepsilon \to 0$ can be handled in exactly the same way if we scale the energy by $1/(\varepsilon N_\varepsilon)$ (see also Theorem 14). Going back to (1.8) the natural scaling of the resolved shear stress in this regime is $S_\varepsilon \sim \varepsilon N_\varepsilon$. The regimes $\varepsilon N_\varepsilon \sim \ln 1/\varepsilon$ and $\varepsilon N_\varepsilon \gg \ln 1/\varepsilon$ will be discussed in a forthcoming paper [8].

Finally, we remark that while our analysis is phrased in terms of statics the same energy functional arises in the approximation of the evolution through a sequence of

minimization problems at discrete times (see [10, Chapter 4]). Actually in this case it is more natural to consider a "soft" pinning condition which can be treated exactly along the same lines (see the appendix). In this case the "pinning energy" represents the energy dissipated in crossing one of the pinning sites. A full understanding of dislocation dynamics and its macroscopic consequences, such as hysteresis, will of course require a much better understanding of local minimizers (and the energy barriers between them). To do this rigorously seems currently out of reach. Nonetheless we can identify (in this paper and in [8]) in a rigorous way the relevant scaling regimes for the competition of elastic energy, applied stress, pinning energy, dissipation, and line energy of the dislocations, and we believe that this will be helpful for further studies.

2. The nonlocal energy. This section will be devoted to recalling some basic properties of the nonlocal part of the energy. By minimizing an elastic energy on $\mathbf{R}^3_+$ we get, as in the periodic case, a nonlocal energy equivalent to the $H^{\frac{1}{2}}(\mathbf{R}^2)$ seminorm. Indeed, as above, similar considerations give an energy of the form

$$\int_{\mathbf{R}^2} \left(\frac{\lambda_2^2}{|\lambda|} + \frac{1}{1-\nu} \frac{\lambda_1^2}{|\lambda|} \right) |\widehat{u}(\lambda)|^2 d\lambda,$$

which can be written in spatial variables as

$$\frac{1}{2} \iint_{\mathbf{R}^2 \times \mathbf{R}^2} \Gamma_\nu(x-y)|u(x)-u(y)|^2 \, dx \, dy \,, \tag{2.1}$$

where the Fourier transform of the kernel $\Gamma_\nu(t)$, with $t \in \mathbf{R}^2$, is $-\frac{1}{4}(\frac{\lambda_2^2}{|\lambda|} + \frac{1}{1-\nu}\frac{\lambda_1^2}{|\lambda|})$ and it can be computed explicitly, i.e.,

$$\Gamma_\nu(t) = \frac{1}{2\pi(1-\nu)|t|^3} \left(\nu + 1 - 3\nu \frac{t_2^2}{|t|^2} \right) . \tag{2.2}$$

In particular it is homogeneous of degree -3 and is positive if $\nu < \frac{1}{2}$. Clearly we also have

$$[u]^2_{H^{\frac{1}{2}}(\mathbf{R}^2)} \le \frac{1}{2} \iint_{\mathbf{R}^2 \times \mathbf{R}^2} \Gamma_\nu(x-y)|u(x)-u(y)|^2 \, dx \, dy \, \le \frac{1}{1-\nu} [u]^2_{H^{\frac{1}{2}}(\mathbf{R}^2)} \,, \tag{2.3}$$

where

$$[u]_{H^{\frac{1}{2}}(\mathbf{R}^2)} := \left(\frac{1}{4\pi} \iint_{\mathbf{R}^2 \times \mathbf{R}^2} \frac{|u(x)-u(y)|^2}{|x-y|^3} \, dx \, dy \right)^{\frac{1}{2}} = \left(\int_{\mathbf{R}^2} |\lambda| |\widehat{u}(\lambda)|^2 \, d\lambda \right)^{\frac{1}{2}} .$$

PROPOSITION 1. *Let $K_\nu(t)$ be the anisotropic kernel defined above for the periodic case, and let $\Gamma_\nu(t)$ be the corresponding kernel in $\mathbf{R}^2$. Then there exists a constant $C > 0$ such that*

$$|\Gamma_\nu(t) - K_\nu(t)| \le C$$

on $\{t \in \mathbf{R}^2 \,:\, |t_i| \le 3/4\}$.

Proof. By the Poisson summation formula (see, e.g., Stein and Weiss [15, Corollary 2.6]) we have

$$K_\nu(t) = \sum_{z\in\mathbf{Z}^2} \Gamma_\nu(t+z)\,. \tag{2.4}$$

In particular, for any $t \in \mathbf{R}^2$ such that $|t_i| \le 3/4$ we get

$$|K_\nu(t) - \Gamma_\nu(t)| = \sum_{z\in\mathbf{Z}^2\setminus\{0\}} \Gamma_\nu(t+z) \le \sum_{z\in\mathbf{Z}^2\setminus\{0\}} \frac{c}{|z|^3} \le C\,. \qquad \square$$

Remark 2. By Proposition 1, using the homogeneity of Γ_ν we deduce that for every $\delta > 0$

$$\lim_{\varepsilon\to 0} \varepsilon^3 K_\nu(\varepsilon t) = \Gamma_\nu(t)$$

uniformly on $\{t \in \mathbf{R}^2 : |t| \le \delta\}$.

From the definition of $[\cdot]_{H^{\frac12}}$ as a trace seminorm we can deduce a Poincaré-type inequality for functions in $H^{\frac12}(T^2)$. For a given bounded domain $D \subseteq \mathbf{R}^3$ a refinement of the classical Poincaré inequality permits us to estimate the L^2 norm of a function in $H^1(D)$ with the L^2 norm of its gradient, as long as the set where the function is zero is not too small. There exists a constant C such that for every $w \in H^1(D)$

$$\int_D |w|^2 dx' \le \frac{C}{\mathrm{Cap}(N(w))} \int_D |\nabla w|^2 dx'\,, \tag{2.5}$$

where $\mathrm{Cap}(N(w))$ denote the harmonic capacity (with respect to $\mathbf{R}^3$) of the set $N(w) = \{x' \in D : w(x) = 0\}$ (see [16, Corollary 4.5.2] or [7, Theorem 2.9]). (Note that in view of (2.5) the set $N(w)$ is well defined, since the pointwise value of w can be specified up to a set of zero harmonic capacity using its quasi-continuous representative.)

Proposition 3. *There exists a constant C_0 such that for every $u \in H^{\frac12}(T^2)$, with $u = 0$ on $E \subseteq T^2$, we have*

$$\int_{T^2} |u|^2 dx \le C_0 \left(1 + \frac{1}{\mathrm{Cap}(E\times\{0\})}\right) [u]^2_{H^{\frac12}(T^2)}\,, \tag{2.6}$$

where $\mathrm{Cap}(E\times\{0\})$ *denote the harmonic capacity of* $E\times\{0\}$ *as a subset of* $\mathbf{R}^3$.

Proof. The proof follows immediately by applying (2.5) to the harmonic extension $\widetilde{u}$ of u in $D = (0,1)^3$ and by the fact that

$$\int_{T^2} |u|^2 dx \le c\,\|\widetilde{u}\|^2_{H^1(D)}\,. \qquad \square$$

Remark 4. Given an arbitrary $H^{\frac12}(Q)$ function we can extend by reflection to a periodic function on Q_2, and applying the above inequality we get that there exists a constant C_1 such that

$$\int_Q |u|^2 dx \le C_1 \left(1 + \frac{1}{\mathrm{Cap}(E\times\{0\})}\right) \iint_{Q\times Q} \frac{|u(x)-u(y)|^2}{|x-y|^3} dx\,dy\,. \tag{2.7}$$

3. The cell problem. In this section we will define a suitable notion of capacity which will be the natural tool to study the asymptotics of our problem. We call the following set function the "$H^{\frac{1}{2}}$ dislocation capacity of an open set E with respect to Ω at the integer level $a \in \mathbf{Z}$":

(3.1)
$$D^{\nu}_{\frac{1}{2}}(a,E,\Omega) := \inf\Bigg\{ \int_{\mathbf{R}^2} \mathrm{dist}^2(\zeta,\mathbf{Z})\,dx + \iint_{\mathbf{R}^2\times\mathbf{R}^2} \Gamma_\nu(x-y)|\zeta(x)-\zeta(y)|^2\,dx\,dy : \zeta = a \text{ on } E\,,\ \zeta = 0 \text{ on } \mathbf{R}^2\setminus\Omega \Bigg\}.$$

We denote by $D^{\nu}_{\frac{1}{2}}(a,E)$ the "$H^{\frac{1}{2}}$ dislocation capacity of an open set E with respect to $\mathbf{R}^2$ at the integer level $a \in \mathbf{Z}$"; namely,

(3.2)
$$D^{\nu}_{\frac{1}{2}}(a,E) := \inf\Bigg\{ \int_{\mathbf{R}^2} \mathrm{dist}^2(\zeta,\mathbf{Z})\,dx + \iint_{\mathbf{R}^2\times\mathbf{R}^2} \Gamma_\nu(x-y)|\zeta(x)-\zeta(y)|^2\,dx\,dy : \zeta = a \text{ on } E\,,\ \zeta \in L^4(\mathbf{R}^2) \Bigg\}.$$

The condition $\zeta \in L^4(\mathbf{R}^2)$ is the natural condition at ∞, in view of the following Sobolev inequality:

(3.3)
$$\|u\|_{L^4(\mathbf{R}^2)} \le C^*[u]_{H^{\frac{1}{2}}(\mathbf{R}^2)} \qquad \forall\ u \in C_0^\infty(\mathbf{R}^2)\,.$$

Remark 5. Denote

(3.4)
$$I(\zeta) := \int_{\mathbf{R}^2} \mathrm{dist}^2(\zeta,\mathbf{Z})\,dx + \iint_{\mathbf{R}^2\times\mathbf{R}^2} \Gamma_\nu(x-y)|\zeta(x)-\zeta(y)|^2\,dx\,dy\,.$$

Using the fact that the kernel Γ_ν is positive (under the assumption $\nu < \frac{1}{2}$), it is easy to check that both terms in the energy are decreasing under truncations by integers. For every $a, b \in \mathbf{R}$ we set $a \wedge b = \min(a,b)$ and $a \vee b = \max(a,b)$. Then for every $t \in \mathbf{Z}$, we have $I(\zeta \wedge t) \le I(\zeta)$ (and $I(\zeta \vee t) \le I(\zeta)$). Moreover, $I(\zeta \wedge t) < I(\zeta)$ (and $I(\zeta \vee t) < I(\zeta)$) unless $\zeta \wedge t = \zeta$ a.e. (or $\zeta \vee t = \zeta$ a.e.).

PROPOSITION 6. *Let Ω be a bounded open subset of $\mathbf{R}^2$, and let $E \subseteq \Omega$ be an open set such that $D^{\nu}_{\frac{1}{2}}(a,E,\Omega) < +\infty$, with $a \in \mathbf{Z}$. Then there exists a minimum point $\zeta \in H^{\frac{1}{2}}(\Omega)$ for* (3.1) *and it satisfies $0 \le \zeta \le a$. We will call each minimum point a $D^{\nu}_{\frac{1}{2}}$-capacitary potential of E with respect to Ω.*

Proof. In order to obtain the existence let ζ_k be a minimizing sequence for (3.1). By Remark 5 we may assume that $0 \le \zeta_k \le a$. Now let $\xi_k : \mathbf{R}^2 \to \mathbf{Z}$ such that

$$\int_{\mathbf{R}^2} \mathrm{dist}^2(\zeta_k,\mathbf{Z})\,dx = \int_{\mathbf{R}^2} |\zeta_k - \xi_k|^2 dx\,,$$

i.e., $\xi_k = \mathbf{P}_{\mathbf{Z}}\zeta_k$. Clearly $\xi_k = 0$ on $\mathbf{R}^2 \setminus \Omega$, and

$$\lim_{k\to\infty} \int_{\mathbf{R}^2} |\zeta_k - \xi_k|^2\,dx + \iint_{\mathbf{R}^2\times\mathbf{R}^2} \Gamma_\nu(x-y)|\zeta_k(x)-\zeta_k(y)|^2\,dx\,dy = D^{\nu}_{\frac{1}{2}}(a,E,\Omega)\,.$$

By (2.3) and the assumption $D^{\nu}_{\frac12}(a,E,\Omega)<+\infty$, the sequence $\{(\zeta_k,\xi_k)\}$ is bounded in $H^{\frac12}(\mathbf{R}^2)\times L^2(\Omega,\mathbf{Z})$, and we may assume that ζ_k and ξ_k converge weakly to a pair (ζ,ξ). By lower semicontinuity the limiting pair minimizes the energy and clearly satisfies $0\le\zeta\le a$ and $\zeta=0$ on $\mathbf{R}^2\setminus\Omega$. □

Remark 7. If $\Omega=B_R$, $E=B_r$, and $\nu=0$ (in the "isotropic case") the capacitary potential is unique and radially symmetric. This follows immediately by the fact that both terms of the energy defining $D_{\frac12}(a,B_r,B_R)$ are rotation invariant and the $H^{\frac12}$ seminorm is strictly decreasing under radial rearrangements. (For results on rearrangements for nonlocal energies see, e.g., [9] or the review paper [3].)

PROPOSITION 8. *There exists a minimizer for problem* (3.2)*, the $D^{\nu}_{\frac12}$-capacitary potential of E with respect to $\mathbf{R}^2$. If, moreover, E is bounded, then every minimizer converges to zero uniformly at infinity.*

Proof. As above, the existence follows by considering a minimizing sequence and remarking that the "boundary condition" $\zeta-a\in L^4(\mathbf{R}^2)$ is preserved, in view of (3.3).

Let now ζ be a potential of a bounded set E with respect to $\mathbf{R}^2$. In the case $\nu=0$ the decay at infinity follows by a comparison argument with the radially symmetric case (see Remark 7). Let us consider now the general case. Let L be the linear operator representing the quadratic form defined by (2.1), so that

$$I(\zeta)=\langle L\zeta,\zeta\rangle+\int_{\mathbf{R}^2}\mathrm{dist}^2(\zeta,\mathbf{Z})\,dx\,.$$

We will see that there exist a function $\psi\in L^4(\mathbf{R}^2)$ and a measure μ supported on $\overline{E}$ such that ζ is the solution in the sense of distributions of

$$L\zeta=\mu+\psi\,. \tag{3.5}$$

This would follow immediately from the Euler–Lagrange equation if $\mathrm{dist}^2(\cdot,\mathbf{Z}^2)$ was a C^1 function. For the case at hand we can argue as follows.

Fix $\eta\in C_0^\infty(\mathbf{R}^2)$, with $\eta\ge0$ on E, and compute the variation of $I(\zeta)$ in direction η. We have

$$2\langle L\zeta,\eta\rangle+t\langle L\eta,\eta\rangle+\int_{\mathbf{R}^2}\frac{\mathrm{dist}^2(\zeta+t\eta,\mathbf{Z})-\mathrm{dist}^2(\zeta,\mathbf{Z})}{t}\,dx\ge0\,. \tag{3.6}$$

Since $\mathrm{dist}(\cdot,\mathbf{Z})$ is a Lipschitz function,

$$\limsup_{t\to0}\frac{\mathrm{dist}^2(\zeta(x)+t\eta(x),\mathbf{Z})-\mathrm{dist}^2(\zeta(x),\mathbf{Z})}{t}\le2\mathrm{dist}(\zeta(x),\mathbf{Z})\,\eta(x)\qquad\text{a.e. } x\in\mathbf{R}^2,$$

and hence, by Fatou's lemma, we get

$$\langle L\zeta,\eta\rangle+\int_{\mathbf{R}^2}\mathrm{dist}(\zeta,\mathbf{Z})\,\eta\,dx\ge0\qquad\forall\,\eta\in C_0^\infty(\mathbf{R}^2)\,,\quad\eta\ge0\,.$$

Thus there exists a positive measure $\widetilde{\mu}$ such that

$$L\zeta=\widetilde{\mu}-\mathrm{dist}(\zeta,\mathbf{Z}) \tag{3.7}$$

in the sense of distributions. Now consider $\eta\in C_0^\infty(\mathbf{R}^2)$, with $\eta=0$ on E. We can apply the above argument to η and $-\eta$ and get

$$|\langle L\zeta,\eta\rangle|\le\int_{\mathbf{R}^2}\mathrm{dist}(\zeta,\mathbf{Z})|\eta|\,dx\qquad\forall\eta\in C_0^\infty(\mathbf{R}^2\setminus\overline{E})\,.$$

By density this holds also for $\eta \in C_0^0(\mathbf{R}^2 \setminus \overline{E})$, and since $\operatorname{dist}(\zeta, \mathbf{Z}) \in L^4(\mathbf{R}^2)$ we deduce that the restriction of $\widetilde{\mu}$ to $\mathbf{R}^2 \setminus \overline{E}$ is absolutely continuous with respect to the Lebesgue measure and its density is a L^4 function. Thus $\widetilde{\mu}$ can be written as the sum of a measure μ supported on $\overline{E}$ and a function belonging to $L^4(\mathbf{R}^2)$ which together with (3.7) gives (3.5).

Now let G_ν be the Green function of the operator $L+I$. We will need only rather mild decay properties of G_ν at ∞. To verify these it suffices to note that the Fourier transform of G_ν is given by

$$\widehat{G_\nu}(\lambda) = \frac{1}{1 + \frac{\lambda_2^2}{|\lambda|} + \frac{1}{1-\nu}\frac{\lambda_1^2}{|\lambda|}},$$

since $\widehat{L\zeta}(\lambda) = (\frac{\lambda_2^2}{|\lambda|} + \frac{1}{1-\nu}\frac{\lambda_1^2}{|\lambda|})\,\widehat{\zeta}(\lambda)$. One easily sees that $\widehat{G_\nu}$ and its first two derivatives are in $L^1(\mathbf{R}^2)$. Hence G_ν is continuous, and $G_\nu(x) \le C/(1+|x|^2)$. In particular, $G_\nu \in L^{4/3}(\mathbf{R}^2)$.

Since ψ and ζ belong to $L^4(\mathbf{R}^2)$, for every $\varepsilon > 0$ we can write $\psi + \zeta = \psi_1 + \psi_2$, where ψ_1 has compact support and $\|\psi_2\|_4 \le \varepsilon$. Thus we have

$$\zeta(x) = G_\nu * \mu(x) + G_\nu * \psi_1(x) + G_\nu * \psi_2(x) \qquad \text{a.e. } \mathbf{R}^2\,.$$

We can estimate the L^∞ norm of the last term of the right-hand side using the Hölder inequality and get

$$\|G_\nu * \psi_2\|_{L^\infty(\mathbf{R}^2)} \le \|G_\nu\|_{L^{\frac{4}{3}}(\mathbf{R}^2)} \|\psi_2\|_{L^4(\mathbf{R}^2)} \le C\varepsilon\,.$$

On the other hand, the decay of G_ν and the fact that μ and ψ_1 have compact support guarantee that for $|x|$ big enough ζ is uniformly small. This concludes the proof of the proposition. □

The following proposition shows that as $a \to \infty$ the Peierls potential term becomes negligible and the dislocation capacity converges to the $H^{\frac{1}{2}}$-capacity, defined for any open set $E \subseteq \mathbf{R}^2$ as

(3.8)

$$\operatorname{Cap}^\nu_{H^{\frac{1}{2}}}(E) = \inf\left\{ \int\!\!\int_{\mathbf{R}^2\times\mathbf{R}^2} \Gamma_\nu(x-y)|\eta(x)-\eta(y)|^2\,dx\,dy :\ \eta = 1 \text{ on } E\,,\ \eta \in L^4(\mathbf{R}^2) \right\}$$

(see, for instance, [1]).

PROPOSITION 9. *For any bounded open set $E \subseteq \mathbf{R}^2$ there exists a positive constant C_E such that*

$$a^2 \operatorname{Cap}^\nu_{H^{\frac{1}{2}}}(E) \le D^\nu_{\frac{1}{2}}(a,E) \le a^2 \operatorname{Cap}^\nu_{H^{\frac{1}{2}}}(E) + 2a^{3/2} \operatorname{Cap}^\nu_{H^{\frac{1}{2}}}(E) + C_E a \tag{3.9}$$

for every $a \in \mathbb{N}$. In particular,

$$\lim_{a\to\infty} \frac{D^\nu_{\frac{1}{2}}(a,E)}{a^2} = \operatorname{Cap}^\nu_{H^{\frac{1}{2}}}(E)\,. \tag{3.10}$$

Proof. The first inequality in (3.9) is trivial. In order to prove the estimate from above let η_E be the $H^{\frac{1}{2}}$-potential of E, i.e., the minimum point for (3.8). Using the fact that $\widehat{\Gamma_\nu}(\lambda)$ is homogeneous of degree 1, nonvanishing, and smooth on the unit

sphere, one can easily check that η_E decays at infinity as $1/|x|$. Fix $a \in \mathbb{N}$ and define the function

$$v_a(x) = \begin{cases} \frac{a}{a-\sqrt{a}}(a\eta_E(x) - \sqrt{a}) & \text{if } x \in E_a = \{\eta_E > 1/\sqrt{a}\}, \\ 0 & \text{otherwise.} \end{cases}$$

The function v_a is admissible in the definition of $D^\nu_{\frac{1}{2}}(a, E)$; thus

(3.11)
$$D^\nu_{\frac{1}{2}}(a, E) \le \int_{E_a} \operatorname{dist}^2(\tfrac{a}{a-\sqrt{a}}(a\eta_E - \sqrt{a}), \mathbf{Z})\, dx + \iint_{\mathbf{R}^2\times\mathbf{R}^2} \Gamma_\nu(x-y)|v_a(x) - v_a(y)|^2 dx\, dy\,.$$

By the decay of η_E at infinity we have that there exists a constant C_E such that $|E_a| \le C_E a$. Thus

$$\int_{E_a} \operatorname{dist}^2(\tfrac{a}{a-\sqrt{a}}(a\eta_E - \sqrt{a}), \mathbf{Z})\, dx \le C_E a\,.$$

Moreover,

$$\begin{aligned} &\iint_{\mathbf{R}^2\times\mathbf{R}^2} \Gamma_\nu(x-y)|v_a(x) - v_a(y)|^2 dx\, dy \\ &\quad \le \frac{a^2}{(1-\frac{1}{\sqrt{a}})^2} \iint_{\mathbf{R}^2\times\mathbf{R}^2} \Gamma_\nu(x-y)|\eta_E(x) - \eta_E(y)|^2 dx\, dy \\ &\quad = \frac{a^2}{(1-\frac{1}{\sqrt{a}})^2} \mathrm{Cap}^\nu_{H^{\frac{1}{2}}}(E)\,. \end{aligned}$$

After possibly modifying the value of C_E this yields (3.8). □

We can extend the dislocation capacity to the class of all subsets of Ω by setting

$$D^\nu_{\frac{1}{2}}(a, E, \Omega) = \inf\{D^\nu_{\frac{1}{2}}(a, A, \Omega) \,:\, A \text{ open}\,,\ A \supseteq E\} \tag{3.12}$$

for any set $E \subseteq \Omega$.

PROPOSITION 10. *The dislocation capacity satisfies the following properties:*

(1) $D^\nu_{\frac{1}{2}}(a, E, \Omega) \le D^\nu_{\frac{1}{2}}(a, F, \Omega)$ *if* $E \subseteq F \subseteq \Omega$;
(2) $D^\nu_{\frac{1}{2}}(a, E, \Omega) \le D^\nu_{\frac{1}{2}}(a, E, \Omega')$ *if* $E \subseteq \Omega' \subseteq \Omega$;
(3) $D^\nu_{\frac{1}{2}}(a, E, \Omega) = D^\nu_{\frac{1}{2}}(-a, E, \Omega)$ *for every* $a \in \mathbf{Z}$;
(4) *If* $0 \le a \le b$, $a, b \in \mathbf{Z}$, *then* $D^\nu_{\frac{1}{2}}(a, E, \Omega) \le D^\nu_{\frac{1}{2}}(b, E, \Omega)$;
(5) *(Subadditivity). Given two open subsets of* Ω, E_1 *and* E_2,

$$D^\nu_{\frac{1}{2}}(a, E_1 \cup E_2, \Omega) \le D^\nu_{\frac{1}{2}}(a, E_1, \Omega) + D^\nu_{\frac{1}{2}}(a, E_2, \Omega)$$

for every $a \in \mathbf{Z}$;
(6) *(Continuity on increasing sequences of sets). Given a sequence of subsets* $E_n \subseteq \Omega$, *such that* $E_n \subseteq E_{n+1}$, *and letting* $E = \bigcup_n E_n$, *we have*

$$\lim_{n\to\infty} D^\nu_{\frac{1}{2}}(a, E_n, \Omega) = D^\nu_{\frac{1}{2}}(a, E, \Omega)$$

for every $a \in \mathbf{Z}$;

(7) *(Continuity on decreasing sequences of compact sets). Given a sequence of compact subsets of* Ω, K_n, *such that* $K_n \supseteq K_{n+1}$, *and letting* $K = \bigcap_n K_n$, *we have*

$$\lim_{n\to\infty} D^{\nu}_{\frac{1}{2}}(a, K_n, \Omega) = D^{\nu}_{\frac{1}{2}}(a, K, \Omega)$$

for every $a \in \mathbf{Z}$.

Proof. The monotonicity properties (1)–(4) can be checked directly by the definition. In order to prove property (5) let ζ_1 and ζ_2 be capacitary potentials of E_1 and E_2, respectively. Clearly the function $\zeta_1 \vee \zeta_2$ is a good competitor for the $D^{\nu}_{\frac{1}{2}}$-capacity of $E_1 \cup E_2$. The conclusion follows by the explicit computation of the energy, remarking that

$$|\zeta_1(x) - \zeta_2(y)|^2 \le |\zeta_1(x) - \zeta_1(y)|^2 \vee |\zeta_2(x) - \zeta_2(y)|^2$$

if $\zeta_1(x) \ge \zeta_2(x)$ and $\zeta_2(y) \ge \zeta_1(y)$.

Let us prove property (6). By (1) we have

$$\lim_{n\to\infty} D^{\nu}_{\frac{1}{2}}(a, E_n, \Omega) \le D^{\nu}_{\frac{1}{2}}(a, E, \Omega)\,.$$

For the reverse inequality it is enough to consider the case of open sets. Let $\zeta_n \in H^{\frac{1}{2}}(\mathbf{R}^2)$ be a sequence of capacitary potentials, i.e., $\zeta_n = a$ a.e. on E_n and

$$I(\zeta_n) = D^{\nu}_{\frac{1}{2}}(a, E_n, \Omega) \le D^{\nu}_{\frac{1}{2}}(a, E, \Omega)\,.$$

Thus the $H^{\frac{1}{2}}$ seminorm of ζ_n is bounded. In view of (3.3), ζ_n is bounded in L^4, and hence in $H^{\frac{1}{2}}_{\mathrm{loc}}(\mathbf{R}^2)$. Thus (a subsequence of) ζ_n converges strongly in $L^2_{\mathrm{loc}}(\mathbf{R}^2)$. By the lower semicontinuity of $I(\cdot)$ we get

$$I(\zeta) \le \liminf_{n\to\infty} I(\zeta_n) = \liminf_{n\to\infty} D^{\nu}_{\frac{1}{2}}(a, E_n, \Omega)\,.$$

Since $E = \bigcup_n E_n$ is also open, $\zeta = a$ a.e. in E, and hence is a good competitor for the definition of $D^{\nu}_{\frac{1}{2}}(a, E, \Omega)$. Thus

$$D^{\nu}_{\frac{1}{2}}(a, E, \Omega) \le I(\zeta) \le \liminf_{n\to\infty} D^{\nu}_{\frac{1}{2}}(a, E_n, \Omega),$$

which concludes the proof.

Finally, the proof of property (7) follows directly from the definition. In fact, for any fixed $\varepsilon > 0$ there exists an open set $A \supseteq K$ such that $D^{\nu}_{\frac{1}{2}}(a, A, \Omega) \le D^{\nu}_{\frac{1}{2}}(a, K, \Omega) + \varepsilon$. Since the sets K_n are decreasing and compact, there exists an index n_0 such that $K_n \subseteq A$ for every $n \ge n_0$. The conclusion follows by the monotonicity of the dislocation capacity. □

Remark 11. The properties proved above show that the dislocation capacity is a Choquet capacity (see, for instance, [1] or [11] for a general capacity theory). Note, however, that we define the dislocation capacity starting from open sets instead of compact sets, since this is more convenient for the present purpose.

In the following we will mostly consider the dislocation capacity of a ball with respect to either a concentric ball or $\mathbf{R}^2$. In order to simplify the notation we will state and prove some properties of the capacity in this particular case, although most of them hold in a more general situation.

PROPOSITION 12. *Let $a \in \mathbf{Z}$ and $r > 0$. Then*

$$\lim_{R\to\infty} D^\nu_{\frac12}(a, B_r, B_R) = D^\nu_{\frac12}(a, B_r)\,.$$

Proof. By Proposition 10 we know that $D^\nu_{\frac12}(a, B_r, B_R)$ is decreasing in R. Thus the limit always exists, and

$$\lim_{R\to\infty} D^\nu_{\frac12}(a, B_r, B_R) \geq D^\nu_{\frac12}(a, B_r)\,.$$

In order to prove the reverse inequality, fix $a \in \mathbf{Z}$ positive, and let ζ be a capacitary potential of B_r with respect to $\mathbf{R}^2$. We may assume that $0 \leq \zeta \leq a$, and by Proposition 8 we know that ζ decays to zero uniformly at infinity; i.e., for every $\varepsilon > 0$ there exists $R_0 > 0$ such that

$$|\zeta(x)| < \varepsilon \qquad \text{if} \quad |x| \geq R_0\,.$$

Fix $\varepsilon > 0$, and let us define the function ζ_ε as follows:

$$\zeta_\varepsilon(x) = \begin{cases} \zeta(x) & \text{if } x \in \{\zeta \geq 1\},\\ \dfrac{\zeta(x)-\varepsilon}{1-\varepsilon} & \text{if } x \in \{\varepsilon \leq \zeta < 1\},\\ 0 & \text{if } x \in \{\zeta < \varepsilon\}. \end{cases}$$

We can write ζ_ε as $\zeta_\varepsilon(x) = \max\{\min\{\zeta(x), \frac{\zeta(x)-\varepsilon}{1-\varepsilon}\}, 0\}$, and hence $\zeta_\varepsilon \in H^{\frac12}(\mathbf{R}^2)$. Moreover, $\zeta_\varepsilon(x) = a$ on B_r, and $\zeta_\varepsilon(x) = 0$ in $\mathbf{R}^2 \setminus B_{R_0}$. This implies that ζ_ε is an admissible function in the definition of $D^\nu_{\frac12}$ and

(3.13)

$$D^\nu_{\frac12}(a, B_r, B_R) \leq \int_{\mathbf{R}^2} \operatorname{dist}^2(\zeta_\varepsilon, \mathbf{Z})\,dx + \iint_{\mathbf{R}^2\times\mathbf{R}^2} \Gamma_\nu(x-y)|\zeta_\varepsilon(x)-\zeta_\varepsilon(y)|^2\,dx\,dy = I(\zeta_\varepsilon)$$

for every $R \geq R_0$. Let us compute $I(\zeta_\varepsilon)$ and show that $I(\zeta_\varepsilon) \leq I(\zeta) + o(1)$, as $\varepsilon \to 0$. Denote by E_t the level set $\{\zeta \leq t\}$, and let us first estimate the dislocation part of the energy

$$\begin{aligned}\int_{\mathbf{R}^2} \operatorname{dist}^2(\zeta_\varepsilon, \mathbf{Z})\,dx &= \int_{\mathbf{R}^2\setminus E_1} \operatorname{dist}^2(\zeta, \mathbf{Z})\,dx + \frac{1}{(1-\varepsilon)^2}\int_{E_1\setminus E_\varepsilon} \operatorname{dist}^2(\zeta-\varepsilon, (1-\varepsilon)\mathbf{Z})\,dx\\ &= \int_{\mathbf{R}^2\setminus E_1} \operatorname{dist}^2(\zeta, \mathbf{Z})\,dx + \frac{1}{(1-\varepsilon)^2}\int_{E_1\setminus E_{\frac{1+\varepsilon}{2}}} |\zeta-1|^2 dx + \frac{1}{(1-\varepsilon)^2}\int_{E_{\frac{1+\varepsilon}{2}}\setminus E_\varepsilon} |\zeta-\varepsilon|^2 dx\,.\end{aligned}$$

Since

$$\int_{E_{\frac{1+\varepsilon}{2}}\setminus E_\varepsilon} |\zeta-\varepsilon|^2 dx \leq \int_{E_{\frac{1+\varepsilon}{2}}\setminus E_\varepsilon} |\zeta|^2 dx,$$

we have

$$\limsup_{\varepsilon\to0} \int_{\mathbf{R}^2} \operatorname{dist}^2(\zeta_\varepsilon, \mathbf{Z})\,dx \leq \int_{\mathbf{R}^2} \operatorname{dist}^2(\zeta, \mathbf{Z})\,dx\,.$$

To estimate the nonlocal term in $I(\zeta_\varepsilon)$ it suffices to note that $\zeta_\varepsilon = \psi_\varepsilon \circ \zeta$ with $\operatorname{Lip}\psi_\varepsilon \leq \frac{1}{1-\varepsilon}$. Hence $|\zeta_\varepsilon(x)-\zeta_\varepsilon(y)|^2 \leq (1-\varepsilon)^{-2}|\zeta(x)-\zeta(y)|^2$, and we get

$$\limsup_{\varepsilon\to0} \iint_{\mathbf{R}^n\times\mathbf{R}^n} \Gamma_\nu(x-y)|\zeta_\varepsilon(x)-\zeta_\varepsilon(y)|^2\,dx\,dy \leq \iint_{\mathbf{R}^n\times\mathbf{R}^n} \Gamma_\nu(x-y)|\zeta(x)-\zeta(y)|^2\,dx\,dy\,.$$

Thus the conclusion follows from (3.13). ∎

4. Compactness and Γ-convergence result. In this section we will study the Γ-convergence of the functional E_ε defined in (1.8) with a pinning condition on N_ε balls of radius εR and centered in uniformly distributed and well-separated points x_i^ε, $i \in I_\varepsilon \subset \mathbb{N}$, in the regime where $N_\varepsilon \varepsilon$ is bounded.

For every $i \in I_\varepsilon$ and $r > 0$ we denote by B_r^i the ball in $\mathbf{R}^2$ of radius r and center x_i^ε. (B_r always denotes the ball of radius r centered at 0.)

In order to get a nontrivial result we rescale the function E_ε and prove that the functional $F_\varepsilon(u) := E_\varepsilon(u)/N_\varepsilon\varepsilon$, i.e.,

$$F_\varepsilon(u) = \begin{cases} \dfrac{1}{N_\varepsilon \varepsilon^2} \displaystyle\int_{T^2} \operatorname{dist}^2(u, \mathbf{Z})\, dx + \dfrac{1}{N_\varepsilon \varepsilon} \displaystyle\iint_{T^2 \times T^2} K_\nu(x-y)|u(x)-u(y)|^2\, dx\, dy \\ \qquad\qquad \text{if } \ u \in H^{\frac{1}{2}}(T^2)\,, \\ \qquad\qquad u = 0 \text{ on } \bigcup_i B^i_{R\varepsilon}, \\ +\infty \qquad\qquad \text{otherwise}\,, \end{cases}$$

Γ-converges, with respect to the strong L^2 topology, to the functional

$$F(u) = \begin{cases} D^\nu_{\frac{1}{2}}(u, B_R) & \text{if } u = \text{const.} \in \mathbf{Z}\,, \\ +\infty & \text{otherwise}\,. \end{cases} \tag{4.1}$$

For every subset E of $(0,1)^2$ we denote $I_\varepsilon(E) := \{i \in I_\varepsilon \,:\, x_i^\varepsilon \in E\}$. For the centers of the balls we assume the following conditions:

- (*Uniformly distributed*). There exists a constant $L > 0$ such that

$$|\#(I_\varepsilon(Q)) - N_\varepsilon |Q|| \le L \tag{4.2}$$

  for every open square $Q \subset (0,1)^2$.
- (*Well separated*). There exists $\beta < 1$ such that

$$\operatorname{dist}(x_i^\varepsilon, x_j^\varepsilon) > 6\varepsilon^\beta \tag{4.3}$$

  for every $i, j \in I_\varepsilon$, $i \neq j$, and for every $\varepsilon \in (0, \varepsilon_0)$.
- (*Finite capacity density*). There exists a constant $\Lambda \ge 0$ (possibly zero) such that $N_\varepsilon \varepsilon \to \Lambda$.

Remark 13. For brevity we refer to the constant Λ as the *capacity density* of the obstacles. (More correctly $\Lambda \mathrm{Cap}^\nu_{H^{\frac{1}{2}}}(B_R) \sim \Lambda R$ is the capacity density.) Note that in order to get a Γ-convergence result the capacity density does not need to be constant. One could also consider either a case where the obstacles are not uniformly distributed in space or the case where the radii of the obstacles are varying (i.e., $B_\varepsilon^i = B(x_\varepsilon^i, R_\varepsilon^i \varepsilon)$). This would lead in general to a nonconstant capacity density $\Lambda(x)$. In this case, condition (4.2) should be replaced by

$$\left| \sum_{x_\varepsilon^i \in Q} R_\varepsilon^i - \frac{1}{\varepsilon} \int_Q \Lambda(x)\, dx \right| \le L, \tag{4.4}$$

with $\Lambda \in L^\infty$.

THEOREM 14. *Assume that $N_\varepsilon \to +\infty$ and that the balls $B^i_{R\varepsilon}$ are uniformly distributed, are well separated, and have finite capacity density. Then*

(i) *every sequence $\{u_\varepsilon\}$ such that $\sup_\varepsilon F_\varepsilon(u_\varepsilon) < \infty$ is precompact in L^2, and every cluster point is an integer constant;*

(ii) *for every $u \in L^2(\mathbf{R}^2)$ there exists a sequence $\{u_\varepsilon\}$ strongly converging in L^2 to u such that*

$$\lim_{\varepsilon\to 0} F_\varepsilon(u_\varepsilon) = F(u)\,;$$

(iii) *every sequence $\{u_\varepsilon\}$ strongly converging in L^2 to some function u satisfies*

$$\liminf_{\varepsilon\to 0} F_\varepsilon(u_\varepsilon) \geq F(u)\,.$$

Remark 15. As noted in the introduction, by general facts about Γ-convergence (see, e.g., [5, Proposition 6.20]), we can easily include an applied stress S^ε. If S^ε converges to some S strongly in L^2, then the functional

$$F_\varepsilon(u) - \int_{T^2} S^\varepsilon u\, dx$$

Γ-converges to $F(u) - \int_{T^2} S\, u\, dx$.

Proof of (i) *(compactness).* Since $\sup_\varepsilon F_\varepsilon(u_\varepsilon) \leq C < +\infty$, by (1.6) we have

$$[u_\varepsilon]_{H^{\frac12}(T^2)} \leq C N_\varepsilon \varepsilon\,. \tag{4.5}$$

Moreover, $u_\varepsilon = 0$ on $\bigcup_i B^i_{R\varepsilon} = E_\varepsilon$. This obstacle condition is enough to deduce an L^2 estimate via Poincaré's inequality (2.7). Roughly speaking, the idea is the standard fact that the capacity is almost additive on a union of small well-separated sets and the three-dimensional harmonic capacity of a disc is proportional to its radius, i.e.,

$$\mathrm{Cap}(E_\varepsilon \times \{0\}) \approx \sum_i \mathrm{Cap}(B^i_{R\varepsilon} \times \{0\}) \approx N_\varepsilon \varepsilon\,.$$

In order to carry out this argument rigorously we cover the unit square with a lattice of small squares and apply the Poincaré inequality to each of them. The right estimate follows if we choose the side of each square small but big enough to contain at least one obstacle.

Fix $r_\varepsilon = \sqrt{\frac{L+1}{N_\varepsilon}}$. ($L$ is the constant given by (4.2).) With a little abuse of notation we denote by $Q^{r_\varepsilon}_j$ the squares of a lattice on $(0,1)^2$ of size approximatively r_ε. Applying the Poincaré inequality (2.7), scaled to the square $Q^{r_\varepsilon}_j$, we get

$$\int_{Q^{r_\varepsilon}_j} |u_\varepsilon|^2 dx \leq C_0 r_\varepsilon \left(1 + \frac{r_\varepsilon}{\mathrm{Cap}((\{u_\varepsilon = 0\} \cap Q^{r_\varepsilon}_j) \times \{0\})}\right) [u]^2_{H^{\frac12}(Q^{r_\varepsilon}_j)}\,. \tag{4.6}$$

By our choice of r_ε and assumption (4.2) we have

$$1 \leq \#(I_\varepsilon(Q^{r_\varepsilon}_j)) \leq 2L + 1,$$

and thus $\mathrm{Cap}((\{u_\varepsilon = 0\} \cap Q^{r_\varepsilon}_j) \times \{0\}) > CR\varepsilon$. Taking the sum over all j in (4.6), by (4.5), we get

$$\int_{T^2} |u_\varepsilon|^2 dx \leq \sum_j C_0 r_\varepsilon \left(1 + \frac{r_\varepsilon}{CR\varepsilon}\right) [u]^2_{H^{\frac12}(Q^{r_\varepsilon}_j)} \leq C r_\varepsilon \left(1 + \frac{r_\varepsilon}{CR\varepsilon}\right) N_\varepsilon \varepsilon \leq C\,.$$

Thus u_ε is precompact in the weak topology of $H^{\frac12}$ and, by the compact embedding, in the strong topology of L^2.

Finally, let u be a cluster point. Assume for simplicity that the whole sequence u_ε converges to u. In particular, since $\sup_\varepsilon F_\varepsilon(u_\varepsilon) \le C$, we have

$$\lim_{\varepsilon\to 0}\int_{T^2} \operatorname{dist}^2(u_\varepsilon, \mathbf{Z})\,dx = 0\,.$$

This implies that $u \in H^{\frac12}(T^2, \mathbf{Z})$. Then u must be constant. This is obvious in the case when $N_\varepsilon\varepsilon \to 0$ but is true in general for any function in $H^{\frac12}$ taking values in $\mathbf{Z}$. (This fact can be easily checked in one dimension, where jumps are not permitted, and then extended to any dimension by slicing.) □

Proof of (ii) *(the upper bound).* It is clearly enough to prove the result for any constant function $u = a \in \mathbf{Z}$. (Otherwise the upper bound is trivial.) In order to construct a sequence $\{u_\varepsilon\}$ which converges strongly to a in L^2 and satisfies

$$\lim_{\varepsilon\to 0} F_\varepsilon(u_\varepsilon) = D^\nu_{\frac12}(a, B_R),$$

fix $\rho_\varepsilon > 0$ such that $\varepsilon \le \rho_\varepsilon \ll \varepsilon^{(\beta+1)/2}$, and let ζ_ε be a $H^{\frac12}$-dislocation capacitary potential of B_R with respect to $B_{\frac{\rho_\varepsilon}{\varepsilon}}$ at level a. Define

$$u^i_\varepsilon(x) = a - \zeta_\varepsilon\left(\frac{x - x^i_\varepsilon}{\varepsilon}\right)$$

and

$$u_\varepsilon(x) = \begin{cases} \sum_i u^i_\varepsilon(x)\chi_{B^i_{\varepsilon^\beta}}(x) & \text{if } x \in \bigcup_i B^i_{\varepsilon^\beta},\\ a & \text{otherwise.}\end{cases}$$

It is easy to check that the sequence u_ε converges to a in L^2. In order to control the nonlocal term in the energy let us first show that long-range interactions are negligible. Indeed, using the fact that $u^i_\varepsilon = a$ outside $B^i_{\rho_\varepsilon}$ and the properties of the kernel K_ν we have

$$\begin{aligned}\iint_{\substack{T^2\times T^2\\ |x-y|>\varepsilon^\beta}} & K_\nu(x-y)|u_\varepsilon(x) - u_\varepsilon(y)|^2 dx\,dy\\ &\le 2\sum_i \int_{B^i_{\rho_\varepsilon}}\int_{T^2} \chi_{|x-y|>\varepsilon^\beta}\, K_\nu(x-y)\,|u^i_\varepsilon(x) - u_\varepsilon(y)|^2 dx\,dy\\ &\le 2a^2 N_\varepsilon \int_{B_{\rho_\varepsilon}}\int_{T^2} \chi_{|y|>\varepsilon^\beta}\, K_\nu(y)\,dy\,dx \le C\frac{\rho^2_\varepsilon N_\varepsilon}{\varepsilon^\beta}\,.\end{aligned}$$

The constant C depends on a, but since a is fixed we suppress this dependence in the following. Since $(T^2 \setminus \cup_i B^i_{2\varepsilon^\beta}) \times T^2 \subseteq \left[(T^2 \setminus \cup_i B^i_{\varepsilon^\beta}) \times (T^2 \setminus \cup_i B^i_{\varepsilon^\beta})\right] \cup \{|x-y| > \varepsilon^\beta\}$ and $u_\varepsilon(x) = a$ outside $\cup_i B^i_{\varepsilon^\beta}$, by our choice of ρ_ε and (4.3) we have

(4.7)

$$\begin{aligned}F_\varepsilon(u_\varepsilon) &\le \sum_i \frac{1}{N_\varepsilon\varepsilon}\left(\frac1\varepsilon\int_{B^i_{\varepsilon^\beta}} \operatorname{dist}^2(u^i_\varepsilon, \mathbf{Z})\,dx + \iint_{B^i_{3\varepsilon^\beta}\times B^i_{3\varepsilon^\beta}} K_\nu(x-y)\,|u^i_\varepsilon(x) - u^i_\varepsilon(y)|^2 dx\,dy\right) + o(1)\\ &= \frac{1}{\varepsilon^2}\int_{B_{\varepsilon^\beta}} \operatorname{dist}^2\left(\zeta_\varepsilon\left(\frac{x}{\varepsilon}\right), \mathbf{Z}\right) dx + \frac1\varepsilon\iint_{B_{3\varepsilon^\beta}\times B_{3\varepsilon^\beta}} K_\nu(x-y)\left|\zeta_\varepsilon\left(\frac{x}{\varepsilon}\right) - \zeta_\varepsilon\left(\frac{y}{\varepsilon}\right)\right|^2 dx\,dy + o(1)\\ &= \int_{B_{\varepsilon^\beta - 1}} \operatorname{dist}^2(\zeta_\varepsilon(x), \mathbf{Z})\,dx + \iint_{B_{3\varepsilon^\beta-1}\times B_{3\varepsilon^\beta-1}} \varepsilon^3 K_\nu(\varepsilon(x-y))\,|\zeta_\varepsilon(x) - \zeta_\varepsilon(y)|^2\,dx\,dy + o(1).\end{aligned}$$

Now, by Proposition 1, for ε small enough we have

$$|\varepsilon^3 K_\nu(\varepsilon(x-y)) - \Gamma_\nu(x-y)| \le C\varepsilon^3 \qquad \forall\, x\,, y \in B_{3\varepsilon^{\beta-1}}\,,$$

and hence

$$\begin{aligned}(4.8)\quad &\iint_{B_{3\varepsilon^{\beta-1}}\times B_{3\varepsilon^{\beta-1}}} \varepsilon^3 K_\nu(\varepsilon(x-y))\ |\zeta_\varepsilon(x)-\zeta_\varepsilon(y)|^2\,dx\,dy \\ &\qquad \le \iint_{B_{3\varepsilon^{\beta-1}}\times B_{3\varepsilon^{\beta-1}}} \Gamma_\nu(x-y)\ |\zeta_\varepsilon(x)-\zeta_\varepsilon(y)|^2\,dx\,dy + C\varepsilon^3\varepsilon^{4(\beta-1)}\,.\end{aligned}$$

Thus by the definition of ζ_ε we have

$$F_\varepsilon(u_\varepsilon) \le D_\nu^{\frac12}(a, B_R, B_{\frac{\rho_\varepsilon}{\varepsilon}}) + C\varepsilon^{4\beta-1},$$

which for the choice of $\beta > \frac14$ together with Proposition 12 gives

$$\limsup_{\varepsilon\to 0} F_\varepsilon(u_\varepsilon) \le D_\nu^{\frac12}(a, B_R).$$

(Note that if (4.3) holds for some β, it also holds for all larger β.) ∎

The proof of the lower bound is based on the following key lemma.

LEMMA 16. *Given $\mathcal{R} : \mathbf{R}_+ \to \mathbf{R}_+$, with $\mathcal{R}(\varepsilon) \to \infty$ as $\varepsilon \to 0$, there exists a function $\omega : \mathbf{R}_+ \times \mathbf{R}_+ \to \mathbf{R}_+$, with $\omega(\varepsilon,\delta) \to 0$ as $(\varepsilon,\delta) \to (0,0)$, such that the following statement holds. Let $a \in \mathbf{Z}$. If $\zeta \in H^{\frac12}(B_{\mathcal{R}(\varepsilon)})$ satisfies*

$$(4.9)\qquad \fint_{B_{\mathcal{R}(\varepsilon)}} |\zeta - a|dx \le \delta$$

and $\zeta = 0$ on B_R, then

$$\begin{aligned}(4.10)\quad J_\varepsilon(\zeta) &:= \int_{B_{\mathcal{R}(\varepsilon)}} \operatorname{dist}^2(\zeta, \mathbf{Z})\,dx + \iint_{B_{\mathcal{R}(\varepsilon)}\times B_{\mathcal{R}(\varepsilon)}} K^\varepsilon(x-y)|\zeta(x)-\zeta(y)|^2\,dx\,dy \\ &\ge D^\nu_{\frac12}(a, B_R) - \omega(\varepsilon,\delta)\,,\end{aligned}$$

where $K^\varepsilon(t) = \varepsilon^3 K_\nu(\varepsilon t)$.

Proof. Assume for a contradiction that there exist $(\varepsilon_k,\delta_k) \to (0,0)$, $\eta > 0$, and $\zeta_k \in H^{\frac12}(B_{\mathcal{R}(\varepsilon_k)})$, with $\zeta_k = 0$ on B_R such that

$$J_{\varepsilon_k}(\zeta_k) \le D^\nu_{\frac12}(a, B_R) - \eta$$

and

$$(4.11)\qquad \fint_{B_{\mathcal{R}(\varepsilon_k)}} |\zeta_k - a|dx \to 0 \qquad \text{as } k \to \infty\,.$$

Denote $B^k = B_{\mathcal{R}(\varepsilon_k)}$. By the Sobolev embedding there exists a sequence of real numbers a_k such that

$$\left(\int_{B^k} |\zeta_k - a_k|^4 dx\right)^{\frac12} \le C \iint_{B^k\times B^k} K^{\varepsilon_k}(x-y)|\zeta_k(x)-\zeta_k(y)|^2\,dx\,dy \le C\,.$$

Hence by the Hölder inequality we have

$$\fint_{B^k} |\zeta_k - a_k| dx \le \left(\fint_{B^k} |\zeta_k - a_k|^4 dx \right)^{\frac{1}{4}} \to 0,$$

and thus we deduce that $a_k \to a$ as $k \to \infty$. In conclusion there exists a function ζ such that for every $r > 0$ we have that ζ_k converge weakly to ζ in $H^{\frac{1}{2}}(B_r)$ and in $L^4(B_r)$ and strongly in $L^2(B_r)$. Moreover,

$$\int_{B_r} |\zeta - a|^4 dx \le \liminf_{k\to\infty} \int_{B_r} |\zeta_k - a_k|^4 dx \le \liminf_{k\to\infty} \int_{B^k} |\zeta_k - a_k|^4 dx \le C; \tag{4.12}$$

i.e., ζ is a good competitor for the definition of $D^{\nu}_{\frac{1}{2}}(a, B_R)$. In addition we have

$$\lim_{k\to\infty} \int_{B_r} \operatorname{dist}^2(\zeta_k, \mathbf{Z})\, dx = \int_{B_r} \operatorname{dist}^2(\zeta, \mathbf{Z})\, dx$$

and

$$\lim_{k\to\infty} \iint_{B_r\times B_r} \Gamma_\nu(x-y)|\zeta_k(x) - \zeta_k(y)|^2\, dx\, dy = \iint_{B_r\times B_r} \Gamma_\nu(x-y)|\zeta(x) - \zeta(y)|^2\, dx\, dy\,.$$

Finally, by Proposition 1 and the homogeneity of Γ_ν we have

$$|K^\varepsilon(x-y) - \Gamma_\nu(x-y)| \le C_r \frac{\varepsilon^3}{|x-y|^3} \qquad \text{if } |x-y| \le \frac{3}{4\varepsilon},$$

and hence

$$\left| \iint_{B_r\times B_r} \left(K^{\varepsilon_k}(x-y) - \Gamma_\nu(x-y)\right) |\zeta_k(x) - \zeta_k(y)|^2\, dx\, dy \right| \le C_r \varepsilon_k^3 \|\zeta_k\|_{H^{\frac{1}{2}}(B_r)}\,.$$

Thus for every $r > 0$ we get

$$D^{\nu}_{\frac{1}{2}}(a, B_R) - \eta \ge \limsup_{k\to\infty} J_{\varepsilon_k}(\zeta_k) \ge \int_{B_r} \operatorname{dist}^2(\zeta, \mathbf{Z})\, dx + \iint_{B_r\times B_r} \Gamma_\nu(x-y)|\zeta(x) - \zeta(y)|^2\, dx\, dy$$

so that

$$D^{\nu}_{\frac{1}{2}}(a, B_R) - \eta \ge \int_{\mathbf{R}^2} \operatorname{dist}^2(\zeta, \mathbf{Z})\, dx + \iint_{\mathbf{R}^2\times\mathbf{R}^2} \Gamma_\nu(x-y)|\zeta(x) - \zeta(y)|^2\, dx\, dy\,.$$

This is a contradiction, in view of the fact that ζ is a good test function in the definition of $D^{\nu}_{\frac{1}{2}}(a, B_R)$, and the proof is complete. □

A second key point for the proof of the lower bound is to show that if the sequence $u_\varepsilon - a$ is close to zero in some ball B_r, then it is close to zero also at a smaller scale. This is a consequence of the following proposition.

PROPOSITION 17. *There exists a positive constant C such that for every $0 < \rho < r$ the following inequality holds:*

$$\fint_{B_\rho} |u|\, dx \le \fint_{B_r} |u|\, dx + \frac{C}{\sqrt{\rho}} [u]_{H^{\frac{1}{2}}(B_r)} \tag{4.13}$$

for all $u \in H^{\frac{1}{2}}(B_r)$.

Proof. Let us first show that there exists a constant C such that for any $u \in H^{\frac{1}{2}}(B_1)$

$$\fint_{B_\theta} |u|\,dx \le \fint_{B_1} |u|\,dx + C\,[u]_{H^{\frac{1}{2}}(B_1)} \tag{4.14}$$

for every $\theta \in [\frac{1}{2}, 1]$. By the Hölder inequality and the Sobolev embedding there exists a constant c such that

$$\|u-c\|_{L^1(B_\theta)} \le \|u-c\|_{L^1(B_1)} \le C\,[u]_{H^{\frac{1}{2}}(B_1)}\,. \tag{4.15}$$

Moreover, the constant c can be estimated as follows:

$$c = \fint_{B_1} c \le \fint_{B_1} |u|\,dx + \fint_{B_1} |u-c|\,dx \le \fint_{B_1} |u|\,dx + C\,[u]_{H^{\frac{1}{2}}(B_1)}\,,$$

and hence (4.14) follows by (4.15). By a scaling argument we obtain

$$\fint_{B_{\theta r}} |u|\,dx \le \fint_{B_r} |u|\,dx + \frac{C}{\sqrt{r}}[u]_{H^{\frac{1}{2}}(B_r)}$$

for every $r > 0$, $\theta \in [\frac{1}{2}, 1]$, and $u \in H^{\frac{1}{2}}(B_r)$. Finally, any $\rho < r$ can be written as $\rho = \theta 2^{-k} r$ for some $\theta \in (\frac{1}{2}, 1)$ and $k \in \mathbb{N} \cup \{0\}$, so that the conclusion follows by an iteration procedure, with a slightly modified constant C. □

Proof of (iii) *(lower bound).* Let u_ε be a sequence in $H^{\frac{1}{2}}(T^2)$, and assume that u_ε converges to u strongly in L^2. In order to prove the lower bound we may assume that $\liminf_\varepsilon F_\varepsilon(u_\varepsilon) = \lim_\varepsilon F_\varepsilon(u_\varepsilon) < +\infty$. Thus by (i) (compactness) we have that $u = a \in \mathbf{Z}$. Since the energy decreases under truncation we may assume that $0 \le u_\varepsilon \le a$.

Consider a lattice of squares Q_j^ε of size approximatively $1/\sqrt{N_\varepsilon}$. Let $\widehat{Q}_j^\varepsilon$ be concentric squares twice the size. Since each point is contained at most in nine of the squares $\widehat{Q}_j^\varepsilon$, we have

$$\sum_j \iint_{\widehat{Q}_j^\varepsilon \times \widehat{Q}_j^\varepsilon} \frac{|u_\varepsilon(x) - u_\varepsilon(y)|^2}{|x-y|^3}\,dx\,dy \le C N_\varepsilon \varepsilon$$

and

$$\sum_j \int_{\widehat{Q}_j^\varepsilon} |u_\varepsilon - a| dx \le \omega_\varepsilon\,,$$

where $\omega_\varepsilon \to 0$ as $\varepsilon \to 0$. Let $\theta > 0$. Then there exist a set of indices J^ε such that $\frac{1}{N_\varepsilon}\#(J^\varepsilon) \ge 1 - \theta$ and a constant C_θ such that

$$\iint_{\widehat{Q}_j^\varepsilon \times \widehat{Q}_j^\varepsilon} \frac{|u_\varepsilon(x) - u_\varepsilon(y)|^2}{|x-y|^3}\,dx\,dy \le C_\theta \varepsilon \tag{4.16}$$

and

$$\fint_{\widehat{Q}_j^\varepsilon} |u_\varepsilon - a| dx \le C_\theta \omega_\varepsilon \tag{4.17}$$

for all $j \in J^\varepsilon$. Let $0 < \delta < 1$. By applying Proposition 17 with $\rho = \varepsilon^\beta$, with $\frac{1}{2} < \beta < 1$, and $r = \frac{1}{\sqrt{N_\varepsilon}}$, for each $x_\varepsilon^i \in Q_j^\varepsilon$ we also have

$$\fint_{B^i_{\varepsilon^\beta}} |u_\varepsilon - a|^2 dx \le \delta \qquad \text{if } \varepsilon \le \varepsilon_0(\delta, \theta)\,. \tag{4.18}$$

Then by Lemma 16 and the assumption that $\operatorname{dist}(x_i^\varepsilon, x_j^\varepsilon) > 6\varepsilon^\beta$

$$F_\varepsilon(u_\varepsilon) \ge \frac{1}{N_\varepsilon}\left[\sum_{j\in J^\varepsilon} \#(I_\varepsilon(Q_j^\varepsilon))\right]\left(D_\nu^{\frac{1}{2}}(a, B_R) - \omega(\varepsilon,\delta)\right). \tag{4.19}$$

The uniform distribution of the obstacles (see condition (4.2)) implies that

$$\begin{aligned}\sum_{j\in J^\varepsilon} \#(I_\varepsilon(Q_j^\varepsilon)) &= N_\varepsilon - \sum_{j\notin J^\varepsilon} \#(I_\varepsilon(Q_j^\varepsilon)) \ge N_\varepsilon - \sum_{j\notin J^\varepsilon} (N_\varepsilon |Q_j^\varepsilon| + L)\\ &= N_\varepsilon - (L+1)\#(\{j:\ j\notin J^\varepsilon\})\,.\end{aligned}$$

Since $\#(\{j:\ j\notin J^\varepsilon\}) \le N_\varepsilon\theta$, we get

$$F_\varepsilon(u_\varepsilon) \ge (1-\theta(L+1))\left(D_\nu^{\frac{1}{2}}(a, B_R) - \omega(\varepsilon,\delta)\right),$$

and this yields the required lower bound taking the limit as $\varepsilon \to 0$; then $\delta \to 0$, and, finally, $\theta \to 0$. □

Appendix. Finite pinning condition. We can model the hardening mechanism due to obstacles such as secondary dislocations by assuming a weaker pinning condition given by a concentrated force. Namely, we assume that a crossing of an obstacle by a dislocation costs a finite amount of energy, i.e.,

$$\lambda_0 \int_{B^i_{R\varepsilon}} \varepsilon\psi_\varepsilon^i |u|\, dx\,,$$

where $\psi_\varepsilon^i(x) = \varepsilon^{-2}\psi(\frac{x-x_\varepsilon^i}{\varepsilon})$, with $\operatorname{supp}\psi \subseteq B_R(0)$ and $\int_{B_R}\psi\, dx = 1$, and $\lambda_0\varepsilon\psi_\varepsilon^i$ is the force concentrated on each obstacle B_ε^i. Then we can consider the following functional:

$$\widetilde{F}_\varepsilon(u) = \begin{cases} \dfrac{1}{\varepsilon}\displaystyle\int_{T^2} \operatorname{dist}^2(u, \mathbf{Z})\, dx \;+\; \iint_{T^2\times T^2} K_\nu(x-y)|u(x)-u(y)|^2\, dx\, dy \\ \qquad\qquad + \displaystyle\sum_i \lambda_0 \int_{B^i_{R\varepsilon}} \varepsilon\psi_\varepsilon^i |u|\, dx \\ \qquad\qquad\qquad\qquad \text{if } u \in H^{\frac{1}{2}}(T^2)\,, \\ +\infty \qquad\qquad\qquad\qquad \text{otherwise}\,. \end{cases} \tag{A.1}$$

With our scaling assumptions the total force due to the obstacles is finite and is given by

$$\sum_i \lambda_0 \int_{B^i_{R\varepsilon}} \varepsilon\psi_\varepsilon^i\, dx = N_\varepsilon\varepsilon\lambda_0 \approx \Lambda\lambda_0\,.$$

In order to study the Γ-limit of the functional $\widetilde{F}_\varepsilon$, another natural notion of capacity has to be defined, i.e.,

(A.2)
$$\widetilde{D}^\nu_{\frac12}(a,\lambda_0,\psi) := \inf\Bigg\{ \int_{\mathbf{R}^2} \operatorname{dist}^2(\zeta,\mathbf{Z})\,dx + \iint_{\mathbf{R}^2\times\mathbf{R}^2} \Gamma_\nu(x-y)|\zeta(x)-\zeta(y)|^2\,dx\,dy + \\ +\lambda_0 \int_{\mathbf{R}^2} \psi|\zeta|\,dx : \quad \zeta - a \in L^4(\mathbf{R}^2) \Bigg\}.$$

Again one can check that this is a Choquet capacity and satisfies

$$\widetilde{D}^\nu_{\frac12}(a,\lambda_0,\psi) \le D^\nu_{\frac12}(a,B_R)\,.$$

Moreover,

$$\lim_{\lambda_0\to\infty} \widetilde{D}^\nu_{\frac12}(a,\lambda_0,\psi) = D^\nu_{\frac12}(a,B_R)\,. \tag{A.3}$$

Indeed, $\widetilde{D}^\nu_{\frac12}(a,\lambda_0,\psi)$ is increasing in λ_0; thus the limit always exists. If $\widetilde{\zeta}_{\lambda_0}$ is a sequence of potentials for $\widetilde{D}^\nu_{\frac12}(a,\lambda_0,\psi)$, one can check that, up to a subsequence, it converges weakly in $H^{\frac12}$ to a function $\widetilde{\zeta}$ which is a good competitor for $D^\nu_{\frac12}(a,B_R)$.

Using this notion of capacity we can perform the same analysis as above and prove the following theorem.

THEOREM 18. *Assume that $N_\varepsilon \to +\infty$ and that the balls $B^i_{R_\varepsilon}$ are uniformly distributed, are well separated, and have finite capacity density. Denote by $\widetilde{F}$ the functional*

$$\widetilde{F}(u) = \begin{cases} \widetilde{D}^\nu_{\frac12}(u,\lambda_0\psi) & \text{if } u = \text{const.} \in \mathbf{Z}\,, \\ +\infty & \text{otherwise}\,. \end{cases} \tag{A.4}$$

Then

(i) *every sequence $\{u_\varepsilon\}$ such that $\sup_\varepsilon \widetilde{F}_\varepsilon(u_\varepsilon) < \infty$ is precompact in L^2, and every cluster point is an integer constant;*

(ii) *for every $u \in L^2(\mathbf{R}^2)$ there exists a sequence $\{u_\varepsilon\}$ strongly converging in L^2 to u such that*
$$\lim_{\varepsilon\to0} \widetilde{F}_\varepsilon(u_\varepsilon) = \widetilde{F}(u)\,;$$

(iii) *every sequence $\{u_\varepsilon\}$ strongly converging in L^2 to some function u satisfies*
$$\liminf_{\varepsilon\to0} \widetilde{F}_\varepsilon(u_\varepsilon) \ge \widetilde{F}(u)\,.$$

Remark 19. The new dislocation capacity for the weak pinning condition is linear for a big enough. In order to see this we can rewrite it for positive a as follows:

(A.5)
$$\widetilde{D}(a,\lambda_0,\psi) = \inf\Bigg\{ \int_{\mathbf{R}^2} \operatorname{dist}^2(w,\mathbf{Z})\,dx + \iint_{\mathbf{R}^2\times\mathbf{R}^2} \Gamma_\nu(x-y)|w(x)-w(y)|^2\,dx\,dy \\ +\ \lambda_0 \int_{\mathbf{R}^2} \psi(w+a)\,dx : \quad w \in L^4(\mathbf{R}^2) \text{ and } \quad -a \le w \le 0 \Bigg\}.$$

Then consider the following minimum problem:

(A.6)
$$D_0(\lambda_0,\psi) := \inf\Bigg\{\int_{\mathbf{R}^2} \operatorname{dist}^2(w,\mathbf{Z})\,dx + \iint_{\mathbf{R}^2\times\mathbf{R}^2} \Gamma_\nu(x-y)|w(x)-w(y)|^2\,dx\,dy + \lambda_0\int_{\mathbf{R}^2}\psi w\,dx : \quad w\in L^4(\mathbf{R}^2)\Bigg\},$$

and let w_0 be a minimum point. As in the proof of Proposition 8 one can prove that there exists an L^4 function f_0 such that

$$Lw_0 = f_0 - \frac{\lambda_0}{2}\psi$$

in the sense of distributions. Using the Green function G_ν of $L+I$ we show that w_0 is bounded. In fact,

$$w_0(x) = G_\nu * f_0(x) + G_\nu * w_0(x) - \frac{\lambda_0}{2}G_\nu * \psi(x),$$

and the conclusion follows by the Hölder inequality. Now let a_0 the smallest positive integer such that

$$w_0 \geq -a_0\,.$$

Clearly we have

$$\widetilde{D}(a,\lambda_0,\psi) = D_0(\lambda_0,\psi) + \lambda_0 a \qquad \forall a \geq a_0\,.$$

In particular, if we minimize our energy subject to an external force S, i.e.,

$$\min_{a\in\mathbf{Z}}\ \widetilde{F}(a) - a\int_{T^2} S\,dx\,,$$

then the minimum exists if and only if $|\int_{T^2} S\,dx| \leq \lambda_0$. If the force S is greater than the total resistance of the obstacles, then no equilibrium states exist.

Acknowledgment. We would like to thank M. Ortiz for bringing this problem to our attention and for many stimulating discussions.

REFERENCES

[1] D. R. Adams and L. I. Hedberg, *Function Spaces and Potential Theory*, Grundlehren Math. Wiss. 314, Springer-Verlag, Berlin, 1996.

[2] G. Alberti, G. Bouchitté, and P. Seppecher, *Phase transition with the line-tension effect*, Arch. Rational Mech. Anal., 144 (1998), pp. 1–46.

[3] A. I. Baernstein, *A unified approach to symmetrization*, in Partial Differential Equations of Elliptic Type (Cortona, 1992), Sympos. Math. XXXV, Cambridge University Press, Cambridge, UK, 1994, pp. 47–91.

[4] D. Cioranescu and F. Murat, *Un terme étrange venu d'ailleurs*, in Nonlinear Partial Differential Equations and Their Applications, College de France Seminar, Vol. II (Paris, 1979/1980), Res. Notes Math. 60, Pitman, Boston, London, 1982, pp. 98–138, 389–390. English translation in *Topics in the Mathematical Modelling of Composite Materials*, A. Cherkaev and R.V. Kohn, eds., Birkhäuser Boston, Boston, MA, 1997, pp. 45–94.

[5] G. Dal Maso, *An Introduction to* Γ*-Convergence*, Progr. Nonlinear Differential Equations Appl. 8, Birkhäuser Boston, Boston, MA, 1993.
[6] G. Dal Maso and U. Mosco, *Wiener's criterion and* γ*-convergence*, Appl. Math. Optim., 15 (1987), pp. 15–63.
[7] J. Frehse, *Capacity methods in the theory of partial differential equations*, Jahresber. Deutsch. Math.-Verein., 84 (1982), pp. 1–44.
[8] A. Garroni and S. Müller, *A Variational Model for Dislocation in the Line Tension Limit*, preprint, Max-Planck Institute, Leipzig, Germany, 2004.
[9] A. M. Garsia and E. Rodemich, *Monotonicity of certain functionals under rearrangement*, Ann. Inst. Fourier (Grenoble), 24 (1974), pp. 67–116.
[10] M. Koslowski, A. M. Cuitino, and M. Ortiz, *A phase-field theory of dislocation dynamics, strain hardening and hysteresis in ductile single crystal*, J. Mech. Phys. Solids, 50 (2002), pp. 2597–2635.
[11] N. S. Landkov, *Foundations of Modern Potential Theory*, Springer-Verlag, New York, Heidelberg, 1972.
[12] A. Marchenko and E. Y. Kruslov, *New results of boundary value problems for regions with closed-grained boundaries*, Uspekhi Mat. Nauk, 33 (1978).
[13] L. Modica and S. Mortola, *Il limite nella* γ*-convergenza di una famiglia di funzionali ellittici*, Boll. Un. Mat. Ital. A (5), 14 (1977), pp. 526–529.
[14] L. Modica and S. Mortola, *Un esempio di* γ^-*-convergenza*, Boll. Un. Mat. Ital. B (5), 14 (1977), pp. 285–299.
[15] E. M. Stein and G. Weiss, *Introduction to Fourier Analysis on Euclidean Spaces*, Princeton Math. Ser. 32, Princeton University Press, Princeton, NJ, 1971.
[16] W. Ziemer, *Weakly Differentiable Functions*, Springer-Verlag, New York, 1989.

SIAM J. MATH. ANAL.
Vol. 36, No. 6, pp. 1965–1990

QUASI-PERIODIC SOLUTIONS FOR 1D SCHRÖDINGER EQUATIONS WITH HIGHER ORDER NONLINEARITY*

ZHENGUO LIANG† AND JIANGONG YOU†

Abstract. In this paper, one-dimensional (1D) nonlinear Schrödinger equations

$$iu_t - u_{xx} + mu + \nu|u|^4u = 0,$$

with Dirichlet boundary conditions are considered. It is proved that for all real parameters m, the above equation admits small-amplitude quasi-periodic solutions corresponding to b-dimensional invariant tori of an associated infinite-dimensional dynamical system. The proof is based on infinite-dimensional KAM theory, partial normal form, and scaling skills.

Key words. quasi-periodic solutions, infinite-dimensional KAM theory, partial normal form

AMS subject classifications. 37K55, 35Q55

DOI. 10.1137/S0036141003435011

1. Introduction and main result. In this paper, we will prove that one-dimensional (1D) nonlinear Schrödinger equation

$$iu_t - u_{xx} + mu + \nu|u|^4u = 0 \tag{1.1}$$

subject to Dirichlet boundary conditions

$$u(0,t) = u(\pi,t) = 0, \tag{1.2}$$

admits small-amplitude quasi-periodic solutions for all m. Equation (1.1) with $m = 0$ and negative ν is called "focusing" while (1.1) with $m = 0$ and positive ν is called "defocusing." Under some initial-boundary conditions they have been considered by many authors (see [2, 3, 4, 11]). Throughout this paper, we suppose $\nu > 0$ in (1.1). As we will see later, the sign of ν is immaterial for our results.

We study the equation (1.1) as a Hamiltonian system on $\mathcal{P} = W_0^1([0,\pi])$, the Sobolev space of all complex valued L^2-functions on $[0,\pi]$ with an L^2-derivative and vanishing boundary values. Let

$$\phi_j(x) = \sqrt{\frac{2}{\pi}}\sin jx, \lambda_j = j^2 + m, j \geq 1$$

be the basic modes and their frequencies for the linear equation $iu_t = u_{xx} - mu$ with Dirichlet boundary conditions. Then every solution is the superposition of oscillations of the basic modes, with the coefficients moving on circles,

$$u(t,x) = \sum_{j\geq 1} q_j(t)\phi_j(x), q_j(t) = q_j^0 e^{i\lambda_j t}.$$

*Received by the editors September 20, 2003; accepted for publication (in revised form) July 2, 2004; published electronically June 30, 2005. The work was supported by the National Natural Science Foundation of China (19925107) and the Special Funds for Major State Basic Research Projects of China (973 projects).

http://www.siam.org/journals/sima/36-6/43501.html

†Department of Mathematics, Nanjing University, Nanjing 210093, People's Republic of China (jyou@nju.edu.cn).

Together they move on a rotational torus of finite or infinite dimension, depending on how many modes are excited. In particular, for every choice

$$J = \{j_1 < j_2 < \cdots < j_b\} \subset \mathbb{N}$$

of b basic modes there is an invariant linear space E_J of complex dimension b which is completely foliated into rotational tori,

$$E_J = \{u = q_1\phi_{j_1} + \cdots + q_b\phi_{j_b} : q \in C^b\} = \bigcup_{I \in \overline{P^b}} \mathcal{T}_J(I),$$

where $P^b = \{I : I_j > 0 \text{ for } 1 \le j \le b\}$ and

$$\mathcal{T}_J(I) = \{u = q_1\phi_{j_1} + \cdots + q_b\phi_{j_b} : |q_j|^2 = 2I_j \text{ for} 1 \le j \le b\}.$$

In addition, each such torus is linearly stable and all solutions have vanishing Lyapunov exponents. This is the linear situation.

Upon restoration of the nonlinearity $\nu|u|^4u$, we show that there exists a Cantor set $\mathcal{C} \subset \mathcal{P}^b$, a specially chosen index set $\mathcal{I} = n_1 < n_2 < \cdots < n_b \subset \mathbb{N}$ (we will call it an admissible set, for more specific see section 3) and a family of b-tori

$$\mathcal{T}_{\mathcal{I}}[\mathcal{C}] = \cup_{I \in \mathcal{C}} \mathcal{T}_{\mathcal{I}}(I) \subset E_{\mathcal{I}}$$

over $\mathcal{C}$, and a Whitney smooth embedding

$$\Phi : \mathcal{T}_{\mathcal{I}}[\mathcal{C}] \hookrightarrow \mathcal{P},$$

such that the restriction of Φ to each $\mathcal{T}_{\mathcal{I}}(I)$ in the family is an embedding of a rotational b-torus for the nonlinear equation. In [10], The image $\mathcal{E}_{\mathcal{I}}$ of $\mathcal{T}_{\mathcal{I}}[\mathcal{C}]$ is called a Cantor manifold of rotational b-tori given by the embedding $\Phi : \mathcal{T}_{\mathcal{I}}[\mathcal{C}] \to \mathcal{E}_{\mathcal{I}}$.

THEOREM 1 (main theorem). *Consider the* 1*D nonlinear Schrödinger equation* (1.1) *with* (1.2). *Then for any admissible index set* $\mathcal{I} = \{n_1 < n_2 < \cdots < n_b\} \subset \mathbb{N}$ *and* $m \in R$, *there exists a positive-measure Cantor manifold* $\mathcal{E}_{\mathcal{I}}$ *of real analytic, linearly stable, Diophantine b-tori for the nonlinear Schrödinger equation given by a Whitney smooth embedding* $\Phi : \mathcal{T}_{\mathcal{I}}[\mathcal{C}] \to \mathcal{E}_{\mathcal{I}}$, *which is a higher order perturbation of the inclusion map* $\Phi_0 : E_{\mathcal{I}} \hookrightarrow \mathcal{P}$ *restricted to* $\mathcal{T}_{\mathcal{I}}[\mathcal{C}]$.

Remark 1. The existence of admissible sets will be proved in the appendix. In fact there exist infinite admissible index sets $\mathcal{I}$.

Remark 2. The result remains true for more general nonlinearities $f(|u|^2)u$, where $f(0) = f'(0) = 0, f''(0) \neq 0$. Our method essentially applies to the nonlinearities $f(|u|^2)u$, where $f(0) = f^{(1)}(0) = \cdots = f^{(k-1)}(0), f^{(k)}(0) \neq 0, k \ge 1$, but the proof would be much more complicated.

Remark 3. The frequencies of the diophantine tori are also under control. They are $\omega(I) = (\lambda_{n_1}, \lambda_{n_2}, \ldots, \lambda_{n_b}) + \frac{1}{\pi^2}(10I_1^2 + 18I_2^2 + \cdots + 18I_b^2 + 36I_1(I_2 + \cdots + I_b) + 48(I_2I_3 + \cdots + I_{b-1}I_b), \ldots, 18I_1^2 + \cdots + 18I_{b-1}^2 + 10I_b^2 + 36I_b(I_1 + \cdots + I_{b-1}) + 48(I_1I_2 + \cdots + I_{b-2}I_{b-1})) + O(||I||^{\frac{13}{6}})$.

Remark 4. The technique of this paper is not restricted to the nonlinear Schrodinger equation. It applies equally well to the nonlinear 1D beam equations

$$u_{tt} + u_{xxxx} = f(u)$$

with hinged boundary conditions, where f is a real analytic, odd function of u of the form $f(u) = au^3 + \sum_{k \ge 5} f_k u^k,\ a \neq 0$. Our result is an improvement on [6]. Details

will be given in another paper. Unfortunately, our technique can't be applied to the complete resonant 1D wave equation

$$u_{tt} + u_{xx} = u^3 \tag{1.3}$$

with Dirichlet boundary conditions. From the proof, one sees that that superlinear growth of the eigenvalues $\lambda_j \sim j^2$ is crucial. For (1.3), the admissible set does not exist and one can't obtain the desired partial Birkhoff normal form by eliminating all the unpleasant terms, which include 2 or 3 tangential coordinates.

The rest of the paper is organized as follows. In section 2 the Hamiltonian function is written in infinitely many coordinates, which is then put into partial normal form in section 3. In section 4 we improve an infinite dimensional KAM theorem, which is developed by many people (see Kuksin [7, 8, 9], Wayne [16], Pöschel [13], Chierchia and You [5]). Measure estimates are given in section 5. Some propositions are proved in the appendix.

2. The Hamiltonian. For simplicity, we choose $\nu = 1$. Other cases can be rescaled into this case. The Hamiltonian of the nonlinear Schrödinger equation is

$$H = \frac{1}{2}\langle Au, u\rangle + \frac{1}{6}\int_0^\pi |u|^6 dx, \tag{2.1}$$

where $A = -d^2/dx^2 + m$. We rewrite H as a Hamiltonian in infinitely many coordinates by making the ansatz

$$u(x) = \sum_{j\geq 1} q_j \phi_j, \quad \phi_j = \sqrt{\frac{2}{\pi}}\sin jx, \quad j \geq 1.$$

The coordinates are taken from the Hilbert spaces $\mathcal{H}^{a,\rho}$ of all complex-valued sequences $q = (q_1, q_2, \ldots)$ with

$$\|q\|^2_{a,\rho} = \sum_{j\geq 1} |q_j|^2 j^{2a} e^{2j\rho} < \infty.$$

Fix $\rho > 0$ and $a \geq 0$ later. One then obtains the Hamiltonian

$$H = \Lambda + G = \frac{1}{2}\sum_{j\geq 1} \lambda_j |q_j|^2 + \frac{1}{6}\int_0^\pi |u|^6 dx \tag{2.2}$$

on the phase space $\mathcal{H}^{a,\rho}$ with symplectic structure $\frac{i}{2}\sum_j dq_j \wedge d\bar{q}_j$. Its equations of motion are

$$\dot{q}_j = 2i\frac{\partial H}{\partial \bar{q}_j}, \quad j \geq 1. \tag{2.3}$$

They are the classical Hamiltonian equations of motion for the real and imaginary parts of $q_j = x_j + iy_j$ written in complex notation. Rather than discussing the above formal validity, we shall, following [10] or [5], use the following elementary observation.

LEMMA 1. *Let I be an interval and let*

$$t \in I \to q(t) \equiv (\{q_j(t)\}_{j\geq 1})$$

be an analytic solution of (2.3) *such that*

$$\sup_{t\in I}\sum_{j\geq 1}|q_j(t)|^2 j^{2a}e^{2j\rho}<\infty \tag{2.4}$$

for some $\rho>0$ *and* $a\geq 0$. *Then*

$$u(t,x)\equiv\sum_{j\geq 1}q_j(t)\phi_j(x),$$

is an analytic solution of (1.1).

For the proof, refer to Lemma 1 in [10].

Next, we consider the regularity of the gradient of G. To this end, let $\mathcal{H}_b^2$ and L^2, respectively, be the Hilbert spaces of all bi-infinite, square summable sequences with complex coefficients and all square-integrable complex valued functions on $[-\pi,\pi]$. Let

$$\mathcal{F}:\mathcal{H}_b^2\to L^2,\quad q\mapsto \mathcal{F}q=\frac{1}{\sqrt{2\pi}}\sum_j q_j e^{\mathrm{i}jx}$$

be the inverse discrete Fourier transform, which defines an isometry between the two spaces. The subspaces $\mathcal{H}_b^{a,\rho}\subset\mathcal{H}_b^2$ consist, by definition, of all bi-infinite sequences with finite norm

$$\|q\|_{a,\rho}^2=|q_0|^2+\sum_j|q_j|^2|j|^{2a}e^{2|j|\rho}.$$

Through $\mathcal{F}$ they define subspaces $W^{a,\rho}\subset L^2$ that are normed by setting $\|\mathcal{F}q\|_{a,\rho}=\|q\|_{a,\rho}$.

LEMMA 2. *For* $a>\frac{1}{2}$ *and* $\rho\geq 0$, *the space* $\mathcal{H}_b^{a,\rho}$ *is a Hilbert algebra with respect to convolution of sequences, and*

$$\|q*p\|_{a,\rho}\leq c\|q\|_{a,\rho}\|p\|_{a,\rho}$$

with a constant c *depending only on* a. *Consequently,* $W^{a,\rho}$ *is a Hilbert algebra with respect to a multiplication of functions.*

For the proof, see [10].

LEMMA 3. *For* $a>\frac{1}{2}$ *and* $\rho\geq 0$, *the gradient* G_q *is real analytic as a map from some neighborhood of the origin in* $\mathcal{H}^{a,\rho}$ *into* $\mathcal{H}^{a,\rho}$, *with*

$$\|G_q\|_{a,\rho}=O(\|q\|_{a,\rho}^5).$$

The proof is similar with Lemma 3 in [10], which we omit.

By the elementary computation, one can get

$$\begin{aligned}G&=\frac{1}{6}\int_0^\pi|u(x)|^6\,dx\\&=\frac{1}{6}\sum_{i,j,k,l,m,n}G_{ijklmn}q_iq_jq_k\bar{q}_l\bar{q}_m\bar{q}_n\end{aligned}$$

with

$$G_{ijklmn}=\int_0^\pi\phi_i\phi_j\phi_k\phi_l\phi_m\phi_n\,dx. \tag{2.5}$$

It is not difficult to verify that $G_{ijklmn} = 0$ unless $i \pm j \pm k \pm l \pm m \pm n = 0$, for some combination of plus and minus signs. For simplicity, we denote $G_{ijk} = G_{iijjkk}, G_i = G_{iiiiii}$. If we choose $n_1, n_2, \ldots, n_b$ satisfying

$$n_i \neq n_j + n_k, \ \forall i, j, k \in \{1, 2, \ldots, b\}, \tag{2.6}$$

one can get

$$G_{n_1} = \cdots = G_{n_b} = \frac{5}{2\pi^2}, \quad G_{n_i n_j n_j} = \frac{3}{2\pi^2}, \quad G_{n_i n_j n_k} = \frac{1}{\pi^2}$$

and

$$G_{n_i n_j l} = \frac{1}{4\pi^2}(4 - \delta^{n_j}_{n_i+l} - \delta^{n_i}_{n_j+l} - \delta^{l}_{n_i+n_j}), \quad G_{n_i n_i l} = \frac{1}{4\pi^2}(6 - \delta^{2n_i}_{l}),$$

where $i \neq j, \ j \neq k, \ k \neq i, \ i, j, k \in \{1, 2, \ldots, b\}, \ l \notin \{n_1, n_2, \ldots, n_b\}$, and for $v \in \mathbb{Z}$

$$\delta^v_i = \begin{cases} 1, & i = v, \\ 0, & \text{otherwise.} \end{cases}$$

3. Partial Birkhoff normal form. We shall use the KAM iteration to get the desired result. Since the quadratic part of the Hamiltonian

$$H = \Lambda + G = \frac{1}{2}\sum_{j\geq 1} \lambda_j |q_j|^2 + \frac{1}{6} \sum_{i\pm j\pm k\pm l\pm m\pm n=0} G_{ijklmn} q_i q_j q_k \bar{q}_l \bar{q}_m \bar{q}_n$$

does not provide any "twist" required by KAM theory, we shall use the normal form technique to get the "twisted" integrable terms from the sixth order terms. To get finite dimensional KAM tori, we shall first fix finite many sites $\{n_1, n_2, \ldots, n_b\}$, and call $q = (q_{n_1}, \ldots, q_{n_b})$ tangential variables. All the other variables, denoted by w, are called normal variables. For our purpose, the sixth order terms with at most two normal variables

$$q_i q_j q_k q_l \bar{q}_m \bar{q}_n, \qquad q_i q_j q_k \bar{q}_l q_m w_n, \qquad q_i q_j q_k q_l w_m w_n$$

must be put into normal form, i.e., the terms that remain after normal form procedure must have the form of $|q_i|^2|q_j|^2|q_k|^2$ or $|q_i|^2|q_j|^2|w_k|^2$. The other sixth order terms are left since they can be scaled into higher perturbations. Such kind of normal form is called a partial Birkhoff normal form since we don't normalize all sixth order terms. In order to get the desired partial Birkhoff normal form, we have to carefully choose $\{n_1, n_2, \ldots, n_b\}$.

For fixed $\{n_1, n_2, \ldots, n_b\}$, we define the index sets $\Delta_*, * = 0, 1, 2$ and Δ_3 in the following way: Δ_* is the set of index (i, j, k, l, m, n) such that there exist right $*$ components not in $\{n_1, n_2, \ldots, n_b\}$. Δ_3 is the set of the index (i, j, k, l, m, n) such that there exist at least three components not in $\{n_1, n_2, \ldots, n_b\}$. Define the resonance sets $\mathcal{N} = \{(i, j, k, i, j, k)\} \cap \Delta_2$ and $\mathcal{M} = \{(i, j, k, i, j, k)\} \cap \Delta_2$. For our convenience, rewrite $G = G^0 + G^1 + G^2 + \hat{G}$, where

$$G^* = \frac{1}{6} \sum_{i\pm j\pm k\pm l\pm m\pm n=0, (i,j,k,l,m,n)\in\Delta_*} G_{ijklmn} q_i q_j q_k \bar{q}_l \bar{q}_m \bar{q}_n, \tag{3.1}$$

and $* = 0, 1, 2$.

DEFINITION 1. *The index set* $\mathcal{I} = \{n_1 < n_2 < \cdots < n_b\}$ *is said to be admissible if and only if* $n_1, n_2, \ldots, n_b$ *satisfy the following Assumptions* A, B, C *and* (2.6).

A. If $i \pm j \pm k \pm l \pm m \pm n = 0, (i, j, k, l, m, n) \in \Delta_0 \setminus \mathcal{N}$, *then*

$$\lambda_i + \lambda_j + \lambda_k - \lambda_l - \lambda_m - \lambda_n \neq 0.$$

B. If $i \pm j \pm k \pm l \pm m \pm n = 0, (i, j, k, l, m, n) \in \Delta_1$, *then*

$$\lambda_i + \lambda_j + \lambda_k - \lambda_l - \lambda_m - \lambda_n \neq 0.$$

C. If $i \pm j \pm k \pm l \pm m \pm n = 0, (i, j, k, l, m, n) \in \Delta_2 \setminus \mathcal{M}$, *then*

$$\lambda_i + \lambda_j + \lambda_k - \lambda_l - \lambda_m - \lambda_n \neq 0.$$

PROPOSITION 1. *There exist infinite many admissible index sets.*

When $b = 2$, we can construct some of the admissible index sets clearly. Denote

$$\mathcal{S} = \{n_1 \leq n_2 | n_1 \equiv 5 \text{ or } 9 (\mathrm{mod} 14),\ n_2 \equiv 8 (\mathrm{mod} 14), n_2 \geq 11 n_1^2\}.$$

PROPOSITION 2. *If* $b = 2$, *any element in* $\mathcal{S}$ *is the admissible index set.*

The proofs of Propositions 1 and 2 are given in the appendix.

Next we transform the Hamiltonian (2.2) into the partial Birkhoff form of order six so that the infinite KAM Theorem (see section 4) can be applied.

LEMMA 4. *For any given admissible index set* $\{n_1 < n_2 < \cdots < n_b\}$, *there exists a real analytic, symplectic change of coordinates* X_F^1 *in some neighborhood of the origin that takes the Hamiltonian* $H = \Lambda + G$ *into*

$$H \circ X_F^1 = \Lambda + \overline{G} + \hat{G} + K,$$

where $X_{\overline{G}}, X_{\hat{G}}$ *and* X_K *are real analytic vector fields in a neighborhood of the origin in* $\mathcal{H}^{a,\rho}$,

$$\begin{aligned}
\overline{G} = {} & \frac{5}{12\pi^2}(|q_{n_1}|^6 + \cdots + |q_{n_b}|^6) \\
& + \frac{9}{4\pi^2}(|q_{n_1}|^4 |q_{n_2}|^2 + \cdots + |q_{n_1}|^4 |q_{n_b}|^2 \\
& + |q_{n_2}|^4 |q_{n_1}|^2 + |q_{n_2}|^4 |q_{n_3}|^2 + \cdots + |q_{n_2}|^4 |q_{n_b}|^2 + \cdots + |q_{n_b}|^4 |q_{n_{b-1}}|^2) \\
& + \frac{6}{\pi^2}(|q_{n_1}|^2 |q_{n_2}|^2 |q_{n_3}|^2 + \cdots + |q_{n_{b-2}}|^2 |q_{n_{b-1}}|^2 |q_{n_b}|^2) \\
& + \frac{3}{2}\left(\sum_{i \neq n_1, n_2, \ldots, n_b} G_{n_1 n_1 i} |q_{n_1}|^4 |q_i|^2 + \cdots + \sum_{i \neq n_1, n_2, \ldots, n_b} G_{n_b n_b i} |q_{n_b}|^4 |q_i|^2\right) \\
& + 6\left(\sum_{i \neq n_1, n_2, \ldots, n_b} G_{n_1 n_2 i} |q_{n_1}|^2 |q_{n_2}|^2 |q_i|^2 + \ldots + \sum_{i \neq n_1, n_2, \ldots, n_b} G_{n_{b-1} n_b i} |q_{n_{b-1}}|^2 |q_{n_b}|^2 |q_i|^2\right),
\end{aligned}$$

$|K| = O(||q||_{a,\rho}^8)$ *and* $\hat{G} = \frac{1}{6} \sum_{i \pm j \pm k \pm l \pm m \pm n = 0, (i,j,k,l,m,n) \in \Delta_3} G_{ijklmn} q_i q_j q_k \bar{q}_l \bar{q}_m \bar{q}_n$.

Proof. Let $\Gamma = X_F^t|_{t=1}$ be the time 1-map of the flow of the Hamiltonian vector field X_F given by the Hamiltonian

$$\begin{aligned} F &= F^0 + F^1 + F^2 \\ &= \frac{1}{6}\Bigg\{ \sum_{i,j,k,l,m,n} F^0_{ijklmn} q_i q_j q_k \bar{q}_l \bar{q}_m \bar{q}_n \\ &\quad + \sum_{i,j,k,l,m,n} F^1_{ijklmn} q_i q_j q_k \bar{q}_l \bar{q}_m \bar{q}_n \\ &\quad + \sum_{i,j,k,l,m,n} F^2_{ijklmn} q_i q_j q_k \bar{q}_l \bar{q}_m \bar{q}_n \Bigg\} \end{aligned}$$

with coefficients

$$\mathrm{i}F^0_{ijklmn} = \begin{cases} \frac{G_{ijklmn}}{\lambda_i+\lambda_j+\lambda_k-\lambda_l-\lambda_m-\lambda_n} & i \pm j \pm k \pm l \pm m \pm n = 0, (i,j,k,l,m,n) \in \Delta_0 \setminus \mathcal{N}, \\ 0 & otherwise, \end{cases}$$

$$\mathrm{i}F^1_{ijklmn} = \begin{cases} \frac{G_{ijklmn}}{\lambda_i+\lambda_j+\lambda_k-\lambda_l-\lambda_m-\lambda_n} & i \pm j \pm k \pm l \pm m \pm n = 0, (i,j,k,l,m,n) \in \Delta_1, \\ 0 & otherwise, \end{cases}$$

$$\mathrm{i}F^2_{ijklmn} = \begin{cases} \frac{G_{ijklmn}}{\lambda_i+\lambda_j+\lambda_k-\lambda_l-\lambda_m-\lambda_n} & i \pm j \pm k \pm l \pm m \pm n = 0, (i,j,k,l,m,n) \in \Delta_2 \setminus \mathcal{M}, \\ 0 & otherwise. \end{cases}$$

Note our Assumptions A, B, C, the remained proof is just a copy of Lemma 4 of [10]. □

Now our Hamiltonian is $H = \Lambda + \overline{G} + \hat{G} + K$. Introduce the symplectic polar and complex coordinates by setting

$$q_j = \begin{cases} \sqrt{2(\xi_j + y_j)} e^{-\mathrm{i}x_j}, & j = n_1, n_2, \ldots, n_b \\ \sqrt{2} z_j, & j \neq n_1, n_2, \ldots, n_b \end{cases}$$

depending on parameters $\xi \in \Pi = [0,1]^b$. The precise domain will be specified later. In order to simplify the expression, we substitute $\xi_{n_j}, j = 1, 2, \ldots, b$ by $\xi_j, j = 1, 2, \ldots, b$. Then one gets

$$\frac{i}{2} \sum_{j \geq 1} dq_j \wedge d\bar{q}_j = \sum_{j = n_1, n_2, \ldots, n_b} dx_j \wedge dy_j + i \sum_{j \neq n_1, n_2, \ldots, n_b} dz_j \wedge d\bar{z}_j.$$

The new Hamiltonian

$$H = \Lambda + \overline{G} + \hat{G} + K = \langle \omega(\xi), y \rangle + \langle \Omega(\xi) z, \bar{z} \rangle + \tilde{G} + \hat{G} + K$$

with frequencies $\omega(\xi) = \alpha' + A(\xi)$, $\Omega(\xi) = \beta' + B(\xi)$, where

$$\alpha' = (\lambda_{n_1}, \lambda_{n_2}, \ldots, \lambda_{n_b}), \beta' = (\lambda_i)_{i \neq n_1, \ldots, n_b},$$

$$\begin{aligned} A(\xi) = \frac{1}{\pi^2} (&10\xi_1^2 + 18\xi_2^2 + \cdots + 18\xi_b^2 + 36\xi_1(\xi_2 + \cdots + \xi_b) + 48(\xi_2\xi_3 + \cdots + \xi_{b-1}\xi_b), \ldots, \\ &18\xi_1^2 + \cdots + 18\xi_{b-1}^2 + 10\xi_b^2 + 36\xi_b(\xi_1 + \cdots + \xi_{b-1}) + 48(\xi_1\xi_2 + \cdots + \xi_{b-2}\xi_{b-1})), \end{aligned}$$

$$B(\xi) = (12G_{n_1 n_1 i}\xi_1^2 + \cdots + 12G_{n_b n_b i}\xi_b^2 + 48G_{n_1 n_2 i}\xi_1\xi_2 + \cdots + 48G_{n_{b-1} n_b i}\xi_{b-1}\xi_b)_{i \neq n_1, \ldots, n_b},$$

and the remainder $\tilde{G} = O(|y|^3) + O(|\xi||y|^2) + O(|\xi||y||z|^2_{a,\rho}) + O(|y|^2|z|^2_{a,\rho}), \hat{G} = O(|\xi|^{\frac{3}{2}}|z|^3_{a,\rho}), K = O(|\xi|^4)$. Rescaling ξ by $\epsilon^6\xi$,$z, \bar{z}$ by $\epsilon^4 z, \epsilon^4\bar{z}$, and y by $\epsilon^8 y$, one obtains a Hamiltonian given by the rescaled Hamiltonian

$$\begin{aligned}\tilde{H}(x, y, z, \bar{z}, \xi) &= \epsilon^{-20} H(x, \epsilon^8 y, \epsilon^4 z, \epsilon^4\bar{z}, \epsilon^6\xi, \epsilon)\\ &= \langle\tilde{\omega}(\xi), y\rangle + \langle\tilde{\Omega}(\xi)z, \bar{z}\rangle + \epsilon\tilde{P}(x, y, z, \bar{z}, \xi, \epsilon),\end{aligned}$$

where $\tilde{\omega}(\xi) = \epsilon^{-12}\alpha' + A(\xi),\ \tilde{\Omega} = \epsilon^{-12}\beta' + B(\xi),\ \xi \in [1, 2]^b$. For simplicity, we rewrite $\tilde{H}$ by H, $\tilde{\omega}$ by ω, $\tilde{\Omega}$ by Ω, and $\tilde{P}$ by P.

In what follows, we use the KAM iteration which involves infinite many steps of coordinate transformations to prove the existence of the KAM tori. To make this quantitative we introduce the following notations and spaces.

Define

$$D(r, s) = \{(x, y, z, \bar{z}) : |Imx| < s, |y| < r^2, ||z||_{a,\rho} < r, ||\bar{z}||_{a,\rho} < r\}$$

a complex neighborhood of $\mathbb{T}^b \times \{y = 0\} \times \{z = 0\} \times \{\bar{z} = 0\}$, where $|\cdot|$ denotes the sup-norm for complex vectors. For a p ($p \geq 1$) order Whitney smooth function $F(\xi)$, define

$$\begin{aligned}||F||^* &= \max\left\{\sup_{\xi\in\Pi}|F|, \ldots, \sup_{\xi\in\Pi}\left|\frac{\partial^p F}{\partial\xi^p}\right|\right\},\\ ||F||_* &= \max\left\{\sup_{\xi\in\Pi}\left|\frac{\partial F}{\partial\xi}\right|, \ldots, \sup_{\xi\in\Pi}\left|\frac{\partial^p F}{\partial\xi^p}\right|\right\}.\end{aligned}$$

If $F(\xi)$ is a vector function from ξ to $\mathcal{H}^{a,\rho}(R^n)$ which is p order Whitney smooth on ξ, define $||F||^*_{a,\rho} = ||(||F_i(\xi)||^*)_i||_{a,\rho}(||F||^*_{R^n} = \max_i(||F_i(\xi)||^*))$. If $F(\eta, \xi)$ is a vector function from $D \times \Pi$ to $\mathcal{H}^{a,\rho}$, define $||F||^*_{a,\rho,D} = \sup_{\eta\in D}||F||^*_{a,\rho}$. We usually omit D for brevity. For functions F, associate a Hamiltonian vector field defined as $X_F = \{F_y, -F_x, iF_{\bar{z}}, -iF_z\}$. Denote the weighted norm for X_F by letting

$$|X_F|^*_{r,D(r,s)} = ||F_y||^* + \frac{1}{r^2}||F_x||^* + \frac{1}{r}\|F_z\|^*_{a,\rho} + \frac{1}{r}\|F_{\bar{z}}\|^*_{a,\rho}.$$

4. An infinite dimensional KAM theorem. Theorem 1 is a direct result of Theorem 2 and measures estimates in section 5. Consider small perturbations of an infinite dimensional Hamiltonian in the parameter dependent normal form

$$N = \langle\omega(\xi), y\rangle + \langle\Omega(\xi)z, \bar{z}\rangle$$

on a phase space

$$\mathcal{P}^{a,\rho} = \mathbb{T}^n \times \mathbb{R}^n \times \mathcal{H}^{a,\rho} \times \mathcal{H}^{a,\rho} \ni (x, y, z, \bar{z}),$$

where

$$\omega_j = \frac{j^d + \cdots}{\epsilon^t} + O(\xi^p)^1, \quad \Omega_j = \frac{j^d + \cdots}{\epsilon^t} + O(\xi^p),$$

$t, p \in \mathbb{N}, \rho > 0, a \geq 0$. Suppose that $||\omega||_* \leq M_1, ||\Omega_j||_* \leq M_2, M_1 + M_2 \geq 1$. Define $M = (M_1 + M_2)^p$. The parameter set Π is $[1, 2]^n$.

[1] $O(\xi^p)$ means pth order terms in $\xi_1, \ldots, \xi_b$

For the Hamiltonian $H = N + P$, there exists n-dimensional, linearly stable torus $\mathcal{T}_0^n = \mathbb{T}^n \times \{0,0,0\}$ with frequencies $\omega(\xi)$ when $P = 0$. Our aim is to prove the persistence of a large portion of this family of linearly stable rotational tori under small perturbations. Suppose that the perturbation P is real analytic in the space variables, C^p in ξ, and for each $\xi \in \Pi$ its Hamiltonian vector field $X_P = (P_y, -P_x, iP_{\bar{z}}, -iP_z)^T$ defines near $\mathcal{T}_0^n$ a real analytic map $X_P : \mathcal{P}^{a,\rho} \to \mathcal{P}^{a,\rho}$. Under the previous assumptions, we have the following theorem.

THEOREM 2. *Suppose that $H = N + P$ satisfies*

$$|X_P|^*_{r,D(s,r)} \le \gamma s^{2(1+\mu)}, \tag{4.1}$$

where γ depends on n, p, τ and M, $\mu = (p+1)\tau + p + \frac{n}{2}$. Then there exists a Cantor set $\Pi_\epsilon \subset \Pi$, a Whitney smooth family of torus embeddings $\Phi : \mathbb{T}^n \times \Pi_\epsilon \to \mathcal{P}^{a,\rho}$, and a Whitney smooth map $\omega_* : \Pi_\epsilon \to \mathbb{R}^n$, such that for each $\xi \in \Pi_\epsilon$, the map Φ restricted to $\mathbb{T}^n \times \{\xi\}$ is a real analytic embedding of a rotational torus with frequencies $\omega_*(\xi)$ for the Hamiltonian H at ξ.

Each embedding is real analytic on $|\mathrm{Im} x| < \frac{s}{2}$, and

$$\|\Phi - \Phi_0\|^*_r \le c\epsilon^{\frac{1}{2}},$$
$$||\omega_* - \omega||^* \le c\epsilon,$$

uniformly on that domain and Π_ϵ, where Φ_0 is the trivial embedding $\mathbb{T}^n \times \Pi \to \mathcal{T}_0^n$. Moreover, there exist Whitney smooth maps ω_m and Ω_m on Π for $m \ge 1$ satisfying $\omega_1 = \omega$, $\Omega_1 = \Omega$ and

$$||\omega_m - \omega||^* \le c\epsilon, \tag{4.2}$$

$$\|\Omega_m - \Omega\|^* \le c\epsilon. \tag{4.3}$$

Remark. Note that in the theorem, we didn't claim that the measure of Π_ϵ is positive. For positive measure, one needs further information of the frequencies $\omega(\xi)$ and $\Omega(\xi)$. We shall come back to this point in section 5.

Since the proof of Theorem 2 is essentially standard, we only state the main step of KAM iteration. The more detailed steps can be found in [13] and other papers.

4.1. Solving the linearized equations and KAM step. At each step of KAM iteration, the symplectic coordinate change Φ is obtained as the time 1-map $X_F^t|_{t=1}$ of the flow of Hamiltonian vector field X_F. Its generating function F and some normal correction $\hat{N}$ to the given normal form N are solutions of the linear equation

$$\{F, N\} + \hat{N} = R, \tag{4.4}$$

where

$$R = \sum_{2m+|q+\bar{q}|\le 2} R_{km q\bar{q}} y^m z^q \bar{z}^{\bar{q}} e^{i\langle k,x\rangle}, R_{km q\bar{q}} = P_{km q\bar{q}},$$

and the coefficients $R_{km q\bar{q}}$ depend on ξ such that $X_R : P^{a,\rho} \to P^{a,\rho}$ is real analytic and Whitney smooth in ξ. Below we solve the linear equation and estimate the generating function F.

LEMMA 5. *Suppose that uniformly on* $\Pi_+ \subset \Pi$,

$$|\langle k,\omega\rangle| \geq \frac{\epsilon^\beta}{A_k} \text{ for } k \neq 0, \tag{4.5}$$

$$|\langle k,\omega\rangle + \Omega_i| \geq \frac{\epsilon^\beta}{A_k}, \tag{4.6}$$

$$|\langle k,\omega\rangle + \Omega_i + \Omega_j| \geq \frac{\epsilon^\beta(|i-j|+1)}{A_k}, \tag{4.7}$$

$$|\langle k,\omega\rangle + \Omega_i - \Omega_j| \geq \frac{\epsilon^\beta(|i-j|+1)}{A_k}, i \neq j. \tag{4.8}$$

Then the linear equation has solution F *and* $\hat{N}$, *which satisfy* $[F] = 0$, $[\hat{N}] = \hat{N}$. *Moreover,*

$$|X_{\hat{N}}|^*_{r,D(s,r)} \leq |X_R|^*_{r,D(s,r)}, |X_F|^*_{r,D(s-\sigma,r)} \leq \frac{cM}{\epsilon^{(p+1)\beta}\sigma^\mu}|X_R|^*_{r,D(s,r)}, \tag{4.9}$$

where $A_k = 1 + |k|^\tau$, β *will be denoted later.*

For the proof, refer to [13].

LEMMA 6. *If* $|X_F|^*_{r,D(s-\sigma,r)} \leq \sigma$, *then for any* $\xi \in \Pi_+$, *the flow* $X_F^t(\cdot,\xi)$ *exists on* $D(s-2\sigma,\frac{r}{2})$ *for* $|t| \leq 1$ *and maps* $D(s-2\sigma,\frac{r}{2})$ *into* $D(s-\sigma,r)$. *Moreover, for* $|t| \leq 1$,

$$|X_F^t - id|^*_{r,D(s-2\sigma,\frac{r}{2})}, \sigma||DX_F^t - Id||^*_{r,r,D(s-3\sigma,\frac{r}{4})} \leq c|X_F|^*_{r,D(s-\sigma,r)},$$

where D *is the differentiation operator with respect to* $(x,y,z,\bar{z})$, id *and* Id *are identity mapping and unit matrix, and the operator norm*

$$||A(\xi,\eta)||_{\bar{r},r,D(s,r)} = \sup_{\eta\in D(s,r)}\sup_{w\neq 0}\frac{||A(\xi,\eta)w||_{a,\bar{r}}}{||w||_{a,r}},$$

$$||A||^*_{r,r} = \max\left\{||A||_{r,r}, \ldots, \left|\left|\frac{\partial^p A}{\partial\xi^p}\right|\right|_{r,r}\right\}.$$

For the proof, see [14].

Below we consider the new perturbation under the sympletic transformation $\Phi = X_F^t|_{t=1}$. Let $|X_P|^*_{r,D(s,r)} \leq \epsilon$. From the above we have

$$R = \sum_{2|m|+|q+\bar{q}|\leq 2} R_{kmq\bar{q}} y^m z^q \bar{z}^{\bar{q}} e^{i\langle k,x\rangle}.$$

Thus $|X_R|^*_{r,D(s,r)} \leq \cdot|X_P|^*_{r,D(s,r)} \leq \cdot\epsilon$, and for $\eta \leq \frac{1}{8}$,

$$|X_{P-R}|^*_{\eta r,D(s,4\eta r)} \leq \cdot\eta|X_P|^*_{r,D(s,r)} \leq \cdot\eta\epsilon. \tag{4.10}$$

Since

$$\hat{N} = \sum_{2|m|+|q+\bar{q}|\leq 2, q=\bar{q}} P_{0mq\bar{q}} y^m z^q \bar{z}^{\bar{q}} e^{i\langle k,x\rangle},$$

the new normal form is

$$N_+ = N + \hat{N} = \langle \omega_+, y\rangle + \langle \Omega_+ z, \bar{z}\rangle.$$

By Lemma 5, one has $|X_{\hat{N}}|^*_{r,D(s,r)} \leq \cdot\epsilon$. Therefore,

$$||\omega_+ - \omega||^*, ||\Omega_+ - \Omega||^* \leq \cdot\epsilon, \tag{4.11}$$

where $||\Omega||^* = \max_{j\geq 1} ||\Omega_j||^*$. If $\frac{cM\epsilon^{1-\beta(p+1)}}{\sigma^{\mu+1}} \leq 1$, by Lemmas 5 and 6, it follows that for $|t| \leq 1$,

$$\frac{1}{\sigma}|X_F^t - id|^*_{r,D(s-2\sigma,\frac{r}{2})}, ||DX_F^t - Id||^*_{r,r,D(s-3\sigma,\frac{r}{4})} \leq \frac{cM\epsilon^{1-(p+1)\beta}}{\sigma^{\mu+1}}. \tag{4.12}$$

Under the transformation $\Phi = X_F^1$, $(N+R)\circ\Phi = N_+ + R_+$, where $R_+ = \int_0^1 \{(1-t)\hat{N} + tR, F\}\circ X_F^t$. Thus, $H\circ\Phi = N_+ + R_+ + (P-R)\circ\Phi = N_+ + P_+$, where the new perturbation

$$P_+ = R_+ + (P-R)\circ\Phi = (P-R)\circ\Phi + \int_0^1 \{\bar{R}(t), F\}\circ X_F^t dt,$$

where $\bar{R}(t) = (1-t)\hat{N} + tR$. Hence, the Hamiltonian vector field of the new perturbation is

$$X_{P_+} = (X_F^1)^*(X_{P-R}) + \int_0^1 (X_F^t)^*[X_{\bar{R}(t)}, X_F]dt.$$

For the estimate of X_{P_+}, we need the following lemma.

LEMMA 7. *If the Hamiltonian vector field $W(\cdot,\xi)$ on $V = D(s-4\sigma, 2\eta r)$ depends on the parameter $\xi \in \Pi_+$ with $||W||^*_{r,V} < +\infty$, and $\Phi = X_F^t : U = D(s-5\sigma,\eta r) \to V$, then $\Phi^* W = (D\Phi)^{-1} W\circ\Phi$ and if $\frac{cM\epsilon^{1-(p+1)\beta}}{n^2\sigma^{\mu+1}} \leq 1$, we have $||\Phi^* W||^*_{\eta r,U} \leq c||W||^*_{\eta r,V}$.*

For the proof, see [14].

Now we estimate X_{P_+}. By Lemma 7, if $\frac{cM\epsilon^{1-(p+1)\beta}}{n^2\sigma^{\mu+1}} \leq 1$,

$$|X_{P_+}|^*_{\eta r,D(s-5\sigma,\eta r)} \leq c|X_{P-R}|^*_{\eta r,D(s-4\sigma,2\eta r)} + c\int_0^1 |[X_{\bar{R}(t)}, X_F]|^*_{\eta r,D(s-4\sigma,2\eta r)}dt.$$

By Cauchy's inequality and Lemma 6, one obtains

$$\begin{aligned}|[X_{\bar{R}(t)}, X_F]|^*_{\eta r,D(s-4\sigma,2\eta r)} &\leq \frac{cM\epsilon^{2-(p+1)\beta}}{\eta^2\sigma^{\mu+1}} \\ &= cM\eta\epsilon,\end{aligned}$$

where one chooses $\eta^3 = \frac{\epsilon^{1-(p+1)\beta}}{\sigma^{\mu+1}}$. Combining (4.10) we have

$$|X_{P_+}|^*_{\eta r,D(s-5\sigma,\eta r)} \leq cM\eta\epsilon.$$

4.2. Iteration and proof of Theorem 2. To iterate the KAM step infinitely we must choose suitable sequences. For $m \geq 1$ set

$$\epsilon_{m+1} = \frac{cM(m)\epsilon_m^{\frac{4}{3}-\frac{1}{3}(p+1)\beta}}{\sigma_m^{\frac{1}{3}(1+\mu)}}, \quad \sigma_{m+1} = \frac{\sigma_m}{2}, \quad \eta_m^3 = \frac{\epsilon_m^{1-(p+1)\beta}}{\sigma_m},$$

where $\beta = \frac{1}{2(p+1)}$. Furthermore, $s_{m+1} = s_m - 5\sigma_m$, $r_{m+1} = \eta_m r_m$, $M(m) = (M_1 + M_2 + 2c(\epsilon_1 + \cdots + \epsilon_{m-1}))^p$, and $D_m = D(s_m, r_m)$. As initial value fix $\sigma_1 = \frac{s_1}{20} \leq \frac{1}{2}$. Choose

$$\epsilon_1 \leq \frac{\gamma_0^6 \sigma_1^{2(\mu+1)}}{c^6 M^{6p}}, \tag{4.13}$$

where $\gamma_0 \leq \frac{1}{c(M+1)^p 2^{6p+4\mu}}$. Finally, let $K_m = K_1 2^{m-1}$ with

$$K_1^{\tau+1} = c^{5-6\beta} M^{6p(1-\beta)} 2^{2(1+\mu)(1-\beta)-3} \gamma_0^{6(\beta-1)}. \tag{4.14}$$

LEMMA 8. *Suppose $H_m = N_m + P_m$ is given on $D_m \times \Pi_m$, where $N_m = \langle \omega_m(\xi), y\rangle + \langle \Omega_m, z\bar{z}\rangle$ is a normal form satisfying*

$$|\langle k, \omega_m\rangle| \geq \frac{\epsilon_m^\beta}{A_k} \text{ for } k \neq 0, \tag{4.15}$$

$$|\langle k, \omega_m\rangle + \Omega_{m,i}| \geq \frac{\epsilon_m^\beta}{A_k}, \tag{4.16}$$

$$|\langle k, \omega_m\rangle + \Omega_{m,i} + \Omega_{m,j}| \geq \frac{\epsilon_m^\beta (|i-j|+1)}{A_k}, \tag{4.17}$$

$$|\langle k, \omega_m\rangle + \Omega_{m,i} - \Omega_{m,j}| \geq \frac{\epsilon_m^\beta (|i-j|+1)}{A_k}, i \neq j, \tag{4.18}$$

for any $\xi \in \Pi_m$, and

$$|X_{P_m}|^*_{r_m, D_m} \leq \epsilon_m.$$

Then there exists a Whitney smooth family of real analytic symplectic coordinate transformations $\Phi_{m+1} : D_{m+1} \times \Pi_m \to D_m$ and a closed subset

$$\Pi_{m+1} = \Pi_m \setminus \bigcup_{|k| > K_m} R_{kl}^{m+1}(\epsilon_{m+1})$$

of Π_m, where

$$R_{kl}^{m+1}(\epsilon_{m+1}) = A_{k1}^{m+1} \cup A_{k2}^{m+1} \cup A_{k3}^{m+1} \cup A_{k4}^{m+1},$$

and

$$A_{k1}^{m+1} = \left\{ \xi \in \Pi_m : |\langle k, \omega_{m+1}\rangle| < \frac{\epsilon_{m+1}^\beta}{A_k} \right\},$$

$$A_{k2}^{m+1} = \bigcup_i B_{ki}^{m+1,1} = \bigcup_i \left\{ \xi \in \Pi_m : |\langle k, \omega_{m+1}\rangle + \Omega_{m+1,i}| < \frac{\epsilon_{m+1}^\beta}{A_k} \right\},$$

$$A_{k3}^{m+1} = \bigcup_{i,j} B_{kij}^{m+1,11} = \bigcup_{i,j} \left\{ \xi \in \Pi_m : |\langle k, \omega_{m+1}\rangle + \Omega_{m+1,i} + \Omega_{m+1,j}| < \frac{\epsilon_{m+1}^\beta (|i-j|+1)}{A_k} \right\},$$

$$A_{k4}^{m+1} = \bigcup_{i \neq j} B_{kij}^{m+1,12} = \bigcup_{i \neq j} \left\{ \xi \in \Pi_m : |\langle k, \omega_{m+1}\rangle + \Omega_{m+1,i} - \Omega_{m+1,j}| < \frac{\epsilon_{m+1}^\beta (|i-j|+1)}{A_k} \right\},$$

such that for $H_{m+1} = H_m \circ \Phi_{m+1} = N_{m+1} + P_{m+1}$ *the same assumptions are satisfied with* $m+1$ *in place of* m.

Proof. Note the value for p_1, ϵ_1, β and σ_1, one verifies that

$$\frac{M_{m+1}\epsilon_{m+1}^{1-(p+1)\beta}}{\sigma_{m+1}^{1+\mu}} \le \frac{1}{2}\frac{M_m\epsilon_m^{1-(p+1)\beta}}{\sigma_m^{1+\mu}} \tag{4.19}$$

for all $m \ge 1$. So the smallness condition of the KAM step is satisfied. For the remained proof, see the iterative lemma in [13]. □

With (4.11) and (4.12), we also obtain the following estimate.

LEMMA 9. *For* $m \ge 1$,

$$\frac{1}{\sigma_m}||\Phi_{m+1} - id||^*_{r_m, D_{m+1}}, ||D\Phi_{m+1} - I||^*_{r_m, r_m, D_{m+1}} \le \frac{cM(m)\epsilon_m^{1-(p+1)\beta}}{\sigma_m^{\mu+1}} \tag{4.20}$$

$$||\omega_{m+1} - \omega_m||^*_{\Pi_m}, ||\Omega_{m+1} - \Omega_m||^*_{\Pi_m} \le c\epsilon_m. \tag{4.21}$$

Proof of Theorem 2. The smallness condition is

$$\epsilon_1 \le \frac{\gamma_0^6}{20^{2(1+\mu)(cM^p)^6}} s_1^{2(1+\mu)}. \tag{4.22}$$

To apply Lemma 8 with $m = 1$, set $s_1 = s$, $r_1 = r, \ldots$, $N_1 = N$, $P_1 = P$,

$$\gamma = \frac{\gamma_0^6}{20^{2(1+\mu)(cM^p)^6}} \text{ and } \epsilon_1 = \gamma s_1^{2(1+\mu)}.$$

The smallness condition is satisfied, because

$$|X_{P_1}|^*_{r_1, D(s_1, r_1)} = |X_P|^*_{r, D(r,s)} \le \gamma s^{2(1+\mu)} = \epsilon_1.$$

The small divisor conditions are satisfied by setting $\Pi_1 = \Pi \setminus \cup_{kl} R^1_{kl}(\epsilon)$, where $k \ne 0$ for A^1_{k1}, and $\Pi_0 = \Pi$. Then the iterative lemma applies. □

Remark. For the rescaled Hamiltonian H, we fix $r = 1$. Then

$$|X_{\epsilon P}|^*_{1, D(s,1)} \le |X_{\epsilon P}|^*_{1, D(1,1)} \le c\epsilon \le \gamma s^{2(1+\mu)},$$

for ϵ small enough. If fix $\rho > 0$ and $a > \frac{1}{2}$ arbitrarily, Theorem 2 can be applied to the rescaled Hamiltonian.

5. Measure estimates. The remaining job is to estimate the measure. We first give the measure estimates for the first step. In our case, the tangent frequencies $\omega_i = \lambda_i + O(\xi^2)(i = n_1, \ldots, n_b)$ and normal frequencies $\Omega_j = \lambda_j + O(\xi^2)(j \ne n_1, \ldots, n_b)$ are second orders in ξ while the ones appeared in the papers such as [10] and [12] are linear in ξ. This is another main difference between our paper and others. To obtain the measure estimates, we have to control the higher order derivatives for $\langle k, \omega\rangle \pm \Omega_i \pm \Omega_j$ etc. One finds that more information from $O(\xi^2)$ is needed to exclude the degenerate cases. The measure estimates in the subsequent steps are based on the techniques developed in [14] and [15].

5.1. Measure estimates in the first step. The thrown parameter sets in the first step are $(\cup_{k\neq 0}A^1_{k1})\bigcup(\cup_k(A^1_{k2}\cup A^1_{k3}\cup A^1_{k4}))$, where

$$A^1_{k1}=\left\{\xi\in\Pi:|\langle k,\omega\rangle|<\frac{\epsilon^\beta}{A_k}\right\}, \tag{5.1}$$

$$A^1_{k2}=\bigcup_i B^{1,1}_{ki}=\bigcup_i\left\{\xi\in\Pi:|\langle k,\omega\rangle+\Omega_i|<\frac{\epsilon^\beta}{A_k}\right\}, \tag{5.2}$$

$$A^1_{k3}=\bigcup_{i,j} B^{1,11}_{kij}=\bigcup_{i,j}\left\{\xi\in\Pi:|\langle k,\omega\rangle+\Omega_i+\Omega_j|<\frac{\epsilon^\beta(|i-j|+1)}{A_k}\right\}, \tag{5.3}$$

$$A^1_{k4}=\bigcup_{i\neq j} B^{1,12}_{kij}=\bigcup_{i\neq j}\left\{\xi\in\Pi:|\langle k,\omega\rangle+\Omega_i-\Omega_j|<\frac{\epsilon^\beta(|i-j|+1)}{A_k}\right\}. \tag{5.4}$$

It is obvious that $|A^1_{02}\cup A^1_{03}\cup A^1_{04}|=0$.

LEMMA 10. *Suppose that $g(x)$ is an* m*th differentiable function on the closure $\bar{I}$ of I, where $I\subset\mathbb{R}$ is an interval. Let $I_h=\{x||g(x)|<h\}, h>0$. If for some constant $d>0$, $|g^m(x)|\geq d$ for any $x\in I$, then $|I_h|\leq ch^{\frac{1}{m}}$, where $|I_h|$ denotes the Lebesgue measure of I_h and $c=2(2+3+\cdots+m+d^{-1})$.*

For the proof, see [15]. The similar method can be found in [1] and [14].

LEMMA 11. *For $\tau>2b+5$, $|\bigcup_{k\neq 0}A^1_{k3}|=O(\epsilon^{\frac{\beta}{2}})$.*

Proof. Suppose $i\geq j$ without losing generalities. When $i\geq c|k|$, one obtains $\frac{|\Omega_i+\Omega_j|}{1+i-j}\geq\frac{c|k|}{8\epsilon^{12}}$. But we know $\frac{|\langle k,\omega\rangle|}{1+(i-j)}\leq\frac{c'|k|}{\epsilon^{12}}$. If c is large enough, then $\frac{|\Omega_i+\Omega_j+\langle k,\omega\rangle|}{1+(i-j)}\geq 1$. This means $A^1_{k3}=\bigcup_{\max\{i,j\}\leq c|k|}B^{1,11}_{kij}$. Define

$$\begin{aligned}f(\xi)=&\frac{k_1}{\pi^2}(10\xi_1^2+18\xi_2^2+\cdots+18\xi_b^2+36\xi_1(\xi_2+\cdots+\xi_b)+48(\xi_2\xi_3+\cdots+\xi_{b-1}\xi_b))+\cdots\\&+\frac{k_b}{\pi^2}(18\xi_1^2+18\xi_2^2+\cdots+10\xi_b^2+36\xi_b(\xi_1+\cdots+\xi_{b-1})+48(\xi_1\xi_2+\cdots+\xi_{b-2}\xi_{b-1}))\\&+(12\xi_1^2G_{n_1n_1i}+\cdots+12\xi_b^2G_{n_bn_bi}+48G_{n_1n_2i}\xi_1\xi_2+\cdots+48G_{n_{b-1}n_bi}\xi_{b-1}\xi_b)\\&+(12\xi_1^2G_{n_1n_1j}+\cdots+12\xi_b^2G_{n_bn_bj}+48G_{n_1n_2j}\xi_1\xi_2+\cdots+48G_{n_{b-1}n_bj}\xi_{b-1}\xi_b).\end{aligned}$$

It follows that

$$\begin{aligned}\frac{\pi^2}{2}\frac{\partial^2 f}{\partial\xi_1^2}&=10k_1+18k_2+\cdots+18k_b+3(c_1+c_1')\\&\vdots\\\frac{\pi^2}{2}\frac{\partial^2 f}{\partial\xi_n^2}&=18k_1+18k_2+\cdots+10k_b+3(c_b+c_b'),\end{aligned}$$

where $c_i,c_i'=5$ or $6, i=1,2,\ldots,b$. We will prove the inequality

$$\max\left(\frac{\pi^2}{2}\left|\frac{\partial^2 f}{\partial\xi_1^2}\right|,\ldots,\frac{\pi^2}{2}\left|\frac{\partial^2 f}{\partial\xi_b^2}\right|\right)\geq 1$$

always holds. If it is not true, one gets that

$$k_1=\frac{-3}{8(9b-4)}((13-9b)(c_1+c_1')+9(c_2+c_2'+\cdots+c_b+c_b')).$$

One can draw the contradictions from the following three cases.

Case 1. Two "5s" in $\{c_1, c_2, \ldots, c_b, c_1', c_2', \ldots, c_b'\}$. In this case, we discuss it from different possibilities.

Subcase a: $c_1 = 5,\ c_1' = 6$. One obtains $k_1 = -\frac{3(9b+26)}{72b-32}$. It is obvious that $k_1 \notin Z$. It is similar for the case $c_1 = 6,\ c_1' = 5$.

Subcase b: $c_1 = c_1' = 6$. One gets $k_1 = -\frac{45}{4(9b-4)}$. It is impossible.

Subcase c: $c_1 = c_1' = 5$. One gets $|k_1| = \frac{3(18b+22)}{8(9b-4)}$. When $b \geq 6$, $0 < |k_1| < 1$. For $b = 2, \ldots, 5$, one can get $k_1 \notin Z$(check directly).

Case 2. Only one "5" in $\{c_1, c_2, \ldots, c_b, c_1', c_2', \ldots, c_b'\}$.

Subcase a: $c_1 = 5$ or $c_1' = 5$. One obtains $|k_1| = \frac{105+27b}{72b-32}$. If $b \geq 4$, then $0 < |k_1| < 1$. It is impossible. If $b = 2, 3$, one has $k_1 \notin Z$(check directly). It also contradicts with the previous assumption.

Subcase b: $c_{k_0} = 5$ or $c_{k_0}' = 5 (k_0 \neq 1)$. One gets $k_1 = -\frac{117}{8(9b-4)}$. It can't happen.

Case 3. No "5s" in $\{c_1, c_2, \ldots, c_b, c_1', c_2', \ldots, c_b'\}$.

One gets $k_1 = -\frac{18}{9b-4}$. If $b \geq 3$, one can get $0 < |k_1| < 1$. When $b = 2$, we obtain $k_1 \notin Z$ directly. It is impossible.

Hence, for any $k \neq 0, i$, there exists some $k_0 \in \{1, 2, \ldots, b\}, s.t., |\frac{\pi^2}{2}\frac{\partial^2 f}{\partial \xi_{k_0}^2}| \geq 1$. Then one obtains

$$\begin{aligned}\left|\bigcup_{k\neq 0} A_{k3}^1\right| &= \left|\bigcup_{k\neq 0}\bigcup_{i,j\leq c|k|} B_{kij}^{1,11}\right|,\\ &\leq \cdot \sum_{k\neq 0}\left(\frac{\epsilon^\beta |k|}{A_k}\right)^{\frac{1}{2}} |k|^2,\\ &\leq \cdot \sum_{l=1}^{+\infty} \frac{1}{l^{\frac{\tau-2b-3}{2}}}\epsilon^{\frac{\beta}{2}},\\ &= O(\epsilon^{\frac{\beta}{2}}). \qquad \square\end{aligned}$$

LEMMA 12. *For* $\tau > 2b + 5$, $|\bigcup_{k\neq 0} A_{k4}^1| = O(\epsilon^{\frac{\beta}{2}})$.

Proof. By the same methods, one obtains that $A_{k4}^1 = \bigcup_{\max\{i,j\}\leq c|k|} B_{kij}^{1,12}$. Following the similar way, we can get

$$k_l = \frac{-3}{8(9b-4)}\left((13-9b)(c_l - c_l') + 9\left(\sum_{m\neq l} c_k - \sum_{m\neq l} c_k'\right)\right),$$

where $l, m \in \{1, \ldots, b\}$.

Case 1. Two "5s" in $\{c_1, c_2, \ldots, c_b, c_1', c_2', \ldots, c_b'\}$.

If for any i, we have $c_i = c_i'$. One gets $k = 0$ in this case. It is impossible. If $\exists i_0, c_{i_0} \neq c_{i_0}'$, there exist two cases. One is $c_{i_0} = 5$, $c_{i_0}' = 6$. The other is $c_{i_0} = 6$, $c_{i_0}' = 5$. In any case, one can get $|k_{i_0}| = \frac{3}{8}$.

Case 2. One "5" in $\{c_1, c_2, \ldots, c_b, c_1', c_2', \ldots, c_b'\}$.

Case 3. No "5s" in $\{c_1, c_2, \ldots, c_b, c_1', c_2', \ldots, c_b'\}$.

We omit the proof for the two cases. The measure estimate is similar as before. We also omit it. $\square$

The following conclusions are obvious according to the above methods.

LEMMA 13. *For $\tau > 2b+2$, $|\bigcup_{k\neq 0} A^1_{k2}| = O(\epsilon^{\frac{\beta}{2}})$.*

LEMMA 14. *For $\tau > 2b$, $|\bigcup_{k\neq 0} A^1_{k1}| = O(\epsilon^{\frac{\beta}{2}})$.*

LEMMA 15. *For $\tau > 2b+5$, $|(\cup_{k\neq 0} A^1_{k1}) \bigcup (\cup_k (A^1_{k2} \cup A^1_{k3} \cup A^1_{k4}))| = O(\epsilon^{\frac{\beta}{2}})$.*

5.2. The total measure. In order to estimate the total measure of the parameter sets Π_ϵ which is thrown in all the steps, we must estimate the measure in the subsequent steps. The thrown parameter set in $m+1$ step is $\bigcup_{|k|>K_m} R^{m+1}_{kl}(\epsilon_{m+1})$, where $\xi \in \Pi_m$. In fact, we may extend ω_m and Ω_m defined in Π_m to Π. The following ω_m and Ω_m are both defined in Π.

LEMMA 16. *For $\tau > 2b+4$ and $K_m \geq \frac{80b}{c_1}$,*

$$\left| \bigcup_{|k|>K_m} A^{m+1}_{k4} \right| = \left| \bigcup_{|k|>K_m} \bigcup_{i\neq j} B^{m+1,12}_{k,ij} \right| = O\left(\epsilon^{\frac{\beta}{2}}_{m+1}\right),$$

where c_1 is a constant which depends on b and will be defined in the following.

Proof. For our convenience, we write ω' and Ω' for ω_{m+1} and Ω_{m+1}. Define $v_1 = (1,0,\ldots,0)^T$ and $v_b = (0,0,\ldots,1)^T$. Define $S = \{(x_1,x_2,\ldots,x_b) \in \mathbb{R}^b : |x_1|+|x_2|+\cdots+|x_b| = 1\}$. Write $A(\xi) = (D^2_{v_1}\omega, D^2_{v_2}\omega, \ldots, D^2_{v_b}\omega)^T$. It is easy to check that $|A(\xi)| = c > 0$, for any $\xi \in \Pi$. For any $(\xi, v) \in \Pi \times S$, $|A(\xi)v|_1 \geq c_1 > 0$. Thus for any $(\xi, v) \in \Pi \times S$, there exists a open neighborhood S_v of v in S, such that for some i, $|\langle D^2_{v_i}\omega, v'\rangle| \geq \frac{c_1}{2b}$, for any $(\xi, v') \in \Pi \times S_v$. Since $\{\Pi \times S_v\}$ covers the compact set $\Pi \times S$, there exist finite covers: $\Pi \times S_1, \ldots, \Pi \times S_{k_0}$ such that $\bigcup_{i=1}^{k_0} \Pi \times S_i \supset \Pi \times S$ and for any $(\xi, v) \in \Pi \times S_i$,

$$|\langle D^2_{\bar{v}}\omega, v\rangle| \geq \frac{c_1}{2b},$$

where $\bar{v} \in \{v_1, v_2, \ldots, v_b\}$.

Now fix $k \neq 0$ and suppose $\frac{k}{|k|} \in S_i$. Then for any $\xi \in \Pi$,

$$\left| \left\langle D^2_{\bar{v}}\omega, \frac{k}{|k|} \right\rangle \right| \geq \frac{c_1}{2b} > 0. \tag{5.5}$$

Define $f(\xi) = \langle k, \omega'\rangle + \Omega'_i - \Omega'_j$. Note

$$\begin{aligned} D^2_{\bar{v}} \frac{f(\xi)}{|k|} = &\left\langle \frac{k}{|k|}, D^2_{\bar{v}}(\omega) \right\rangle + \frac{D^2_{\bar{v}}(\Omega_i - \Omega_j)}{|k|} + \frac{D^2_{\bar{v}}(\Omega'_i - \Omega_i)}{|k|} \\ &+ \frac{D^2_{\bar{v}}(\Omega_j - \Omega'_j)}{|k|} + \left\langle \frac{k}{|k|}, D^2_{\bar{v}}(\omega' - \omega) \right\rangle. \end{aligned} \tag{5.6}$$

We estimate every term in (5.6). From (4.2) and (4.3), one obtains

$$\left| \left\langle \frac{k}{|k|}, D^2_{\bar{v}}(\omega' - \omega) \right\rangle \right| \leq |D^2_{\bar{v}}(\omega' - \omega)| \leq c\epsilon, \tag{5.7}$$

$$\frac{|D^2_{\bar{v}}(\Omega'_i - \Omega_i)|}{|k|} \leq \frac{c\epsilon}{|k|} \leq \frac{1}{|k|}, \tag{5.8}$$

$$\frac{|D^2_{\bar{v}}(\Omega_j - \Omega'_j)|}{|k|} \leq \frac{c\epsilon}{|k|} \leq \frac{1}{|k|}. \tag{5.9}$$

Note $\frac{|D_{\bar v}^2(\Omega_i-\Omega_j)|}{|k|}\le\frac{8}{|k|}$ and (5.7), (5.8), (5.9), (5.5), we arrive at $|D_{\bar v}\frac{f(\xi)}{|k|}|\ge\frac{c_1}{4b}$ when $|k|\ge\frac{80b}{c_1}$. We will show in what follows that when $\max\{i,j\}\ge c|k|$,

$$\frac{|\langle k,\omega'\rangle+\Omega_i'-\Omega_j'|}{|i-j|+1}\ge 1. \tag{5.10}$$

The proof is similar as before. First,

$$\begin{aligned}\frac{|\Omega_i'-\Omega_j'|}{|i-j|+1}&\ge\frac{|\Omega_i-\Omega_j|}{2|i-j|}-\frac{|\Omega_i'-\Omega_i|}{2|i-j|}-\frac{|\Omega_j'-\Omega_j|}{2|i-j|}\\&\ge\frac{c|k|}{2\epsilon^{12}}-M_9-c_*\epsilon\\&\ge\frac{c|k|}{4\epsilon^{12}}.\end{aligned}$$

Moreover,

$$|\langle k,\omega'\rangle|\le|\langle k,\omega\rangle|+|\langle k,\omega'-\omega\rangle|\le\frac{c'|k|}{\epsilon^{12}}.$$

Therefore, when c is large enough and $\max\{i,j\}\ge c|k|$, (5.10) holds. So when $K_m\ge\frac{80b}{c_1}$ and $\tau>2b+4$,

$$\begin{aligned}\left|\bigcup_{|k|>K_m}\bigcup_{i\ne j}B_{k,ij}^{m+1,12}\right|&=\left|\bigcup_{|k|>K_m}\bigcup_{i,j\le c|k|}B_{k,ij}^{m+1,12}\right|\\&\le\sum_{|k|\ge|K_m|}\sum_{i,j\le c|k|}\left(\frac{|i-j|}{A_k|k|}\right)^{\frac12}O(\epsilon_{m+1}^{\frac\beta2}),\\&\le\sum_{l=1}^{+\infty}\frac{1}{l^{\frac\tau2-b-1}}O(\epsilon_{m+1}^{\frac\beta2})\\&=O(\epsilon_{m+1}^{\frac\beta2}).\qquad\square\end{aligned}$$

LEMMA 17. *For $\tau>2b+4$ and $K_m\ge\frac{80b}{c_1}$,*

$$\left|\bigcup_{|k|>K_m}A_{k3}^{m+1}\right|=\left|\bigcup_{|k|>K_m}\bigcup_{i,j}B_{k,ij}^{m+1,11}\right|=O(\epsilon_{m+1}^{\frac\beta2}).$$

LEMMA 18. *For $\tau>2b+1$ and $K_m\ge\frac{80b}{c_1}$,*

$$\left|\bigcup_{|k|>K_m}A_{k2}^{m+1}\right|=\left|\bigcup_{|k|>K_m}\bigcup_{i}B_{k,i}^{m+1,1}\right|=O(\epsilon_{m+1}^{\frac\beta2}).$$

LEMMA 19. *For $\tau>2b-1$,*

$$\left|\bigcup_{k\ne0}A_{k1}^{m+1}\right|=O(\epsilon_{m+1}^{\frac\beta2}).$$

LEMMA 20. *For $\tau > 2b+4$ and $K_m \geq \frac{80b}{c_1}$,*

$$\left| \bigcup_{|k|>K_m} R_{kl}^{m+1}(\epsilon_{m+1}) \right| = O(\epsilon_{m+1}^{\frac{\beta}{2}}).$$

In order to estimate the value for K_1, a series of constants have to be chosen. We know $p = 2$, $\beta = \frac{1}{6}$. One fixes $M, \tau > 2b+5$. It is easy to see that one obtains $K_1 \geq \frac{80b}{c_1}$ when γ_0 is small enough. Now we compute the total measure of the parameter sets Π_ϵ which is thrown in all the steps

$$\begin{aligned} |\Pi_\epsilon| &\leq O(\epsilon_1^{\frac{1}{12}}) + O(\epsilon_2^{\frac{1}{12}}) + \cdots \\ &\leq O(\epsilon_1^{\frac{1}{12}}) = O(\epsilon^{\frac{1}{12}}). \end{aligned}$$

6. Appendix. The existence of infinite admissible index sets isn't obvious since the corresponding tangential frequencies have to satisfy infinite many nonresonance conditions. The main idea of the proof is as follows: Suppose that our conclusions hold when $b = d-1$, we prove that there exists at least one n_d in $[x, x+\sqrt{\frac{x}{9n_{d-1}}}]$($x, n_1$ is large enough) such that $n_1, \ldots, n_{d-1}, n_d$ satisfy all the nonresonance assumptions (see section 3). The idea is to estimate the total number of integers n in $[x, x+\sqrt{\frac{x}{9n_{d-1}}}]$ such that $n_1, n_2, \ldots, n_{d-1}, n$ conflicts with one of our nonresonance assumptions. In fact we can prove that the total number is far less than $\sqrt{\frac{x}{9n_{d-1}}}$. This shows the existence of n_d. In case $d = 2$, we explicitly construct the admissible index sets. The proof of Proposition 1 requires a couple of lemmas. For our convenience, we introduce the set $K^2 = \{k^2 | k \in Z\}$ and define $L = \sqrt{\frac{n_d}{9n_{d-1}}}$.

LEMMA 21. *For any given $n_1, n_2, \ldots, n_{d-1}$ with $n_1 < n_2 < \cdots < n_{d-1}$, $\{n_1, n_2\} \in \mathcal{S}$ and n_d large enough, there exists at most $\frac{L}{8d}$ integers $x_d \in [n_d, n_d+L] \cap Z$ satisfying $5x_d^2 + 2kx_d - 3k^2 \in K^2, k \in \{n_1, n_2, \ldots, n_{d-1}\}$.*

Proof. Note $\sqrt{5} \notin Q$, the conclusion is obvious. □

Similarly, we have the following lemma.

LEMMA 22. *For any given $n_1, n_2, \ldots, n_{d-1}$ with $n_1 < n_2 < \cdots < n_{d-1}$, $\{n_1, n_2\} \in \mathcal{S}$ and n_d large enough, there exist at most $\frac{L}{8d}$ integers $x_d \in [n_d, n_d+L] \cap Z$ satisfying $5x_d^2 - 2kx_d - 3k^2 \in K^2, k \in \{n_1, n_2, \ldots, n_{d-1}\}$.*

For the following two lemmas, it is easy to draw the contradictions from the contrary.

LEMMA 23. *For any given $n_1, n_2, \ldots, n_d$ where $n_1 < n_2 < \cdots < n_{d-1} < n_d$, $n_d >> n_{d-1}^2$ and $\{n_1, n_2\} \in \mathcal{S}$, there exists at most one $x_{ij} \in [n_d, n_d + L] \cap Z$ satisfying*

$$4x_{ij}(n_j + n_i) + (n_j - 3n_i)(n_j + n_i) \in K^2,$$

where $i, j \in \{1, 2, \ldots, d-1\}$.

LEMMA 24. *For any given $n_1, n_2, \ldots, n_d$ with $n_1 < n_2 < \cdots < n_{d-1} < n_d$, $n_d >> n_{d-1}^2$ and $\{n_1, n_2\} \in \mathcal{S}$, there exists at most one $x_{ij} \in [n_d, n_d + L] \cap Z$ satisfying*

$$4x_{ij}(n_j - n_i) + (n_j + 3n_i)(n_j - n_i) \in K^2,$$

where $1 \leq i < j \leq d-1, i, j \in Z$.

LEMMA 25. *For any given $n_1, n_2, \ldots, n_{d-1}, n_d$, where $n_1 < n_2 < \cdots < n_{d-1} < n_d$, $n_d >> n_{d-1}^3$ and $(n_1, n_2) \in \mathcal{S}$, there exists at most $\frac{12L}{\sqrt{n_j - n_i}}$ integers x belonging to $[n_d, n_d + L]$ so that $\frac{x^2 - n_i^2}{n_j - n_i} \in Z$, where $1 \le i < j \le d-1, i, j \in Z$.*

Proof. Rewrite $n_j - n_i = p_1^{k_1} p_2^{k_2} \cdots p_s^{k_s}$, where $p_1, \ldots, p_s$ are different prime numbers, $k_1, k_2, \ldots, k_s \in Z^+$. For $x + n_i, x \in [n_d, n_d + L]$, it is apparent that there exist at most $\frac{L}{p_1^{l_1} p_2^{l_2} \cdots p_s^{l_s}} + 2$ integers including $p_1^{l_1} p_2^{l_2} \cdots p_s^{l_s}$ as factor, where $0 \le l_i \le k_i, i = 1, 2, \ldots, s$. For our convenience, we use $A^{l_1 l_2 \cdots l_s}$ representing the event that $x + n_i$ includes $p_1^{l_1} p_2^{l_2} \cdots p_s^{l_s}$ as factor. Similarly, $B^{l_1 l_2 \cdots l_s}$ represents the event that $x - n_i$ includes $p_1^{k_1 - l_1} p_2^{k_2 - l_2} \cdots p_s^{k_s - l_s}$ as factor. C represents the event $\frac{x^2 - n_i^2}{n_j - n_i} \in Z$. It is apparent that $\cup_{l_1 \cdots l_s} A^{l_1 l_2 \cdots l_s} B^{l_1 l_2 \cdots l_s} = C$. Then the probability of C is

$$\begin{aligned} P(C) &= \sum_{l_1 \cdots l_s} P(A^{l_1 l_2 \cdots l_s} B^{l_1 l_2 \cdots l_s}) \\ &\le \sum_{l_1 \cdots l_s} P(A^{l_1 l_2 \cdots l_s}) P(B^{l_1 l_2 \cdots l_s}) \\ &\le \sum_{l_1 \cdots l_s} \left(\frac{\frac{L}{p_1^{l_1} p_2^{l_2} \cdots p_s^{l_s}} + 2}{L} \right) \left(\frac{\frac{L}{p_1^{k_1 - l_1} p_2^{k_2 - l_2} \cdots p_s^{k_s - l_s}} + 2}{L} \right) \\ &\le 2 \sum_{l_1 \cdots l_s} \frac{1}{p_1^{k_1} p_2^{k_2} \cdots p_s^{k_s}} \\ &\le \frac{2(k_1 + 1) \cdots (k_s + 1)}{p_1^{k_1} p_2^{k_2} \cdots p_s^{k_s}}. \end{aligned}$$

We know that $l + 1 \le p^{\frac{l}{2}} (p \ge 4)$, $l + 1 \le 3^{\frac{l}{2}+1}$, and $l + 1 \le 2^{\frac{l}{2}+1}$, for any $l \ge 1$. Then $P(C) \le \frac{2(k_1+1)\cdots(k_s+1)}{n_j - n_i} \le \frac{12}{\sqrt{n_j - n_i}}$. Now it is easy to see that our conclusion holds. □

Similarly, we have the following lemma.

LEMMA 26. *For any given $n_1, n_2, \ldots, n_d$ where $n_1 < n_2 < \cdots < n_{d-1} < n_d$, $n_d >> n_{d-1}^3$ and $\{n_1, n_2\} \in \mathcal{S}$, there exists at most $\frac{12L}{\sqrt{n_j + n_i}}$ integers x belonging to $[n_d, n_d + L]$ so that $\frac{x^2 - n_i^2}{n_j + n_i} \in Z$, where $L = \sqrt{\frac{n_d}{9 n_{d-1}}}, 1 \le i \le j \le d-1, i, j \in Z$.*

The proof of Proposition 1.

We first admit that Proposition 2 holds (the proof will be delayed to the end). This means that Proposition 1 holds for $b = 2$. Suppose that Proposition 1 holds for $b = d - 1 \ge 2$, we will show that it also holds for d. When $b = d - 1$, one can choose one admissible set made of $n_1, n_2, \ldots, n_{d-1}$. Our aim is to construct n_d so that $\{n_1 < n_2 < \cdots < n_d\}$ is an admissible set for $b = d$. We first construct n_d to satisfy Assumption A. In fact, it is enough when $n_d >> n_i, i \le d-1$. Otherwise one gets

$$\begin{cases} i^2 + j^2 + k^2 = l^2 + m^2 + n^2 \\ i \pm j \pm k \pm l \pm m \pm n = 0, \end{cases}$$

where $i, j, k, l, m, n \in \{n_1, n_2, \ldots, n_d\}$. One can induce the contradictions from different cases. We only prove the case in which there exist two $n_d's$ in $\{i, j, k, l, m, n\}$. For any more or less n_d(at least one n_d), the proof is similar. Note $n_d >> n_i (i \le d-1)$,

one gets $\{i,j,k\}\cap\{l,m,n\}\supset n_d$. Hence, one obtains

$$\begin{cases} j^2+k^2=l^2+m^2 \\ n_d\pm n_d\pm j\pm k\pm l\pm m=0. \end{cases} \tag{6.1}$$

We know that $\pm n_d\pm n_d=0,\pm 2n_d$. For the preceding, from Lemma 5 in [10], it contradicts with our choice of $\{i,j,k,l,m,n\}$. For the last, it is apparent that $|j\pm k\pm l\pm m|<2n_d$. This leads a contradiction to (6.1). If none of $n_d's$ is in $\{i,j,k,l,m,n\}$, this contradicts with the choice of $n_1,n_2,\ldots,n_{d-1}$.

In fact, n_d also satisfies Assumption B under the same condition. If this is not true, then

$$x^2+j^2+k^2-l^2-m^2-n^2=0.$$

The unique index which is different with $n_1,n_2,\ldots,n_d$ is denoted by x. We only prove the case in which there exist three $n_d's$ in $\{j,k,l,m,n\}$. For the other cases (at least one n_d), the method is similar. One can induce the contradictions from the following three cases.

Case 1.

$$\begin{cases} x^2+j^2+k^2=3n_d^2 \\ x\pm j\pm k\pm n_d\pm n_d\pm n_d=0. \end{cases}$$

From $x^2+j^2+k^2=3n_d^2$, we conclude that $x\approx\sqrt{3}n_d$. But from $x\pm j\pm k\pm n_d\pm n_d\pm n_d=0$, we know that $|x|\approx 3n_d$ or n_d. It is impossible.

Case 2.

$$\begin{cases} x^2+n_d^2=m^2+n^2 \\ x\pm n_d\pm n_d\pm n_d\pm m\pm n=0. \end{cases}$$

From $n_d^2>>m^2+n^2$, we know it can't happen.

Case 3.

$$\begin{cases} x^2+j^2=n_d^2+m^2 \\ x\pm j\pm m\pm n_d\pm n_d\pm n_d=0. \end{cases} \tag{6.2}$$

From $x^2+j^2=n_d^2+m^2$, we can get $x\approx n_d$. Hence, (6.2) holds only when $\pm n_d\pm n_d\pm n_d=\pm n_d$. But at this case, from Lemma 5 of [10], one can get $\{x,j\}=\{n_d,m\}$. It can't happen.

If none n_d in $\{j,k,l,m,n\}$, this contradicts with the choice of $n_1,n_2,\ldots,n_{d-1}$.

For Assumption C, one must place much heavier restrictions on n_d. From Lemmas 21–26, we will prove that there exist many integer points x belonging to $[n_d,n_d+\sqrt{\frac{n_d}{9n_{d-1}}}]$ so that $n_1,\ldots,x$ fulfill our Assumption C when n_d and n_1 is large enough. If it isn't true, then

$$i^2+j^2+k^2=l^2+m^2+n^2.$$

The other two indexes different from $n_1,n_2,\ldots,n_d$ are denoted by x,y,

Case 1.

$$\{x,y\}\subset\{i,j,k\} \text{ or } \{x,y\}\subset\{l,m,n\}.$$

Without losing generality, one gets

$$\begin{cases} x^2+y^2=l^2+m^2+n^2-k^2 \\ x\pm y=\pm k\pm l\pm m\pm n. \end{cases}$$

We have to consider several different subcases. For our convenience, we introduce the notation "$|\cdot|$." The equality $|\{k,l,m,n\}|=t, t=1,2,3,4$, means there exist exactly t $n_d's$ in $\{k,l,m,n\}$.

Subcase a.

$$|\{k,l,m,n\}|=1.$$

It is easy to see that the case $x^2+y^2+n_d^2=l^2+m^2+n^2$ can't happen. So only the following case need be considered:

$$\begin{cases} x^2+y^2+k^2=l^2+m^2+n_d^2 \\ x\pm y\pm k\pm l\pm m\pm n_d=0. \end{cases}$$

We only consider the case when

$$x=y\pm k\pm l\pm m\pm n_d. \tag{6.3}$$

For $x=-y\pm k\pm l\pm m\pm n_d$, it is similar. From (6.3), one obtains

$$2y^2+2(\pm k\pm l\pm m\pm n_d)y+(\pm k\pm l\pm m\pm n_d)^2+k^2-l^2-m^2-n_d^2=0.$$

Write $a=\pm k\pm l\pm m$. Note

$$\triangle=4(n_d-a)^2+8(l^2+m^2-a^2-k^2),$$

and $y\in Z$, one gets

$$l^2+m^2=a^2+k^2. \tag{6.4}$$

At the same time, one obtains

$$y=\frac{-(a\pm n_d)\pm(n_d-a)}{2}.$$

By further computations, one knows $|x|=n_d$ or $|y|=n_d$. It is impossible.

Subcase b.

$$|\{k,l,m,n\}|=2.$$

We must consider different cases.

Case I.

$$\begin{cases} x^2+y^2=m^2+n^2 \\ x\pm y\pm m\pm n\pm n_d\pm n_d=0. \end{cases} \tag{6.5}$$

If $\pm n_d\pm n_d=0$, we arrive at $\{x,y\}=\{m,n\}$ from Lemma 5 of [10]. It is impossible. When $\pm n_d\pm n_d=\pm 2n_d$, we get $|x\pm y|<<n_d$ from (6.5). Hence the equality $x\pm y\pm m\pm n\pm 2n_d=0$ can't hold.

Case II.

$$\begin{cases} x^2+y^2+k^2=2n_d^2+n^2 \\ x\pm y\pm k\pm n\pm n_d\pm n_d=0. \end{cases} \tag{6.6}$$

Write $\pm k\pm n=a$. If $\pm n_d\pm n_d=0$, one gets

$$2y^2+2ay+(a^2+k^2-2n_d^2-n^2)=0$$

and

$$\triangle=16n_d^2+8n^2-8k^2-4a^2.$$

From $y\in Z$, one obtains $8n^2-8k^2-4a^2=0$. Then we have $3k^2\pm 2kn-n^2=0$. Hence, $3k=n$ or $k=n$. Only the last case need be considered. But at this case, we get $|y|=n_d$. It is impossible.

If $\pm n_d\pm n_d=2n_d$, from (6.6), one gets $|x|<<\sqrt{3}n_d$. Then the equality $x\pm y\pm k\pm n+2n_d=0$ can't hold. Similarly, the equality $x-y\pm k\pm n-2n_d=0$ can't be true. So the only case need be considered is

$$x+y\pm k\pm n-2n_d=0. \tag{6.7}$$

Denote $\pm k\pm n=a$. From (6.6) and (6.7), one gets

$$2y^2-2y(a+2n_d)+2n_d^2+4an_d+a^2+k^2-n^2=0.$$

If $a=-k-n$, we obtain $\triangle=4(n_i+n_j)(4n_d+n_j-3n_i)=4\triangle_1, i,j\in\{1,2,\ldots,d-1\}$. If $a=-k+n, k\neq n$, we obtain $\triangle=4(n_j-n_i)(4n_d+3n_i+n_j)=4\triangle_2$. Other cases can't happen. In order to draw the contradictions, one removes all the integers belonging to $[n_d,n_d+L]$ which satisfy $\triangle_1\in K^2,\triangle_2\in K^2$. Thanks to Lemmas 23 and 24, we throw at most $2(d-1)^2$ integer points. Then $y\notin Z$.

Subcase c.

$$|\{k,l,m,n\}|=3.$$

Case I.

$$\begin{cases} x^2+y^2=n_d^2+n^2 \\ x\pm y\pm n\pm n_d\pm n_d\pm n_d=0. \end{cases} \tag{6.8}$$

If $\pm n_d\pm n_d\pm n_d=\pm n_d$, from (6.8) and Lemma 5 of [10], one gets $\{x,y\}=\{n_d,n\}$. It is impossible. If $\pm n_d\pm n_d\pm n_d=\pm 3n_d$, from $x^2+y^2=n_d^2+n^2$ one obtains $|\pm x\pm y\pm n|<<\frac{5}{2}n_d$. Hence the equality $x\pm y\pm n\pm 3n_d=0$ can't hold.

Case II.

$$\begin{cases} x^2+y^2+k^2=3n_d^2 \\ x\pm y\pm k\pm n_d\pm n_d\pm n_d=0. \end{cases} \tag{6.9}$$

If $\pm n_d\pm n_d\pm n_d=\pm 3n_d$, from $x^2+y^2+k^2=3n_d^2$, one gets $|x\pm y|^2\le 2(x^2+y^2)\le 6n_d^2$. Hence, one knows $|x\pm y\pm k|<<\sqrt{7}n_d$. The inequality $x\pm y\pm k\pm 3n_d=0$ can't hold. For the case when $\pm n_d\pm n_d\pm n_d=\pm n_d$, we throw all the integers belonging to $[n_d,n_d+L]$ which satisfy $5n_d^2\pm 2kn_d-3k^2\in K^2, k=n_1,\ldots,n_{d-1}$. From Lemmas 21 and 22, the thrown integers are at most $\frac{L}{4}$. Then $y\notin Z$.

Subcase d.

$$|\{k,l,m,n\}| = 4.$$

$$\begin{cases} x^2 + y^2 = 2n_d^2 \\ x \pm y \pm n_d \pm n_d \pm n_d \pm n_d = 0. \end{cases}$$

The discussion is trivial. We omit it.

Subcase e. If which is no n_d in $\{k,l,m,n\}$, this contradicts with the choice of $n_1, n_2, \ldots, n_{d-1}$.

Case 2.

$$\{x,y\} \cap \{i,j,k\} \neq \{x,y\} \text{ and } \{x,y\} \cap \{l,m,n\} \neq \{x,y\}.$$

In this case, one obtains

$$\begin{cases} x^2 - y^2 = m^2 + n^2 - i^2 - j^2 \\ x \pm y \pm i \pm j \pm m \pm n = 0. \end{cases} \tag{6.10}$$

We have to discuss it in several subcases.

Subcase a'.

$$|\{i,j,m,n\}| = 1.$$

In this case, (6.10) is

$$\begin{cases} x^2 - y^2 = m^2 + n_d^2 - i^2 - j^2 \\ x \pm y \pm i \pm j \pm m \pm n_d = 0. \end{cases} \tag{6.11}$$

Without losing generality, we suppose that $x = y \pm i \pm j \pm m \pm n_d$. From (6.11), one gets

$$2y(\pm i \pm j \pm m \pm n_d) + (\pm i \pm j \pm m \pm n_d)^2 + i^2 + j^2 - m^2 - n_d^2 = 0.$$

Write $a = \pm i \pm j \pm m$. If $x = y + a + n_d$, one has

$$y = -a + \frac{a^2 + m^2 - i^2 - j^2}{2(a + n_d)}.$$

If $n_d >> n_{d-1}^2$ and $y \in Z$, one obtains $a^2 + m^2 - i^2 - j^2 = 0$ and $y = -a$. Hence $|x| = n_d$. It is impossible. If $x = y + a - n_d$, the proof is similar.

Subcase b'. $|\{i,j,m,n\}| = 2$.

If $\{x,i,j\} \cap \{y,m,n\} = \{n_d\}$, then

$$\begin{cases} x^2 + i^2 = y^2 + m^2 \\ x \pm y \pm i \pm m \pm n_d \pm n_d = 0. \end{cases} \tag{6.12}$$

When $\pm n_d \pm n_d = 0$, from Lemma 5 of [10], one gets $\{x,i,n_d\} = \{y,m,n_d\}$. It contradicts with our assumptions. When $\pm n_d \pm n_d = \pm 2n_d$, write $a = \pm i \pm m$. Without losing generality, we suppose that

$$x = y - a \pm 2n_d. \tag{6.13}$$

From (6.13) and (6.12), one gets

$$y = \frac{m^2 - i^2 - (-a \pm 2n_d)^2}{2(-a \pm 2n_d)}.$$

Note that $y > 0$, we have

$$y = \frac{1}{2}(a + 2n_d) + \frac{m^2 - i^2}{-2(a + 2n_d)}.$$

If $\{\frac{1}{2}(a + 2n_d)\} = \frac{1}{2}$, one gets $y \notin Z$. It is impossible. If $\{\frac{1}{2}(a + 2n_d)\} = 0$ and $m^2 - i^2 \neq 0$, we know $y \notin Z$. It is also impossible. If $\{\frac{1}{2}(a + 2n_d)\} = 0$ and $m^2 = i^2$, one gets $x = y$. It can't happen.

If $\{x, i, j\} \cap \{y, m, n\} = \emptyset$, we have

$$\begin{cases} x^2 + 2n_d^2 = y^2 + m^2 + n^2 \\ x \pm y \pm m \pm n \pm n_d \pm n_d = 0. \end{cases} \tag{6.14}$$

When $\pm n_d \pm n_d = \pm 2n_d$, write $\pm m \pm n = a$. If $x = y + a \pm 2n_d$, from (6.14) and $y > 0$, we get

$$y = \frac{6n_d - a}{4} + \frac{2m^2 + 2n^2 - a^2}{4(a - 2n_d)}.$$

If $\{\frac{6n_d - a}{4}\} \neq 0$, one gets $y \notin Z$. It is impossible. Only when $\{\frac{6n_d - a}{4}\} = 0$ and $a = 2m$ or $a = -2m$, we gets $y \in Z$. But by further computation, one gets $x < 0$. It is impossible. If $x = -y + a \pm 2n_d$, one get $x < 0$ by similar method. When $\pm n_d \pm n_d = 0$, we throw all the integers x belonging to $[n_d, n_d + L]$ which satisfy $\frac{x^2 - m^2}{n - m} \in Z (1 \leq m < n \leq d - 1)$ or $\frac{x^2 - m^2}{n + m} \in Z (1 \leq m \leq n \leq d - 1)$. From Lemmas 25 and 26, the thrown integers are at most $\frac{24L(d-1)^2}{\sqrt{2n_1}}$. Then $y \notin Z$.

Subcase c′. $|\{i, j, m, n\}| = 3$.

In this case, we get

$$\begin{cases} x^2 + i^2 = y^2 + n_d^2 \\ x \pm y \pm i \pm n_d \pm n_d \pm n_d = 0. \end{cases} \tag{6.15}$$

When $\pm n_d \pm n_d \pm n_d = \pm n_d$, from Lemma 5 of [10], one obtains $\{x, i\} = \{y, n_d\}$. It is impossible. When $\pm n_d \pm n_d \pm n_d = \pm 3n_d$, we suppose $x = y \pm i \pm 3n_d$. For $x = -y \pm i \pm 3n_d$, the method is similar. From (6.15) and $y > 0$, we get

$$y = \frac{4}{9}(3n_d + i) + \frac{4i^2}{9(3n_d - i)} - i,$$

or

$$y = \frac{4}{9}(3n_d - i) + \frac{4i^2}{9(3n_d + i)} + i.$$

Both can't be integers. It is impossible.

Subcase d′. $|\{i, j, m, n\}| = 4$.

We easily get $x = y$. It means $\{x, n_d, n_d\} = \{y, n_d, n_d\}$. It contradicts with our assumptions.

Subcase e'. If there is no n_d in $\{i, j, m, n\}$, this contradicts with the choice of $n_1, n_2, \ldots, n_{d-1}$.

Now we declare that there exist many integers x belonging to $[n_d, n_d + L]$ so that $n_1, \ldots, x$ fulfill our Assumptions A, B, C and (2.6) when n_d is large enough and $n_1 \geq 18432d^2$. In fact the thrown integers are at most

$$\frac{24L(d-1)^2}{\sqrt{2n_1}} + \frac{L}{4} + 2(d-1)^2. \tag{6.16}$$

If $n_1 \geq 18432d^2$, one can get (6.16)$\leq \frac{L}{2}$. Then there exist many x satisfying Assumptions A, B, C and (2.6). From Proposition 2, Proposition 1 is complete. □

The proof of Proposition 2 also requires a couple of lemmas. Since the proof is elementary, we give them without proof as follows.

LEMMA 27. *If* $n_1, n_2 \in \mathbb{N}, n_1 < n_2$, *then* $-7n_2^2 + n_1^2 \pm 6n_1n_2 \notin K^2$.

LEMMA 28. *If* $n_1 \equiv 2$ or $5(mod 7), n_2 \equiv 1(mod 7)$, *then* $-7n_1^2 + n_2^2 \pm 6n_1n_2 \notin K^2$.

LEMMA 29. *If* $n_1, n_2 \in \mathbb{N}, n_2 > 11n_1^2$, *then* $-3n_1 \pm n_2 \dagger n_2^2 - n_1^2, 3n_2 \pm n_1 \dagger n_2^2 - n_1^2$, *where the notation* $a\dagger b$ *means that* a *is not a factor of* b.

LEMMA 30. *If* $n_1 \in 2\mathbb{N} - 1, n_2 \in 2\mathbb{N},$, *then* $5n_1^2 - 3n_2^2 \pm 2n_1n_2 \notin K^2$, $5n_2^2 - 3n_1^2 \pm 2n_1n_2 \notin K^2$, $n_1(2n_2 - n_1) \notin K^2$, $n_2^2 - n_1^2 \notin K^2$.

Since the proof of Proposition 2 is similar with Proposition 1, we omit it. □

Acknowledgments. The authors thank the anonymous referees for several kind suggestions. It is also a pleasure to thank Dr. Jiansheng Geng for many fruitful discussions on this subject.

REFERENCES

[1] D. BAMBUSI, *Birkhoff normal form for some nonlinear PDEs*, Comm. Math. Phys., 234 (2003), pp. 253–283.

[2] J. BOURGAIN, *Nonlinear Schrödinger equations (Park City Lectures)*, AMS, Providence, RI, 1999.

[3] J. COLLIANDER, M. KEEL, G. STAFFILANI, H. TAKAOKA, AND T. TAO, *Global well-posedness for Schrödinger equations with derivative*, SIAM J. Math. Anal., 33 (2001), pp. 649–669.

[4] J. COLLIANDER, M. KEEL, G. STAFFILANI, H. TAKAOKA, AND T. TAO, *A refined global well-posedness result for Schrödinger equations with derivative*, SIAM J. Math. Anal., 34 (2002), pp. 64–86.

[5] L. CHIERCHIA AND J. YOU, *KAM tori for* $1D$ *nonlinear wave equations with periodic boundary conditions*, Comm. Math. Phys., 211 (2000), pp. 497–525.

[6] J. GENG AND J. YOU, *KAM tori of Hamiltonian perturbations of* $1D$ *linear beam equations*, J. Math. Anal. Appl., 277 (2003), pp. 104–121.

[7] S. B. KUKSIN, *Nearly integrable infinite-dimensional Hamiltonian systems*, Lecture Notes in Mathematics 1556, Springer-Verlag, Berlin, 1993.

[8] S. B. KUKSIN, *A KAM-theorem for equations of the Korteweg-de Vries type*, Rev. Math. Math. Phys., 10 (1998), pp. 1–64.

[9] S. B. KUKSIN, *Analysis of Hamiltonian PDEs*, Oxford University Press, Oxford, 2000.

[10] S. B. KUKSIN AND J. PÖSCHEL, *Invariant Cantor manifolds of quasi-periodic oscillations for a nonlinear Schrödinger equation*, Ann. of Math., 143 (1996), pp. 149–179.

[11] F. MERLE, *On uniqueness and continuation properties after blow-up time of self-similar solutions of nonlinear Schrödiger equation with critical exponent and critical mass*, Comm. Pure Appl. Math., 45 (1992), pp. 203–254.

[12] J. PÖSCHEL, *Quasi-periodic solutions for a nonlinear wave equation*, Comment. Math. Helv., 71 (1996), pp. 269–296.

[13] J. PÖSCHEL, *A KAM-theorem for some nonlinear partial differential equations*, Ann. Scuola Norm. Sup. Pisa, Cl. Sci., 23 (1996), pp. 119–148.

[14] J. Xu , J. You, and Q. Qiu, *A KAM theorem of degenerate infinite-dimensional Hamiltonian systems.* I. II. Sci. China Ser. A, 39 (1996), pp. 372–394.
[15] J. Xu, J. You, Q. Qiu, *Invariant tori for nearly integrable Hamiltonian systems with degeneracy*, Math. Z., 226 (1997), pp. 375–387.
[16] C. E. Wayne, *Periodic and quasi-periodic solutions for nonlinear wave equations via KAM theory*, Commun. Math. Phys., 127 (1990), pp. 479–528.

SIAM J. MATH. ANAL.
Vol. 36, No. 6, pp. 1991–2003

RENORMALIZED SOLUTIONS TO A NONLINEAR PARABOLIC-ELLIPTIC SYSTEM*

MARÍA TERESA GONZÁLEZ MONTESINOS† AND FRANCISCO ORTEGÓN GALLEGO‡

Abstract. The aim of this paper is to show the existence of renormalized solutions to a parabolic-elliptic system with unbounded diffusion coefficients. This system may be regarded as a modified version of the well-known thermistor problem; in this case, the unknowns are the temperature in a conductor and the electrical potential.

Key words. renormalized solutions, nonlinear elliptic equations, nonlinear parabolic equations, weak solutions, Caratheodory functions, thermistor problem, Sobolev spaces

AMS subject classifications. 35M10, 35J60, 35K65

DOI. 10.1137/S0036141003423041

1. Introduction.

This paper is concerned with the resolution of the nonlinear parabolic-elliptic system

$$
(1)\qquad \begin{cases} \dfrac{\partial u}{\partial t} - \nabla \cdot (a(u)\nabla u) = \sigma(u)|\nabla \varphi|^2 & \text{in } Q = \Omega \times (0,T), \\ -\nabla \cdot (\sigma(u)\nabla \varphi) = \nabla \cdot F(u) & \text{in } Q, \\ u = 0 & \text{on } \partial\Omega \times (0,T), \\ \varphi = 0 & \text{on } \partial\Omega \times (0,T), \\ u(\cdot,0) = u_0 & \text{in } \Omega, \end{cases}
$$

where $\Omega \subset \mathbb{R}^N$ is a bounded domain, $T > 0$, $a(x,t,s)$, $\sigma(x,t,s)$, and $F(x,t,s)$, $F = (F_1, \dots, F_N)'$, are Caratheodory functions defined in $Q \times \mathbb{R}$. This problem has a similar structure to the so-called thermistor problem arising in electromagnetism ([4, 12]); in that particular context, Ω stands for the domain occupied by the thermistor, u is the temperature, u_0 the initial temperature, φ is a shifted electric potential, $F(x,t,s) = \sigma(s)\nabla\varphi_0(x,t)$, φ_0 is a given function, and σ is a continuous and bounded function. Indeed, the actual electric potential is $\psi = \varphi + \varphi_0$, and thus φ_0 is the electric potential Dirichlet boundary data on $\partial\Omega \times (0,T)$. In our analysis, and from a mathematical standpoint, we will consider more general functions $F(x,t,s)$.

A great deal of attention has been paid to the thermistor problem during the last two decades by several authors ([2, 4, 13, 26], etc.). In these works, many situations and different hypotheses have been considered, but both a and σ are assumed to be bounded in all these referred works.

The goal of this paper is to analyze problem (1) in the case of nonbounded diffusion coefficients a and σ. Moreover, no asymptotic behavior on a, σ, and F is assumed.

Under these general assumptions, one readily realizes that weak solutions (in the sense of distributions) are not well suited in this context. Note that even if u or φ

*Received by the editors February 17, 2003; accepted for publication (in revised form) October 7, 2004; published electronically June 30, 2005. This research was partially supported by Ministerio de Ciencia y Tecnología of the Spanish Government and FEDER European fund under grant BFM2003-01187, and by Consejería de Educación y Ciencia de la Junta de Andalucía, research group FQM–315.
http://www.siam.org/journals/sima/36-6/42304.html

†Departamento de Matemáticas, Facultad de Ciencias Económicas y Empresariales, Universidad de Cádiz, 11002 Cádiz, Spain (mariateresa.gonzalez@uca.es).

‡Departamento de Matemáticas, Facultad de Ciencias del Mar, Universidad de Cádiz, 11510 Puerto Real, Cádiz, Spain (francisco.ortegon@uca.es).

belong to some Banach space of the form $L^q(W^{1,q}(\Omega))$, the terms $a(u)\nabla u$, $\sigma(u)\nabla\varphi$, or $F(u)$ may not belong to any $L^r(Q)$ space, $r \geq 1$. For this reason, we consider the notion of renormalized solutions adapted to our setting. The concept of renormalized solution was fist introduced by DiPerna and Lions ([15, 16]) in the framework of the Fokker–Plank–Boltzmann equations; later on, it was applied to more general situations (for instance, in the resolution of nonlinear elliptic equations ([9, 22, 23]), or in the resolution of nonlinear parabolic equations ([6, 7, 8])).

The fact that a and σ are unbounded is not the only difficulty we may encounter in the resolution of problem (1). Indeed, the parabolic equation needs a special treatment due to the nonlinear right-hand side belonging to $L^1(Q)$.

In order to solve problem (1) under the assumptions stated below, we use truncation and approximate solutions. This work is organized as follows.

In section 2, we set up the notation used in the paper; this leads to the introduction of some functional spaces. We also recall certain compactness results and give an existence theorem for problem (1) in the case of bounded data.

Section 3 enumerates the hypotheses and introduces the concept of renormalized solution adapted to our context. Finally, we give the existence result.

Section 4 develops the proof of the existence result; it is split into three steps, namely: setting of approximate problems, derivation of estimates, and passing to the limit and conclusion.

2. Notation and functional spaces. Let $\Omega \subset \mathbb{R}^N$, $N \geq 1$, be an open bounded domain, and $\partial\Omega$ its boundary. Then we define $\mathcal{D}(\Omega)$ as the space of all C^∞–functions in Ω with compact support.

For $p \in [1, +\infty]$, let $W^{1,p}(\Omega)$ be the first order Sobolev space given as

$$W^{1,p}(\Omega) = \left\{ v \in L^p(\Omega) \,/\, \nabla v \in L^p(\Omega)^N \right\},$$

where the gradient $\nabla v = \left(\frac{\partial v}{\partial x_1}, \ldots, \frac{\partial v}{\partial x_N} \right)'$ is taken in the sense of distributions (here, the prime symbol stands for vector transposition). It is well-known that $W^{1,p}(\Omega)$ is a Banach space with norm

$$\begin{aligned} \|v\|_{W^{1,p}(\Omega)} &= \left(\|v\|^p_{L^p(\Omega)} + \|\nabla v\|^p_{L^p(\Omega)^N} \right)^{1/p}, \quad p \in [1, +\infty), \\ \|v\|_{W^{1,\infty}(\Omega)} &= \|v\|_{L^\infty(\Omega)} + \|\nabla v\|_{L^\infty(\Omega)^N}; \end{aligned}$$

moreover, if $p = 2$, then we write $H^1(\Omega) = W^{1,2}(\Omega)$, which is a Hilbert space.

Since we deal with homogenous Dirichlet boundary conditions, it is interesting to introduce the space $W^{1,p}_0(\Omega)$ defined as the closure of $\mathcal{D}(\Omega)$ with respect to $\|\cdot\|_{W^{1,p}(\Omega)}$, that is,

$$W^{1,p}_0(\Omega) = \overline{\mathcal{D}(\Omega)}^{W^{1,p}(\Omega)}, \quad p \in [1, +\infty).$$

It is known that if $\partial\Omega$ is smooth enough (for instance, Lipschitz continuous), $W^{1,p}_0(\Omega)$ is characterized by the following property:

$$W^{1,p}_0(\Omega) = \{ v \in W^{1,p}(\Omega) \,/\, v_{|\partial\Omega} = 0 \}, \quad p \in [1, +\infty).$$

Also we put $H^1_0(\Omega) = W^{1,2}_0(\Omega)$. $W^{1,p}_0(\Omega)$ and $H^1_0(\Omega)$ are, respectively, Banach and Hilbert spaces. By Poincaré's inequality, the seminorm $|v|_{W^{1,p}(\Omega)} = \|\nabla v\|_{L^p(\Omega)^N}$ is a

norm in $W_0^{1,p}(\Omega)$ equivalent to $\|\cdot\|_{W^{1,p}(\Omega)}$ on $W_0^{1,p}(\Omega)$. The space $W^{-1,p'}(\Omega)$ stands for the dual space of $W_0^{1,p}(\Omega)$, $p \in [1,+\infty)$.

We now introduce some notation according to the parabolic equation of (1). For a Banach space X and $1 \le p \le +\infty$, let $L^p(X)$ denote the space $L^p([0,T];X)$, that is, the set of (equivalence class of) measurable functions $f : [0,T] \to X$ such that $t \in [0,T] \mapsto \|f(t)\|_X$ is in $L^p(0,T)$. If $f \in L^p(X)$, we define

$$\|f\|_{L^p(X)} = \left(\int_0^T \|f(t)\|_X^p \right)^{1/p}, \ 1 \le p < +\infty, \quad \|f\|_{L^\infty(X)} = \operatorname*{ess\,sup}_{t\in[0,T]} \|f(t)\|_X;$$

and thus $\left(L^p(X), \|\cdot\|_{L^p(X)}\right)$ is a Banach space. By Fubini's theorem we can identify the space $L^p(L^p(\Omega))$ with $L^p(Q)$, Q being the cylinder $\Omega \times (0,T)$.

Let X and Y be two Banach spaces, $X \hookrightarrow Y$ with continuous inclusion, and set

$$W = \left\{ v \in L^p(X) \,/\, \frac{\mathrm{d}v}{\mathrm{d}t} \in L^q(Y) \right\}, \ p,q \in [1,+\infty],$$

provided with the standard norm $\|w\|_W = \|w\|_{L^p(X)} + \left\|\frac{\mathrm{d}v}{\mathrm{d}t}\right\|_{L^q(Y)}$. Then $(W, \|\cdot\|_W)$ is a Banach space and the inclusion $W \hookrightarrow C^0([0,T];Y)$ holds and is continuous. However, it will be very interesting and useful to know if a particular compactness embedding involving these spaces holds. The answer is given by the following two lemmas ([24]).

LEMMA 1. *Let X, B, and Y be three Banach spaces such that $X \hookrightarrow B \hookrightarrow Y$, every embedding being continuous and the inclusion $X \hookrightarrow B$ compact. Let $1 \le p < +\infty$ and $1 \le q \le +\infty$. Then, the inclusion $W \hookrightarrow L^p(B)$ holds and is compact.*

LEMMA 2. *Let X, B and Y be as in Lemma 1, and $E \subset L^\infty(X)$ be a bounded set such that*

(i) $\frac{\mathrm{d}v}{\mathrm{d}t} \in L^1(Y)$ *for all $v \in E$, and*

(ii) *there exist $h \in L^1(0,T)$, $s > 1$ and a bounded set $\mathcal{Z} \subset L^s(0,T)$ such that*

$$\left\| \frac{\mathrm{d}v}{\mathrm{d}t} \right\|_Y \le h + z_v, \ \text{for all } v \in E, \ z_v \in \mathcal{Z} \text{ and a.e. in } (0,T).$$

Then, E is relatively compact in $C^0([0,T];B)$.

The approximate problems in section 4.1 are defined via truncation functions. For this purpose, we introduce, for each $j > 0$ in $\mathbb{R}$, the truncation function at height j to be

$$T_j(s) = \operatorname{sign}(s) \min(j, |s|), \quad \operatorname{sign}(s) = \begin{cases} 0 & \text{if } s = 0, \\ s/|s| & \text{if } s \neq 0. \end{cases} \tag{2}$$

We will also make use of the following lemma, due to Boccardo and Gallouët ([10]) and ([19]).

LEMMA 3. *Let (v_n) be a sequence of measurable functions in Q such that*

1. *(v_n) is bounded in $L^\infty(L^1(\Omega))$.*
2. *For all $j > 0$, $n \ge 0$, $T_j(v_n) \in L^2(H_0^1(\Omega))$.*
3. *There exists a constant $C > 0$ such that*

$$\int_{\{m \le |v_n| < m+1\}} |\nabla v_n|^2 \le C \ \text{for all } m, n \ge 0.$$

Then (v_n) is bounded in the space $L^q(W^{1,q}(\Omega))$ for all $q < \frac{N+2}{N+1}$ if $N \geq 2$, and for all $q < 2$ if $N = 1$.

If $g : Q \times \mathbb{R}$ is a Caratheodory function and u is measurable in Q, we write $g(u)$ for the measurable function in Q defined as $(x,t) \in Q \mapsto g(x,t,u(x,t))$.

In what follows, $C > 0$ stands for generic constant values which only depend on initial data.

The introduction of the approximate solutions relies on the following result.

THEOREM 4. *Assume that the Caratheodory functions a, σ and F are such that $a, \sigma \in L^\infty(Q \times \mathbb{R})$, $F \in L^\infty(Q \times \mathbb{R})^N$ and there exist two constant values $a_0 > 0$ and σ_0 satisfying*

$$a(x,t,s) \geq a_0,\ \sigma(x,t,s) \geq \sigma_0,\ \textit{for all } s \in \mathbb{R},\ \text{a.e. } (x,t) \in Q.$$

Finally, let $u_0 \in L^2(\Omega)$. Then, for every $j > 0$, there exists $u \in L^2(H_0^1(\Omega))$ and $\varphi \in L^\infty(H_0^1(\Omega))$ such that

$$\frac{\mathrm{d}u}{\mathrm{d}t} \in L^2(H^{-1}(\Omega)), \quad u(\cdot,0) = u_0 \textit{ in } \Omega,$$

and

$$(3)\quad \begin{cases} \displaystyle\int_0^T \left\langle \frac{\mathrm{d}u}{\mathrm{d}t}, v \right\rangle + \int_Q a(u)\nabla u \nabla v = \int_Q T_j\left(\sigma(u)|\nabla\varphi|^2\right) v,\ \textit{for all } v \in L^2(H_0^1(\Omega)), \\ \displaystyle\int_\Omega \sigma(u)\nabla\varphi\nabla\psi = \int_\Omega F(u)\nabla\psi,\ \textit{for all } \psi \in H_0^1(\Omega),\ \text{a.e. } t \in (0,T). \end{cases}$$

For the proof of this result one may follow the same arguments as in the proof of the existence theorem for the thermistor problem ([4]).

3. The main result. We make the following assumptions:

(H.1) $a, \sigma : Q \times \mathbb{R} \to \mathbb{R}$ and $F : Q \times \mathbb{R} \to \mathbb{R}^N$ are Caratheodory functions and there exists a nondecreasing function $\gamma : \mathbb{R}^+ \to \mathbb{R}^+$ such that

$$\max\left(a(x,t,s), \sigma(x,t,s), |F(x,t,s)|\right) \leq \gamma(|s|),\ \text{for all } s \in \mathbb{R},\ \text{a.e. in } Q.$$

(H.2) There exist two constant values $a_0 > 0$ and $\sigma_0 > 0$ such that

$$a(x,t,s) \geq a_0,\ \sigma(x,t,s) \geq \sigma_0,\ \text{for all } s \in \mathbb{R},\ \text{a.e. in } Q.$$

(H.3) There exists a function $\Gamma \in L^1(Q)$ such that

$$|F(x,t,s)|^2 \leq \Gamma(x,t)\sigma(x,t,s),\ \text{for all } s \in \mathbb{R},\ \text{a.e. in } Q.$$

(H.4) $\displaystyle\max_{k \leq |s| \leq 2k} \operatorname*{ess\,sup}_Q \frac{1}{k}\frac{\sigma(x,t,s)}{a(x,t,s)} = \omega(k)$ as $k \to +\infty$, where $\omega(k)$ stands for a null sequence, that is, $\lim_{k\to\infty} \omega(k) = 0$.

(H.5) $u_0 \in L^1(\Omega)$.

Hypothesis (H.1) is one of the main difficulties in the resolution of problem (1). As it has been stated in section 1, we cannot expect to search for weak solutions. However, assumptions (H.3) and (H.4) give a relation of the asymptotic behavior of $a(s)$, $\sigma(s)$ and $F(s)$ for large values of s.

We introduce now the definition of renormalized solutions to problem (1).

DEFINITION 5. *A couple of functions (u, φ) is called a renormalized solution to problem* (1) *if the following conditions are fulfilled:*

(R.1) $u \in L^1(\Omega)$, $\varphi \in L^2(H_0^1(\Omega))$, *and* $\int_Q \sigma(u)|\nabla\varphi|^2 < +\infty$;
(R.2) $T_M(u) \in L^2(H_0^1(\Omega))$ *for all* $M > 0$;
(R.3) $\lim_{n\to\infty} \int_{\{n\le|u|<n+1\}} a(u)\nabla u\nabla u = 0$;
(R.4) *For all* $S \in C^\infty(\mathbb{R})$ *with* supp S' *compact,*

$$\frac{\partial S(u)}{\partial t} - \nabla\cdot[a(u)\nabla u S'(u)] + S''(u)a(u)\nabla u\nabla u = \sigma(u)|\nabla\varphi|^2 S'(u) \text{ in } \mathcal{D}'(Q),$$
$$S(u(\cdot,0)) = S(u_0) \text{ in } \Omega;$$

(R.5) *For all* $\psi \in L^2(H_0^1(\Omega))$ *such that* $\int_Q \sigma(u)|\nabla\psi|^2 < +\infty$, *we have*

$$\int_Q \sigma(u)\nabla\varphi\nabla\psi = -\int_Q F(u)\nabla\psi.$$

Remark. Properties (R.1)–(R.4) on u are the usual conditions verified by renormalized solutions of parabolic equations ([7]). On the other hand, (R.5) says in particular that the set of test functions in the equation for φ depends upon the solution u.

We can now state the main result of this work.

THEOREM 6. *Under hypotheses* (H.1)–(H.5), *system* (1) *admits a renormalized solution* (u,φ) *in the sense of Definition* 5.

4. Proof of Theorem 6. The proof is divided into three steps: first, we introduce a sequence of approximate problems; then, we derive certain estimates for the approximate solutions; and finally, we pass to the limit and conclude.

4.1. Setting of the approximate problems. For every $j > 0$, we consider the truncation functions defined by

$$a_j(x,t,s) = a(x,t,T_j(s)), \quad \sigma_j(x,t,s) = \sigma(x,t,T_j(s)), \quad F_j(x,t,s) = F(x,t,T_j(s)),$$

where T_j is defined in (2). Thanks to $a_j, \sigma_j \in L^\infty(Q\times\mathbb{R})$ and $F_j \in L^\infty(Q\times\mathbb{R})^N$.

The approximate problems are stated as follows: to find $u_j \in L^2(H_0^1(\Omega))$ and $\varphi_j \in L^\infty(H_0^1(\Omega))$ such that $\frac{\mathrm{d}u_j}{\mathrm{d}t} \in L^2(H^{-1}(\Omega))$, $u_j(\cdot,0) = T_j(u_0)$ in Ω and

$$(4)\quad \begin{cases} \displaystyle\int_0^T \left\langle \frac{\mathrm{d}u_j}{\mathrm{d}t}, v\right\rangle + \int_Q a_j(u_j)\nabla u_j\nabla v = \int_Q T_j\left(\sigma_j(u_j)|\nabla\varphi_j|^2\right) v, \text{ for all } v\in L^2(H_0^1(\Omega)), \\ \displaystyle\int_\Omega \sigma_j(u_j)\nabla\varphi_j\nabla\psi = -\int_\Omega F_j(u_j)\nabla\psi, \text{ for all } \psi\in H_0^1(\Omega), \text{ a.e. } t\in(0,T). \end{cases}$$

By virtue of Theorem 4, we know that for each $j > 0$, there exists (u_j,φ_j) verifying all these conditions.

4.2. Estimates for (u_j) and (φ_j). Choosing $\psi = \varphi_j$ in the equation for φ_j and integrating over Q yields,

$$\int_Q \sigma_j(u_j)|\nabla\varphi_j|^2 = -\int_Q F_j(u_j)\nabla\varphi_j \le \left(\int_Q \sigma_j(u_j)^{-1}|F_j(u_j)|^2\right)^{1/2}\left(\int_Q \sigma_j(u_j)|\nabla\varphi_j|^2\right)^{1/2},$$

hence, using (H.3),

$$(5)\qquad \int_Q \sigma_j(u_j)|\nabla\varphi_j|^2 \le \int_Q \sigma_j(u_j)^{-1}|F_j(u_j)|^2 \le \int_Q \Gamma = C.$$

In this way, the sequence $\left(\sigma_j(u_j)|\nabla\varphi_j|^2\right)$ is bounded in $L^1(Q)$. We may rewrite the parabolic equation of (4) as

$$\left\{ \begin{array}{c} \displaystyle\int_0^T \left\langle \frac{\mathrm{d}u_j}{\mathrm{d}t}, v \right\rangle + \int_Q a_j(u_j)\nabla u_j \nabla v = \int_Q f_j v, \text{ for all } v \in L^2(H_0^1(\Omega)), \\ u_j(\cdot,0) = T_j(u_0), \end{array} \right. \tag{6}$$

where $f_j = T_j\left(\sigma_j(u_j)|\nabla\varphi_j|^2\right)$. Since the sequences (f_j) and $(T_j(u_0))$ are bounded in $L^1(Q)$ and $L^1(\Omega)$, respectively, we may deduce some well-known estimates for the sequence of solutions to (6) (u_j) in suitable Banach spaces ([7, 10]), namely

$$(u_j) \text{ is bounded in } L^\infty(L^1(\Omega)); \tag{7}$$

for all $M > 0$ and $j \geq 1$, there exists a constant $C > 0$, not depending upon M and j, such that

$$\int_Q |\nabla T_M(u_j)|^2 \leq CM, \tag{8}$$

$$\int_{\{M \leq |u_j| < M+1\}} |\nabla u_j|^2 \leq C, \tag{9}$$

and also

$$\int_{\{M \leq |u_j| < M+1\}} a_j(u_j)\,|\nabla u_j|^2 \leq \int_{\{|u_j| > M\}} |f_j| + \int_{\{|u_0| > M\}} |u_0|. \tag{10}$$

Owing to (7), (9), and Lemma 3, we have

$$(u_j) \text{ is bounded in } L^q(W_0^{1,q}(\Omega)), \text{ for all } q < \frac{N+2}{N+1} \text{ if } N \geq 2,\ q < 2 \text{ if } N = 1. \tag{11}$$

As far as the parabolic term $\frac{\mathrm{d}u_j}{\mathrm{d}t}$ is concerned, we proceed as follows. Let $S \in C^\infty(\mathbb{R})$ with supp $S' \subset [-M, M]$. Taking $v = S'(u_j)\phi$, $\phi \in \mathcal{D}(\Omega)$, in (6), it yields

$$\frac{\mathrm{d}S(u_j)}{\mathrm{d}t} - \nabla \cdot [a_j(u_j)\nabla u_j S'(u_j)] + S''(u_j)a_j(u_j)\nabla u_j \nabla u_j = f_j S'(u_j) \text{ in } \mathcal{D}'(\Omega). \tag{12}$$

Thanks to (8) and (H.1) we obtain

$$\left(\frac{\mathrm{d}S(u_j)}{\mathrm{d}t}\right) \text{ is bounded in } L^2(H^{-1}(\Omega)) + L^1(Q).$$

Since $L^2(H^{-1}(\Omega)) + L^1(Q) \hookrightarrow L^1(W^{-1,r}(\Omega))$, $r < \frac{N}{N-1}$, with continuous inclusion, we have

$$\left(\frac{\mathrm{d}S(u_j)}{\mathrm{d}t}\right) \text{ is bounded in } L^1(W^{-1,r}(\Omega)) \text{ for all } r < \frac{N}{N-1}. \tag{13}$$

Furthermore, using (11), we readily have

$$(S(u_j)) \text{ is bounded in } L^q(W_0^{1,q}(\Omega)), \text{ for all } q < \frac{N+2}{N+1}, \text{ if } N \geq 2,\ q < 2 \text{ if } N = 1.$$

Now we apply the compactness result stated in Lemma 1. To do so, we take

$$X = W_0^{1,q}(\Omega), \quad B = L^q(\Omega), \quad Y = W^{-1,r}(\Omega);$$

therefore,

$$(S(u_j)) \text{ is relatively compact in } L^q(Q) \text{ for all } q < \frac{N+2}{N+1}. \tag{14}$$

Property (14) is not enough to deduce the almost everywhere convergence of (u_j) modulo a subsequence. We must also use the estimates derived above. To this end, let $M > 0$ and consider a function $S \in C^\infty(\mathbb{R})$ satisfying

(i) supp S' is compact,
(ii) S is nondecreasing, and
(iii) $S(s) = s$ if $|s| \le M$.

Therefore, we have the identity $T_M(s) = T_M(S(s))$ for all $s \in \mathbb{R}$, and, in particular,

$$T_M(u_j) = T_M(S(u_j)). \tag{15}$$

According to (8), for every $M > 0$ there exist a subsequence, which will be denoted in the same way, and a function $z_M \in L^2(H_0^1(\Omega))$ such that

$$T_M(u_j) \to z_M \text{ weakly in } L^2(H_0^1(\Omega)). \tag{16}$$

On the other hand, from (14), there exist a subsequence, still denoted in the same way, and a function $\varsigma_S \in L^q(Q)$ such that

$$S(u_j) \to \varsigma_S \text{ strongly in } L^q(Q) \text{ and a.e. in } Q. \tag{17}$$

Notice that (15) and (17) imply that $T_M(u_j)$ converges almost everywhere to $T_M(\varsigma_S)$; this fact, together with (16), implies that $z_M = T_M(\varsigma_S)$.

Furthermore, from (11), there exist $u \in L^q(W_0^{1,q}(\Omega))$ and a subsequence of (u_j) such that

$$u_j \to u \text{ weakly in } L^q(W_0^{1,q}(\Omega)), \text{ for all } q < \frac{N+2}{N+1} \text{ if } N \ge 2,\ q < 2 \text{ if } N = 1.$$

All these convergences lead to (modulo a subsequence) the almost everywhere convergence of (u_j). Indeed, this property can be readily derived from the next result ([19]).

LEMMA 7. *Let $q \ge 1$, $A \subset \mathbb{R}^N$ a nonnegligible measurable set, $(w_j) \subset L^q(A)$, $w \in L^q(A)$ be such that*

$$w_j \to w \text{ weakly in } L^q(A).$$

Assume that for every $M > 0$ there exists $v_M \in L^1(A)$ such that

$$T_M(v_j) \to v_M \text{ a.e. in } A,$$

then $T_M(w) = v_M$, for all $M > 0$ (and in particular $w_j \to w$ almost everywhere in A).

Summing up, we have shown the existence of subsequences, still denoted in the same way, (u_j), (φ_j), and functions $u \in L^q(W_0^{1,q}(\Omega))$ and $\varphi \in L^2(H_0^1(\Omega))$ such that

$$u_j \to u \text{ weakly in } L^q(W_0^{1,q}(\Omega)), \text{ for all } q < \frac{N+2}{N+1} \text{ if } N \ge 2,\ q < 2 \text{ if } N = 1, \tag{18}$$

$$T_M(u_j) \to T_M(u) \quad \text{weakly in } L^2(H_0^1(\Omega)), \tag{19}$$

$$u_j \to u \quad \text{a.e. in } Q, \tag{20}$$

$$S(u_j) \to S(u) \quad \text{strongly in } L^r(Q) \text{ for all } r < +\infty, \tag{21}$$

$$\frac{\mathrm{d}S(u_j)}{\mathrm{d}t} \to \frac{\mathrm{d}S(u)}{\mathrm{d}t} \quad \text{in } \mathcal{D}'(Q), \tag{22}$$

$$\varphi_j \to \varphi \quad \text{weakly in } L^2(H_0^1(\Omega)), \tag{23}$$

where (21) and (22) are valid for all $S \in C^\infty(\Omega)$ with supp S' compact, and (23) is obtained from (5) and (H.2).

Now we turn our attention to (φ_j) and φ. First of all, we show that

$$\sigma_j(u_j)^{1/2}\nabla\varphi_j \to \sigma(u)^{1/2}\nabla\varphi \text{ weakly in } L^2(Q)^N. \tag{24}$$

Indeed, from (5), there exist a subsequence and $\Phi \in L^2(Q)^N$ such that

$$\sigma_j(u_j)^{1/2}\nabla\varphi_j \to \Phi \text{ weakly in } L^2(Q)^N. \tag{25}$$

Using (20) and (H.2), it yields

$$\sigma_j(u_j)^{-1/2} \to \sigma(u)^{-1/2} \text{ weakly–}* \text{ in } L^\infty(Q) \text{ and a.e. in } Q. \tag{26}$$

Putting

$$\nabla\varphi_j = \sigma_j(u_j)^{-1/2}\sigma_j(u_j)^{1/2}\nabla\varphi_j, \tag{27}$$

and passing to the limit, gathering (25)–(27), we obtain $\Phi = \sigma(u)^{1/2}\nabla\varphi$, and this shows the statement (24). Notice that, in particular, $\sigma(u)|\nabla\varphi|^2 \in L^1(Q)$.

One of the most delicate parts in the passing to the limit consists in showing the convergence

$$\sigma_j(u_j)^{1/2}\nabla\varphi_j \to \sigma(u)^{1/2}\nabla\varphi \text{ strongly in } L^2(Q)^N. \tag{28}$$

From (24), it is enough to show that

$$\int_Q \sigma_j(u_j)|\nabla\varphi_j|^2 \to \int_Q \sigma(u)|\nabla\varphi|^2. \tag{29}$$

To do this, we first introduce the function $S_k \in W^{1,\infty}(\mathbb{R})$, $k > 0$, defined as

$$S_k(s) = \begin{cases} 1 & \text{if } |s| \le k, \\ (2k - |s|)/k & \text{if } k < |s| \le 2k, \\ 0 & \text{if } |s| > 2k. \end{cases} \tag{30}$$

Note that supp $S_k = [-2k, 2k]$ and $S_k'(s) = \frac{1}{k}\left(\chi_{(-2k,-k)} - \chi_{(k,2k)}\right)$. Then, we take in (4) the test function $\psi = S_k(u_j)T_M(\varphi) \in L^\infty(H_0^1(\Omega))$. The integration over $(0,T)$ leads to

$$\begin{aligned} &\int_Q \sigma_j(u_j)\nabla\varphi_j\nabla T_M(\varphi)S_k(u_j) + \int_Q \sigma_j(u_j)\nabla\varphi_j\nabla u_j S_k'(u_j)T_M(\varphi) \\ &\quad = -\int_Q F_j(u_j)\nabla T_M(\varphi)S_k(u_j) - \int_Q F_j(u_j)\nabla u_j S_k'(u_j)T_M(\varphi); \end{aligned}$$

we call these terms (I)–(IV) and study them separately.
(I). Since $\sigma_j(u_j)S_k(u_j) = \sigma_j(T_{2k}(u_j))S_k(u_j) \in L^\infty(Q)$ and is bounded in this space, using (20) it yields

$$\sigma_j(u_j)S_k(u_j) \to \sigma(u)S_k(u) \text{ weakly–}* \text{ in } L^\infty(Q) \text{ and a.e. in } Q.$$

From (23), making $j \to \infty$, we readily obtain

$$\int_Q \sigma_j(u_j)\nabla\varphi_j\nabla T_M(\varphi)S_k(u_j) \to \int_Q \sigma(u)\nabla\varphi T_M(\varphi)S_k(u).$$

Owing to Lebesgue's theorem, we finally deduce

$$\lim_{M\to\infty}\lim_{k\to\infty}\lim_{j\to\infty}\int_Q \sigma_j(u_j)\nabla\varphi_j\nabla T_M(\varphi)S_k(u_j) = \int_Q \sigma(u)|\nabla\varphi|^2.$$

(II). We first derive another estimate for (u_j). Let $H_k \in W^{1,\infty}(\mathbb{R})$ be the function

$$H_k(s) = \begin{cases} 0 & \text{if } |s| \le k, \\ (|s|-k)/k & \text{if } k < |s| \le 2k, \\ |s|/s & \text{if } |s| > 2k, \end{cases}$$

then put $\tilde{H}_k(s) = \int_0^s H_k(\tau)\mathrm{d}\tau$ and $E_j^k = \{k < |u_j| < 2k\}$. Choosing $v = H_k(u_j)$ in (4) yields

$$\int_\Omega \tilde{H}_k(u_j(T)) + \frac{1}{k}\int_{E_j^k} a_j(u_j)|\nabla u_j|^2 = \int_Q f_j H_k(u_j) + \int_\Omega \tilde{H}_k(T_j(u_0));$$

therefore, for all $j \ge 1$ and $k > 0$, there exists a constant $C > 0$, not depending upon j and k, such that

$$\frac{1}{k}\int_Q a_j(u_j)|\nabla u_j|^2\chi_{E_j^k} \le C,$$

that is,

$$\left(\frac{1}{\sqrt{k}}a_j(u_j)^{1/2}\nabla u_j\chi_{E_j^k}\right) \text{ is bounded (in } j \text{ and } k) \text{ in } L^2(Q)^N. \tag{31}$$

Going back to (II)

$$(II) = \int_Q \sigma_j(u_j)^{1/2}\nabla\varphi_j\sigma_j(u_j)^{1/2}a_j(u_j)^{-1/2}a_j(u_j)^{1/2}\nabla u_j S_k'(u_j)T_M(\varphi),$$

thus

$$\begin{aligned} |(II)| &\le M\int_Q \left|\sigma_j(u_j)^{1/2}\nabla\varphi_j\,\frac{1}{\sqrt{k}}\sigma_j(u_j)^{1/2}a_j(u_j)^{-1/2}\,\frac{1}{\sqrt{k}}a_j(u_j)^{1/2}\nabla u_j\chi_{E_j^k}\right| \\ &\le M\left\|\sigma_j(u_j)^{1/2}\nabla\varphi_j\right\|_{L^2(Q)}\cdot\left\|\frac{1}{\sqrt{k}}a_j(u_j)^{1/2}\nabla u_j\chi_{E_j^k}\right\|_{L^2(Q)} \\ &\qquad\cdot\left\|\frac{1}{\sqrt{k}}\sigma_j(u_j)^{1/2}a_j(u_j)^{-1/2}\chi_{E_j^k}\right\|_{L^\infty(Q)}. \end{aligned}$$

Hence, from (H.4), (5), and (31), we deduce

$$|(II)| \leq C\omega(k),$$

which implies

$$\lim_{k\to\infty} \limsup_{j\to\infty} \int_Q \sigma_j(u_j)\nabla\varphi_j \nabla u_j S_k'(u_j) T_M(\varphi) = 0.$$

(III). Lebesgue's theorem easily shows that

$$\lim_{j\to\infty} \int_Q F_j(u_j)\nabla T_M(\varphi) S_k(u_j) = \int_Q F(u)\nabla T_M(\varphi) S_k(u).$$

We now express this last integral as

$$\int_Q F(u)\sigma(u)^{-1/2}\sigma(u)^{1/2}\nabla T_M(\varphi) S_k(u).$$

Owing to (H.3) and (24) we can apply again Lebesgue's theorem, first in k, then in M, to deduce finally that

$$\lim_{M\to\infty} \lim_{k\to\infty} \lim_{j\to\infty} \int_Q F_j(u_j)\nabla T_M(\varphi) S_k(u_j) = \int_Q F(u)\nabla\varphi. \tag{32}$$

(IV). Following the same techniques as in (II) and (III), it is straightforward that

$$\lim_{k\to\infty} \limsup_{j\to\infty} \int_Q F_j(u_j)\nabla u_j S_k'(u_j) T_M(\varphi) = 0.$$

Gathering (27)–(32),

$$\int_Q \sigma(u)|\nabla\varphi|^2 = -\int_Q F(u)\nabla\varphi. \tag{33}$$

On the other hand, taking $\psi = \varphi_j$ in (4) and integrating over $(0,T)$, we obtain

$$\int_Q \sigma_j(u_j)|\nabla\varphi_j|^2 = -\int_Q F_j(u_j)\nabla\varphi_j;$$

since $F_j(u_j)\nabla\varphi_j = F_j(u_j)\sigma_j(u_j)^{-1/2}\sigma_j(u_j)^{1/2}\nabla\varphi_j$, and bearing in mind (H.3), (20), and (24), we conclude that

$$\int_Q F_j(u_j)\nabla\varphi_j \to \int_Q F(u)\nabla\varphi; \tag{34}$$

putting together (33)–(34) gives directly (29), that is, $\sigma_j(u_j)^{1/2}\nabla\varphi_j \to \sigma(u)^{1/2}\nabla\varphi$ strongly in $L^2(Q)^N$. This also implies that

$$f_j = T_j\left(\sigma_j(u_j)|\nabla\varphi_j|^2\right) \to \sigma(u)|\nabla\varphi|^2 \text{ strongly in } L^1(Q). \tag{35}$$

The last relevant convergence to be shown before passing to the limit in the approximate problems (4) is,

$$T_M(u_j) \to T_M(u) \text{ strongly in } L^2(H_0^1(\Omega)), \text{ for every } M > 0. \tag{36}$$

In fact, this is a consequence of (6), (19), and (35), but it is not an immediate result; for details of the proof of this property the reader is referred to [8].

4.3. Passing to the limit and conclusion. Let u and φ be the limit functions given in (18) and (23). Here we show that both functions verify (R.1)–(R.5) of Definition 5.

In fact, (R.1) and (R.2) have been already obtained.

By virtue of (19), (20), and (35), making $j \to \infty$ in (10) yields

$$\int_{\{M\le|u|<M+1\}} a(u)|\nabla u|^2 \le \int_{\{|u|>M\}} \sigma(u)|\nabla\varphi|^2 + \int_{|u_0|>M} |u_0|;$$

due to hypothesis (H.5) and making $M \to \infty$ in this last expression, we can easily derive (R.3).

In order to obtain (R.4), we just take $v = S(u_j)\phi$ in (4) with $S \in C^\infty(\mathbb{R})$, supp S' compact and $\phi \in \mathcal{D}(\Omega)$. Thanks to the convergence properties derived in the preceding section, we can make $j \to \infty$ and this yields the variational formulation (R.4). Note that the strong convergence of the truncations function $T_M(u_j) \to T_M(u)$ in $L^2(H_0^1(\Omega))$ is essential in this stage. It remains to state the initial condition $S(u(\cdot,0)) = S(u_0)$; to do so, we apply Lemma 2 with the following choices:

$$X = L^\infty(\Omega), \quad B = Y = W^{-1,r}(\Omega), \text{ any } r < \frac{N}{N-1},$$

and put $E = \{S(u_j)\}_{j\ge1}$, supp $S' = [-M, M]$. Obviously, E is bounded in $L^\infty(X)$ and, according to (13), $\frac{\mathrm{d}v}{\mathrm{d}t} \in L^1(Y)$ for all $v \in E$. Also, by virtue of (12), we can write

$$\frac{\mathrm{d}S(u_j)}{\mathrm{d}t} = f_j S'(u_j) - S''(u_j)a_j(T_M(u_j))|\nabla T_M(u_j)|^2 + \nabla\cdot[a_j(T_M(u_j))\nabla T_M(u_j)S'(u_j)]\,.$$

Now, from (20) and (35), $f_jS'(u_j)$ converges strongly in $L^1(Q)$ and from (20) and (36), $S''(u_j)a_j(T_M(u_j))|\nabla T_M(u_j)|^2$ converges strongly in $L^1(Q)$. Owing to Lebesgue's inverse theorem, there exists $\bar h \in L^1(Q)$ such that

$$|\Phi_j| \le \bar h \text{ for all } j \ge 1 \text{ and a.e. in } Q,$$

where $\Phi_j = f_jS'(u_j) - S''(u_j)a_j(u_j)|\nabla u_j|^2$. Consequently,

$$\|\Phi_j\|_{W^{-1,r}(\Omega)} \le C\|\Phi_j\|_{L^1(\Omega)} \le C\|\bar h\|_{L^1(\Omega)}, \text{ for all } r < \frac{N}{N-1},\ j \ge 1, \text{a.e. } t \in (0,T).$$

On the other hand, the last term $\nabla\cdot[a_j(T_M(u_j))\nabla T_M(u_j)S'(u_j)]$ is bounded in $L^2(H^{-1}(\Omega))$, and therefore it is also bounded in $L^2(W^{-1,r}(\Omega))$, for all $r < \frac{N}{N-1}$. Hence, we may take $h = C\|\bar h\|_{L^1(\Omega)} \in L^1(0,T)$ and $s = 2$ to deduce that

$$\left\|\frac{\mathrm{d}S(u_j)}{\mathrm{d}t}\right\|_Y \le h + \|\nabla\cdot[a_j(T_M(u_j))\nabla T_M(u_j)S'(u_j)]\|_Y\,, \text{ for all } j \ge 1, \text{ a.e. } t \in (0,T).$$

By Lemma 2, this means that $(S(u_j))$ is relatively compact in $C^0\left([0,T];W^{-1,r}(\Omega)\right)$ for any $r < \frac{N}{N-1}$ and thus, there exists a subsequence, still denoted in the same way, such that $(S(u_j))$ converges in $C^0\left([0,T];W^{-1,r}(\Omega)\right)$. From (21), this limit must be $S(u)$. In particular,

$$S(u_j(\cdot,0)) \to S(u(0)) \text{ in } W^{-1,r}(\Omega),$$

and since $S(u_j(0)) = S(T_j(u_0)) \to S(u_0)$ in $L^1(\Omega)$-strongly, we deduce the initial condition

$$S(u(\cdot,0)) = S(u_0) \text{ in } W^{-1,r}(\Omega),\ r < \frac{N}{N-1}.$$

Finally, in order to derive (R.5), we just take $\psi = S_k(u_j)T_M(\phi)$ in (3), where S_k is defined in (30) and $\phi \in L^2(H_0^1(\Omega))$ is such that $\int_Q \sigma(u)|\nabla\phi|^2 < +\infty$. In this situation, we can proceed as in (I)–(IV) above: taking the iterate limits, first in j, then in k, then in M, and the last expression becomes (R.5).

This ends the proof of Theorem 6. □

5. Concluding remarks. The diffusion coefficients a and σ are scalar functions in the setting given by hypotheses (H.1)–(H.4). We may consider a more general setting in which a and σ are diffusion matrices of order $N \times N$. The hypotheses on this data read as follows:

(H.1) $a, \sigma : Q \times \mathbb{R} \to \mathbb{R}^{N\times N}$ and $F : Q \times \mathbb{R} \to \mathbb{R}^N$ are Caratheodory functions and there exists a nondecreasing function $\gamma : \mathbb{R}^+ \to \mathbb{R}^+$ such that

$$\max\left(\|a(x,t,s)\|, \|\sigma(x,t,s)\|, |F(x,t,s)|\right) \le \gamma(|s|),\ \text{for all } s \in \mathbb{R}, \text{ a.e. in } Q,$$

where $\|\cdot\|$ stands for the spectral norm.

(H.2) There are two constant values $a_0 > 0$ and $\sigma_0 > 0$ so that

$$a(x,t,s)\xi\xi \ge a_0|\xi|^2,\ \sigma(x,t,s)\xi\xi \ge \sigma_0|\xi|^2,\ \text{for all } s \in \mathbb{R},\ \xi \in \mathbb{R}^N, \text{ a.e. in } Q.$$

(H.3) $\Gamma \in L^1(Q)$ is a function satisfying

$$|\sigma(x,t,s)^{-\mathrm{S}/2}F(x,t,s)|^2 \le \Gamma(x,t),\ \text{for all } s \in \mathbb{R}, \text{ a.e. in } Q.$$

(H.4) $\displaystyle\max_{k\le|s|\le 2k} \operatorname*{ess\,sup}_Q \frac{1}{\sqrt{k}}\|\sigma(x,t,s)^{\mathrm{S}/2}a(x,t,s)^{-\mathrm{S}/2}\| = \omega(k)$ as $k \to +\infty$.

(H.5) $u_0 \in L^1(\Omega)$.

The notation in (H.3) and (H.4) is now explained: for a matrix $B \in \mathbb{R}^{N\times N}$, we denote by B^{S} the symmetric part of B, that is, $B^{\mathrm{S}} = (B + B')/2$. From (H.2), $\sigma(x,t,s)^{\mathrm{S}}$ and $a(x,t,s)^{\mathrm{S}}$ are positive definite; then $\sigma(x,t,s)^{\mathrm{S}/2}$ stands for the unique positive definite square root of $\sigma(x,t,s)$, whereas $a(x,t,s)^{-\mathrm{S}/2}$ represents the inverse matrix of the unique positive definite square root of $a(x,t,s)^{\mathrm{S}}$.

In this situation, the existence result given in Theorem 6 still holds true.

The analysis described in this paper shows that the concept of renormalized solutions may be applied to systems of parabolic-elliptic equations with unbounded diffusion coefficients. The existence result relies on certain assumptions on data, apart from the standard ones, describing the relation of the asymptotic behavior between them.

The uniqueness of renormalized solution to problem (1) is a very complex task to be deduced; this is due to the fact that all known uniqueness results for the thermistor problem are derived from L^∞ estimates verified by u and φ; this regularity may be obtained under certain restrictive assumptions, including for instance $F \in L^\infty$, $a, \sigma \in L^\infty$. In that setting, there is no need to search for renormalized solutions: one reencounters the setting of weak solutions.

Acknowledgment. The authors wish to thanks the referees for useful comments and suggestions which led to improvements in the presentation of this paper.

REFERENCES

[1] R. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] W. Allegretto and H. Xie, *Existence of solutions for the time-dependent thermistor equations*, IMA J. Appl. Math., 48 (1992), pp. 271–281.
[3] W. Allegretto and H. Xie, *A non-local thermistor problem*, European J. Appl. Math., 6 (1993), pp. 83–94.
[4] S. N. Antontsev and M. Chipot, *The thermistor problem: Existence, smoothness uniqueness, blowup*, SIAM J. Math. Anal., 25 (1994), pp. 1128–1156.
[5] S. N. Antontsev and M. Chipot, *Analysis of blowup for the thermistor problem*, Siberian Math. J., 38 (1997), pp. 827–841.
[6] D. Blanchard, *Truncations and mononicity methods for parabolic equations*, Nonlinear Anal., 21 (1993), pp. 725–743.
[7] D. Blanchard and F. Murat, *Renormalized solutions of nonlinear parabolic problems with L^1–data: Existence and uniqueness*, Proc. Roy. Soc. Edinburgh Sect. A, 127 (1997), pp. 1137–1152.
[8] D. Blanchard and H. Redwane, *Renormalized solutions for a class of nonlinear evolution problems*, J. Math. Pures Appl., 77 (1998), pp. 117–151.
[9] L. Boccardo, I. Díaz, D. Giachetti and F. Murat, *Existence of a solution for a weaker form of a nonlinear elliptic equation*, Recent Advances in Nonlinear Elliptic and Parabolic Problems, Pitman Res. Notes Math. 208, Logman, Harlow, UK, 1989.
[10] L. Boccardo and T. Gallouët, *Nonlinear elliptic and parabolic equations involving measure data*, J. Funct. Anal., 87 (1989), pp. 149–169.
[11] L. Boccardo, F. Murat and J. P. Puel, *Existence results for some quasilinear parabolic equations*, Nonlinear Anal., 13 (1989), pp. 373–392.
[12] G. Cimatti, *Remark on existence and uniqueness for the thermistor problem under mixed boundary conditions*, Quatr. Appl. Math., 47 (1989), pp. 117–121.
[13] G. Cimatti, *Existence of weak solutions for the nonstationary problem of the Joule heating of a conductor*, Ann. Mat. Pura Appl., 162 (1992), pp. 33–42.
[14] M. Chipot and G. Cimatti, *A uniqueness result for the thermistor problem*, European J. Appl. Math., 2 (1991), pp. 97–103.
[15] R. J. DiPerna and P. L. Lions, *On the Fokker–Plank–Boltzman equation*, Comm. Math. Phys., 120 (1988), pp. 1–23.
[16] R. J. DiPerna and P. L. Lions, *On the Cauchy problem for Boltzman equations: Global existence and weak stability*, Ann. Math., 130 (1989) pp. 321–366.
[17] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1983.
[18] M. GómezÁrmol and F. Ortegón Gallego, *Existencia y unicidad de solución renormalizada para la ecuación de Kolmogórov*, in Actas de la Jornada Científica en homenaje al Prof. Antonio Valle Sánchez, Sevilla, 1997, pp. 167–183.
[19] M. GómezÁrmol, *Estudio Matemático de Algunos Problemas no Lineales de la Mecánica de Fluidos Incompresibles*, Ph.D. thesis, Universidad de Sevilla, Seville, Spain, 1998.
[20] M. T. González Montesinos, *Estudio Matemático de Algunos Problemas no Lineales del Electromagnetismo Relacionados con el Problema del Termistor*, Ph.D. thesis, Universidad de Cádiz, Cádiz, Spain, 2002.
[21] J. L. Lions, *Quelques Méthodes de Résolution des Problèmes aux Limites Non Linéaires*, Dunod, Paris, 1969.
[22] F. Murat, *Soluciones renormalizadas de ecuaciones en derivadas parciales elípticas no lineales*, Publications du laboratoire d'Analyse Numérique Paris VI, R 93023, Cours a l'Université de Sevilla, Seville, Spain, 1993.
[23] F. Murat, *Équations elliptiques non linéaires avec second membre L^1 ou mesure*, Comptes rendus du 26ème Congrès national d'analyse numérique, Les Karellis, France, 1994.
[24] J. Simon, *Compact sets in the space $L^p(0,T;B)$*, Ann. Mat. Pura Appl., 146 (1987), pp. 65–96.
[25] X. Xu, *A degenerate Stefan–like problem with Joule's heating*, SIAM J. Math. Anal., 23 (1992), pp. 1417–1438.
[26] X. Xu, *A strongly degenerate system involving an equation of parabolic type and an equation of elliptic type*, Comm. Partial Differential Equations, 18 (1993), pp. 199–213.
[27] X. Xu, *The thermistor problem with conductivity vanishing for large temperature*, Proc. Roy. Soc. Edinburgh Sect. A, 124 (1994), pp. 1–21.

SIAM J. MATH. ANAL.
Vol. 36, No. 6, pp. 2004–2019

EXISTENCE OF ENERGY MINIMIZERS FOR MAGNETOSTRICTIVE MATERIALS*

PIOTR RYBKA† AND MITCHELL LUSKIN‡

Abstract. The existence of a deformation and magnetization minimizing the magnetostrictive free energy is given. Mathematical challenges are presented by a free energy that includes elastic contributions defined in the reference configuration and magnetic contributions defined in the spatial frame. The one-to-one a.e. and orientation-preserving property of the deformation is demonstrated, and the satisfaction of the nonconvex saturation constraint for the magnetization is proven.

Key words. calculus of variations, magnetostriction, micromagnetics

AMS subject classifications. 49J45, 74B20, 74F15, 74G65

DOI. 10.1137/S0036141004442021

1. Introduction.

A mathematical model for magnetostrictive materials, in which the deformation and the magnetization are coupled, has been given in [4, 19, 20, 22, 8]. More significant shape change can be obtained from magnetostrictive materials that also undergo a structural phase transformation since the material can then lower its energy by an increase in the volume fraction of the martensitic variant with magnetic easy axis aligned with the applied magnetic field [11, 21]. This shape change is being utilized in emerging applications in actuators, sensors, and micromachines because they can provide a large *work output*/(*cycle* · *volume*) [25].

In this paper, we prove the existence of a deformation and magnetization minimizing the magnetostrictive free energy [4, 19, 20]. A novel feature of this free energy is that the elastic free energy is given in a reference configuration, as usual for elasticity, but the magnetic energies are given in the spatial frame. This requires that we use a free energy for which the Jacobian of the deformation can be controlled. We also must prove that the deformation is one-to-one a.e. and that the magnetization satisfies the nonconvex saturation constraint.

The existence of low energy phase and variant interfaces in single crystal thin films that do not exist in bulk martensitic crystals offers the possibility to develop thin materials with substantially larger strains than are possible in bulk [2]. Magnetostrictive, shape memory, single crystal, thin films such as Ni_2MnGa have recently been grown [11, 21]. We hope that this work will provide the foundation for our future work on the derivation of a rigorous magnetostrictive thin film energy to extend the results developed for micromagnetics [17] and martensitic deformation [2, 7, 6, 5].

The plan of our work is the following. In section 2, we introduce the magnetostriction model and in particular we present the constitutive assumptions. We shall

*Received by the editors March 11, 2004; accepted for publication (in revised form) September 27, 2004; published electronically June 30, 2005.
http://www.siam.org/journals/sima/36-6/44202.html

†Institute of Applied Mathematics and Mechanics, Warsaw University, Banacha 2, 02-097 Warsaw, Poland (rybka@mimuw.edu.pl). This author's work was supported in part by KBN 2P03A 042 24 and by a visiting position at the University of Minnesota.

‡School of Mathematics, University of Minnesota, 206 Church Street SE, Minneapolis, MN 55455 (luskin@math.umn.edu). This author's work was supported in part by NSF DMS-0074043 and DMS-0304326, by AFOSR F49620-98-1-0433, by the Institute for Mathematics and Its Applications, and by the Minnesota Supercomputer Institute.

seek the minimizers in an admissible set $\mathcal{A}$, described in section 2, that incorporates all of the constraints. To prove the existence of minimizers, we demonstrate in Theorem 3.1 that $\mathcal{A}$ is closed in a weak topology. We then show in section 4 that $\mathcal{E}$ is lower semicontinuous on $\mathcal{A}$, and hence that it attains its minimum. This is the content of Theorem 4.2 and is the main result of this paper.

2. The magnetostriction model. We consider a magnetostrictive crystal that in an undeformed state occupies the Lipschitz domain $\Omega \subset \mathbb{R}^3$. The admissible deformations $\bar{y} : \Omega \to \mathbb{R}^3$ of the crystal will be required to have the regularity $y \in W^{2,2}(\Omega;\mathbb{R}^3)$, to be orientation-preserving ($\det \nabla y > 0$ a.e. [18]), and to be one-to-one a.e., where we recall that a mapping $y : \Omega \to \mathbb{R}^3$ is *one-to-one a.e.* if there is a set $G \subset \Omega$ of full measure (that is, $|G| = |\Omega|$) such that $y(x)$ restricted to G is one-to-one. The one-to-one property everywhere of $y(x)$ would seem more appropriate; however, it is the weaker property that we will prove is preserved under weak convergence in $W^{2,2}(\Omega;\mathbb{R}^3)$.

The demonstration that energy-minimizing deformations obtained from the calculus of variations are one-to-one often requires sophisticated techniques [28], such as for problems with cavitation [27] or other problems with low regularity. Since our problem has higher regularity, simpler methods based on Banach's indicatrix are sufficient.

Because of the continuity of $y \in W^{2,2}(\Omega;\mathbb{R}^3)$, it follows that the sets $y(\bar{\Omega})$ and $y(\partial\bar{\Omega})$ are closed and the deformed domain $\mathcal{O}(y) := y(\bar{\Omega}) \setminus y(\partial\bar{\Omega})$ is open. Here and in what follows the bar over a set denotes its closure. We note that $\mathcal{O}(y)$ differs from $y(\Omega)$ on a set of measure zero [14] and $|\mathcal{O}(y)| = |y(\bar{\Omega})|$.

We wish to model a crystal that is attached on a nonempty, open subset of its boundary, $\Gamma \subset \partial\Omega$, so we will assume that admissible deformations $y(x)$ satisfy the boundary condition

$$y(x) = y_0(x) \qquad \text{for all } x \in \Gamma, \tag{2.1}$$

where $y_0 : \bar{\Omega} \to \mathbb{R}^3$ is a C^2 diffeomorphism with positive Jacobian.

The magnetization $m(z)$ of the crystal is naturally defined in spatial coordinates by $m : \mathcal{O}(y) \to \mathbb{R}^3$ and admissible magnetizations will be required to have the regularity $m \in W^{1,2}(\mathcal{O}(y);\mathbb{R}^3)$ which is equivalent to

$$\int_{\mathcal{O}(y)} \left[|\nabla_z m(z)|^2 + |m(z)|^2\right] dz = \int_\Omega \left[|\nabla_z m(y(x))|^2 + |m(y(x))|^2\right] \det \nabla y(x)\, dx < \infty.$$

We will often find it convenient as above to consider the magnetization $m \circ y : \Omega \to \mathbb{R}^3$ described in material coordinates. We will assume that the crystal is at a fixed temperature below the Curie temperature so that

$$|m(y(x))| \det \nabla y(x) = \tau, \qquad x \in \Omega, \tag{2.2}$$

where τ, the saturation magnetization, is a positive constant depending on the temperature.

The applied magnetic field will be given in spatial coordinates by $h : \mathbb{R}^3 \to \mathbb{R}^3$, and we will assume that $h \in L^2(\mathbb{R}^3;\mathbb{R}^3)$.

The free energy of a magnetostrictive crystal can be modeled by [20, 4],

$$\begin{aligned}\mathcal{E}(y,m) &= \int_\Omega \left\{\kappa \left|D^2 y(x)\right|^2 + \Phi\left(\nabla y(x), m\circ y(x)\right)\right\} dx \\ &\quad + \int_{\mathcal{O}(y)} \left\{\alpha \left|\nabla_z m(z)\right|^2 - h(z)\cdot m(z)\right\} dz + e_{\mathrm{mag}}(y,m) \\ &= \int_\Omega \Big\{\kappa \left|D^2 y(x)\right|^2 + \Phi\left(\nabla y(x), m\circ y(x)\right) \\ &\quad + \Big(\alpha \left|\nabla_z m\left(y(x)\right)\right|^2 - h(y(x))\cdot m(y(x))\Big) \det \nabla y(x)\Big\} dx + e_{\mathrm{mag}}(y,m),\end{aligned} \tag{2.3}$$

where the magnetostatic energy $e_{\mathrm{mag}}(y,m)$ is calculated from the magnetic scalar potential $\zeta : \mathbb{R}^3 \to \mathbb{R}$ by

$$e_{\mathrm{mag}}(y,m) = \frac{1}{2}\int_{\mathbb{R}^3} \left|\nabla_z \zeta(z)\right|^2 dz, \tag{2.4}$$

and where the magnetic scalar potential ζ satisfies

$$\mathrm{div}_z\left(-\nabla_z \zeta + \chi_{\mathcal{O}(y)} m\right) = 0, \qquad z \in \mathbb{R}^3. \tag{2.5}$$

We use the definition

$$\int_\Omega \left|D^2 y(x)\right|^2 dx = \int_\Omega \left[\sum_{i,j=1}^3 \left|\frac{\partial^2 y(x)}{\partial x_i \partial x_j}\right|^2\right] dx,$$

and we recall above that $\chi_{\mathcal{O}(y)}(z)$ is the characteristic function of $\mathcal{O}(y)$. The terms in (2.3) represent, from left to right, the surface energy [2, 26, 23], the anistropy energy, the exchange energy [17, 9], the interaction energy due to the applied magnetic field, and the magnetostatic energy. The parameters α and κ are positive material constants depending on the fixed temperature. The anisotropy energy density $\Phi(F,m)$ is a continuous function of the deformation gradient $F \in \mathbb{R}^{3\times 3}_+$ (where $\mathbb{R}^{3\times 3}_+$ denotes the group of 3×3 matrices with positive determinant) and the magnetization $m \in \mathbb{R}^3$. (Since the temperature is assumed to be fixed, we do not explicitly denote the dependence of the anisotropy energy density $\Phi(F,m)$ on temperature.)

We will assume that the anisotropy free energy density $\Phi \in C^2(\mathbb{R}^{3\times 3}_+ \times \mathbb{R}^3; \mathbb{R})$ is of the form

$$\Phi(F,m) = W(F,m) + \psi(\det F), \tag{2.6}$$

where $\psi : (0,\infty) \to \mathbb{R}$ is continuous, convex, and satisfies for

$$q > 2 \qquad \text{and} \qquad c_L > 0$$

the growth conditions

$$c_L(a^{-q} + a^q) \le \psi(a) \quad \text{for all } 0 < a < +\infty, \tag{2.7}$$

and where $W : \mathbb{R}^{3\times 3}_+ \times \mathbb{R}^3 \to \mathbb{R}$ is continuous and satisfies for

$$2 \le r < 6 \qquad \text{and} \qquad 0 < C_L < C_U$$

the growth conditions

$$C_L(|F|^2-1) \le W(F,m) \le C_U(|F|^r+1) \quad \text{for all } F \in \mathbb{R}^{3\times 3}_+ \text{ and } \ m \in \mathbb{R}^3. \tag{2.8}$$

We shall define the set of admissible functions to be

$$\mathcal{A} = \Big\{ (y,m) \in W^{2,2}(\Omega;\mathbb{R}^3) \times W^{1,2}(\mathcal{O}(y);\mathbb{R}^3) : y(x) = y_0(x) \text{ for all } x \in \Gamma,$$
$$\psi(\det \nabla y) \in L^1(\Omega),\ \det \nabla y > 0 \text{ a.e., and } y \text{ is one-to-one a.e.} \Big\},$$

where the growth properties of ψ were given in (2.7). We note that $\mathcal{A}$ is nonempty since we assumed in (2.1) that $y_0 : \bar{\Omega} \to \mathbb{R}^3$ is a C^2 diffeomorphism with positive Jacobian. We also note that $\mathcal{A}$ is not an affine space.

We shall show under the above assumptions that the problem

$$\min\{\mathcal{E}(y,m) : (y,m) \in \mathcal{A} \text{ and } |m(y(x))| \det \nabla y(x) = \tau \text{ for almost all } x \in \Omega\} \tag{2.9}$$

has a solution. We shall see that the terms

$$\int_{\mathcal{O}(y)} |\nabla_z m(z)|^2 \, dz = \int_\Omega |\nabla_z m|^2 \det \nabla y \, dx \quad \text{and} \quad \int_\Omega \psi(\det \nabla y)\, dx$$

in $\mathcal{E}$ require the most care in the analysis.

The anisotropy energy density for magnetostrictive crystals that undergo a structural phase transformation can have the form (2.6). To see this, we note that the anisotropy energy density for magnetostrictive crystals such as Ni_2MnGa is minimized at temperatures below the martensitic transformation on the wells [20, 21]

$$\mathcal{M} = \mathrm{SO}(3)(U_1,m_1) \cup \mathrm{SO}(3)(U_1,-m_1) \cup \cdots \cup \mathrm{SO}(3)(U_N,m_N) \cup \mathrm{SO}(3)(U_N,-m_N),$$

where $\det U_1 > 0$ and where for the symmetry group $\mathcal{G} \subset \mathrm{SO}(3)$ of the high temperature phase we have

$$\{(U_1,m_1),\dots,(U_N,m_N)\} = \{(QU_1Q^T, Qm_1) : Q \in \mathcal{G}\}. \tag{2.10}$$

We note that SO(3) is the group of proper rotations and

$$\mathrm{SO}(3)(U_k,\pm m_k) \equiv \{(RU_k, \pm Rm_k) : R \in \mathrm{SO}(3)\} \qquad \text{for } k = 1,\dots,N.$$

If $W(F,m)$ satisfies the property of frame indifference

$$W(RF,Rm) = W(F,m) \qquad \text{for all } F \in \mathbb{R}^{3\times 3}_+,\ m \in \mathbb{R}^3,\ R \in \mathrm{SO}(3),$$

and the property of material symmetry for the group, $\mathcal{G}$,

$$W(FQ,m) = W(F,m) \qquad \text{for all } F \in \mathbb{R}^{3\times 3}_+,\ m \in \mathbb{R}^3,\ R \in \mathcal{G},$$

then $\Phi(F,m)$ satisfies the property of frame indifference and material symmetry by the invariance of the determinant function. If $W(F,m)$ is minimized on the wells (2.10), that is,

$$W(\hat{F},\hat{m}) < W(F,m) \qquad \text{for all } (\hat{F},\hat{m}) \in \mathcal{M} \text{ and } (F,m) \in \mathbb{R}^{3\times 3}_+ \times \mathbb{R}^3 \setminus \mathcal{M}$$

and $\psi(a)$ is minimized at $\det U_1$, then $\Phi(F,m)$ is also minimized on the wells (2.10).

By rescaling the crystal domain Ω onto a domain of unit size, we can see that the dimensionless size of the surface energy coefficient (κ) and the exchange energy coefficient (α) become large for small crystals [9, 26, 17]. For the large crystal limit [9], the scaled micromagnetic energy converges to a phase theory which is the relaxation of the micromagnetics energy without exchange energy ($\alpha = 0$).

Since energy-minimizing solutions of (2.3) without surface energy ($\kappa = 0$) and exchange energy ($\alpha = 0$) do not generally exist, microstructures defined by minimizing sequences are studied in [20]. The existence of minimizers for a two-dimensional model without surface energy ($\kappa = 0$) and exchange energy ($\alpha = 0$) when the deformation is constrained on the boundary to be affine, $y(x) = Fx$, for $F \in \mathbb{R}^{2\times 2}_+$ in the lamination convex hull of a $\mathcal{M}$ with two wells ($N = 2$) was given in [10] using the method of convex integration.

3. Compactness and closure in the set of admissible functions. In order to demonstrate the existence of minimizers, we will use the direct method of the calculus of variations. For this purpose, we will show that weak limits of elements of $\mathcal{A}$ belong to $\mathcal{A}$. To be precise, we say that the sequence $\{(y_n, m_n)\} \subset \mathcal{A}$ *converges weakly* to $(y, m) \in \mathcal{A}$ if and only if

$$y_n \rightharpoonup y \quad \text{in } W^{2,2}(\Omega;\mathbb{R}^3),$$

and

$$\begin{aligned} \chi_{\mathcal{O}(y_n)} m_n &\rightharpoonup \chi_{\mathcal{O}(y)} m \qquad \text{in } L^2(\mathbb{R}^3;\mathbb{R}^3),\\ \chi_{\mathcal{O}(y_n)} \nabla_z m_n &\rightharpoonup \chi_{\mathcal{O}(y)} \nabla_z m \quad \text{in } L^2(\mathbb{R}^3;\mathbb{R}^{3\times 3}). \end{aligned}$$

It is possible to explain the weak convergence defined above within the general framework of Cartesian currents (see [15]). However, since y_n, m_n enjoy so much smoothness, we shall be content with easier, more direct methods.

It is convenient to introduce another definition. Namely, we say that a sequence $\{(y_n, m_n)\} \subset \mathcal{A}$ is $\mathcal{A}$-*bounded* if there exists a positive constant K, independent of n, such that

(3.1)
$$\int_\Omega \left\{ |D^2 y_n(x)|^2 + |\nabla y_n(x)|^2 + \psi(\det \nabla y_n(x)) \right\} dx + \int_{\mathcal{O}(y_n)} |\nabla_z m_n(z)|^2 \, dz \le K.$$

We now state the main technical result of this section.

THEOREM 3.1. *If a sequence $\{(y_n, m_n)\} \subset \mathcal{A}$ is $\mathcal{A}$-bounded, then there exists a subsequence (not relabeled) and $(y, m) \in \mathcal{A}$ such that (y_n, m_n) converges weakly to (y, m).*

This result is fundamental for our considerations. It guarantees the existence of candidates for minimizers, which are weak limits of minimizing sequences. We shall divide the proof into a number of tasks. We shall first deal with $\{y_n\}$ prior to discussing $\{m_n\}$.

Our main technical tool will be the distribution function. We define

$$\begin{aligned} A^n_t &= \{x \in \Omega : \det \nabla y_n(x) < t\}, \quad t < 1,\\ B^n_t &= \{x \in \Omega : \det \nabla y_n(x) > t\}, \quad t > 1. \end{aligned}$$

LEMMA 3.2. *If the sequence $\{(y_n, m_n)\} \subset \mathcal{A}$ is $\mathcal{A}$-bounded, then*

$$|A^n_t| \le t^q c_L^{-1} K \quad \textit{and} \quad |B^n_t| \le t^{-q} c_L^{-1} K.$$

Proof. Our starting point is simply

$$|A_t^n| = \int_{A_t^n} 1\,dx = \int_{A_t^n} \frac{t}{t}\,dx$$
$$\leq t \int_{A_t^n} \frac{1}{\det \nabla y_n}\,dx \leq t \left(\int_{A_t^n} \frac{1}{(\det \nabla y_n)^q}\,dx \right)^{\frac{1}{q}} \cdot |A_t^n|^{1-\frac{1}{q}}.$$

Hence, we have by (2.7) that

$$|A_t^n| \leq t^q \int_{A_t^n} (\det \nabla y_n)^{-q}\,dx \leq t^q c_L^{-1} \int_{A_t^n} \psi(\det \nabla y_n)\,dx \leq t^q c_L^{-1} K$$

for $t < 1$.

The argument leading to the second estimate is similarly given by

$$|B_t^n| = \int_{B_t^n} 1\,dx = \int_{B_t^n} \frac{t}{t}\,dx$$
$$\leq t^{-1} \int_{B_t^n} \det \nabla y_n\,dx \leq t^{-1} \left(\int_{B_t^n} (\det \nabla y_n)^q\,dx \right)^{\frac{1}{q}} \cdot |B_t^n|^{1-\frac{1}{q}}.$$

Hence, we have by (2.7) that

$$|B_t^n| \leq t^{-q} \int_{B_t^n} (\det \nabla y_n)^q\,dx \leq t^{-q} c_L^{-1} \int_{B_t^n} \psi(\det \nabla y_n)\,dx \leq t^{-q} c_L^{-1} K$$

for $t > 1$. ∎

We next state and prove a main lemma on the convergence of $\{y_n\}$. We note that the convergence result for $\det \nabla y_n$ below is better than that implied by the Sobolev embedding and compactness theorem [1, 16, 13]. It is due to the growth condition (2.7).

LEMMA 3.3. *If the sequence $\{(y_n, m_n)\} \subset \mathcal{A}$ is $\mathcal{A}$-bounded, then there exists a subsequence (not relabeled) such that*

$$y_n \rightharpoonup y \quad \text{in} \quad W^{2,2}(\Omega; \mathbb{R}^3),$$

and

$$\det \nabla y_n \to \det \nabla y \quad \text{in} \quad L^p(\Omega) \quad \text{for } p < q, \tag{3.2}$$

$$\det \nabla y_n \rightharpoonup \det \nabla y \quad \text{in} \quad L^q(\Omega), \tag{3.3}$$

$$\int_\Omega \psi(\det \nabla y)\,dx < \infty, \tag{3.4}$$

$$\det \nabla y > 0 \quad \text{a.e.} \tag{3.5}$$

Proof. In order to show existence of a subsequence (not relabeled) $\{y_n\}_{n=1}^\infty$, which is weakly convergent in $W^{2,2}(\Omega; \mathbb{R}^3)$, it is sufficient to prove a uniform in n bound on $\|y_n\|_{W^{2,2}}$. Due to the $\mathcal{A}$-boundedness of $\{(y_n, m_n)\}$, it is enough to prove a uniform bound on $\|y_n\|_{L^2}$. This easily follows from the boundary condition (2.1), the Poincaré

inequality [1, 16, 13], and the $\mathcal{A}$-boundedness because we have

$$\begin{aligned}\int_\Omega |y_n(x)|^2\,dx &\le 2\int_\Omega |y_n(x)-y_0(x)|^2\,dx + 2\int_\Omega |y_0(x)|^2\,dx\\ &\le 2C(\Omega)\int_\Omega |\nabla\left(y_n(x)-y_0(x)\right)|^2\,dx + 2\int_\Omega |y_0(x)|^2\,dx\\ &\le C\left(\int_\Omega |\nabla y_n(x)|^2\,dx + 1\right) \le C(K+1).\end{aligned}$$

Since the sequence $\{y_n\}$ converges weakly in $W^{2,2}(\Omega;\mathbb{R}^3)$ it follows from the Sobolev embedding theorem [1] that there exists a subsequence such that $y_n \longrightarrow y$ in $W^{1,p}(\Omega;\mathbb{R}^3)$ for $p<6$ and also in $C^{0,\alpha}(\overline{\Omega};\mathbb{R}^3)$ for $0<\alpha<1-\frac{3}{6}=\frac{1}{2}$. Thus, for another subsequence

$$\begin{aligned}\nabla y_n &\longrightarrow \nabla y \qquad \text{a.e.},\\ \det\nabla y_n &\longrightarrow \det\nabla y \qquad \text{a.e.},\\ \det\nabla y_n &\longrightarrow \det\nabla y \qquad \text{in } L^p(\Omega) \quad \text{for } p<2.\end{aligned} \tag{3.6}$$

We now show that we can improve the convergence in (3.6). For that purpose, we are going to show that $\{(\det\nabla y_n)^p\}$ is equi-integrable for any $p<q$, that is, for any given $\epsilon>0$ there is a $\delta>0$ such that, if $V\subset\Omega$ satisfies $|V|<\delta$, then

$$\int_V (\det\nabla y_n)^p\,dx < \epsilon.$$

Indeed, for any $V\subset\Omega$ and $t>1$ we have that

$$\int_V (\det\nabla y_n)^p\,dx = \left(\int_{V\setminus B_t^n} + \int_{V\cap B_t^n}\right)(\det\nabla y_n)^p\,dx. \tag{3.7}$$

Due to the definition of B_t^n, we can see that

$$\begin{aligned}\int_V (\det\nabla y_n)^p\,dx &\le t^p\,|V\setminus B_t^n| + \left(\int_{V\cap B_t^n}(\det\nabla y_n)^q\,dx\right)^{\frac{p}{q}}\cdot |B_t^n|^{1-\frac{p}{q}}\\ &\le t^p|V| + c_L^{-p/q}\left(\int_{B_t^n}\psi\left(\det\nabla y_n\right)\,dx\right)^{\frac{p}{q}}\cdot\left(t^{-q}c_L^{-1}K\right)^{\left(1-\frac{p}{q}\right)}\\ &\le t^p|V| + c_L^{-1}Kt^{-q+p}.\end{aligned}$$

We can now take $t>1$ so large that the second term is less than $\frac{\varepsilon}{2}$. Then we choose δ so small that $t^p\delta<\frac{\varepsilon}{2}$, and the claim follows since $|V|<\delta$.

After we choose an a.e.-convergent subsequence, we deduce by Vitali's theorem that

$$(\det\nabla y_n)^p \to (\det\nabla y)^p \quad \text{ in } L^1(\Omega) \quad \text{for } p<q.$$

Furthermore, we can infer that for this subsequence [3, Theorem 1],

$$\det\nabla y_n \longrightarrow \det\nabla y \ \text{ in } \ L^p(\Omega) \quad \text{for } p<q.$$

Thus, (3.2) follows.

In order to deduce (3.3), we notice that due to (2.7) the sequence $\det \nabla y_n$ is bounded in $L^q(\Omega)$. Hence, it contains a subsequence converging weakly to g. By uniqueness of the limit it follows that $g = \det \nabla y$.

To prove (3.4), we observe that since $\det \nabla y_n \to \det \nabla y$ in $L^1(\Omega)$, we have that there exists a subsequence (not relabeled) such that $\det \nabla y_n \to \det \nabla y$ a.e. Since ψ is continuous, we have also that $\psi(\det \nabla y_n) \to \psi(\det \nabla y)$ a.e. Thus, since sequence $\{(y_n, m_n)\}$ is $\mathcal{A}$-bounded, we have by Fatou's lemma [12] that

$$\int_\Omega \psi(\det \nabla y)\, dx \le \liminf_{n\to\infty} \int_\Omega \psi(\det \nabla y_n)\, dx \le K < \infty.$$

Our next task is to show that (3.5) holds. For this purpose, we use again the technique of distribution functions. Let us define for $t < 1$ the set

$$A_t = \{x \in \Omega :\ \det \nabla y < t\}.$$

It is clear that (3.5) holds once we establish the estimate

$$|A_t| \le c t^q, \tag{3.8}$$

where c is a positive constant. Since

$$\det \nabla y = \det \nabla y_n + (\det \nabla y - \det \nabla y_n)$$

and $\det \nabla y_n > 0$ a.e., we have for all $n \in \mathbb{N}$ that

$$A_t \subset A^n_{2t} \cup \{x \in \Omega :\ |\det \nabla y(x) - \det \nabla y_n(x)| \ge t\} \equiv A^n_{2t} \cup E^n.$$

By Egorov's theorem, for any $t \in (0,1)$ there is a $V \subset \Omega$ such that $|V| < t^q$ and $\det \nabla y_n$ converges uniformly to $\det \nabla y$ on $\Omega \setminus V$. Hence,

$$\begin{aligned} |A_t| &\le |A^n_{2t}| + |E^n \cap V| + |E^n \setminus V| \\ &\le c_L^{-1} K 2^q t^q + t^q + |E^n \setminus V|. \end{aligned}$$

However, for fixed t and sufficiently large n, the set $E^n \setminus V$ is empty. Thus, (3.8) holds with $c = 1 + c_L^{-1} K 2^q$. Hence, (3.5) follows. ∎

We remark that the methods we have presented and the assumptions on ψ allow us to prove similar convergence statements for $(\det \nabla y_n)^{-1}$. Indeed, we have the following lemma.

LEMMA 3.4. *If $1 \le p < q$, then $(\det \nabla y_n)^{-1}$ converges to $(\det \nabla y)^{-1}$ in $L^p(\Omega)$.*

Proof. By previous results, $\det \nabla y_n$ converges a.e., and the limit $\det \nabla y$ is positive a.e. Hence, $(\det \nabla y_n)^{-1} \to (\det \nabla y)^{-1}$ a.e. Due to Vitali's convergence theorem, it is sufficient to check that the sequence $(\det \nabla y_n)^{-p}$ is equi-integrable. To prove this, we suppose that $\epsilon > 0$ is given. Then for $V \subset \Omega$ with $|V| \le \delta$, similar to argument following (3.7), we have that

$$\begin{aligned} \int_V (\det \nabla y_n)^{-p}\, dx &= \left(\int_{V \setminus A^n_t} + \int_{V \cap A^n_t} \right) (\det \nabla y_n)^{-p}\, dx \\ &\le t^{-p} |V| + \left(\int_{V \cap A^n_t} (\det \nabla y_n)^{-q}\, dx \right)^{\frac{p}{q}} |A^n_t|^{1 - \frac{p}{q}} \\ &\le t^{-p} |V| + c_L^{-1} K t^{q-p} \le \epsilon \end{aligned}$$

for a suitable choice of t and δ. ∎

Finally, we want to make sure that the limit mapping y is indeed one-to-one a.e.

LEMMA 3.5. *Let us suppose that $\Omega \subset \mathbb{R}^3$ is open, the deformations $y_n : \Omega \to \mathbb{R}^3$ are one-to-one a.e., and $\det \nabla y_n > 0$ a.e. We also assume that the sequence y_n converges weakly in $W^{2,2}(\Omega)$ to y, $\det \nabla y_n \to \theta$ in $L^1(\Omega)$, and $\theta > 0$ a.e. Then y is one-to-one a.e. (and obviously $\det \nabla y = \theta$).*

Proof. We need a characterization of invertibility a.e. which is easy to apply to the limit y. Let us recall for that purpose the notion of Banach's indicatrix

$$N(y, \Omega, z) = \#\{x \in \Omega : \ y(x) = z\},$$

where we restrict our attention to the continuous representative of y. Of course, we have that $y(\Omega) = \{z \in \mathbb{R}^3 : \ N(y, \Omega, z) \geq 1\}$.

We claim that y is one-to-one a.e. if and only if

$$|\{z \in \mathbb{R}^3 : \ N(y, \Omega, z) \geq 2\}| = 0.$$

Indeed, let us suppose that $y|_G$ is one-to-one and G is of full measure. Since y has the Lusin property (see [14], section 5.2), we deduce that $E = y(\Omega \setminus G)$ has measure zero. Moreover, $N(y, \Omega, z) \geq 2$ if and only if $z \in E$.

On the other hand, let us suppose that $E = \{z \in \mathbb{R}^3 : \ N(y, \Omega, z) \geq 2\}$ has measure zero. We set $G = \Omega \setminus y^{-1}(E)$. We must show that $|y^{-1}(E)| = 0$. By the area formula (see [14], Theorem 5.11), we can see that

$$\int_{y^{-1}(E)} \det \nabla y(x)\, dx = \int_E N(y, \Omega, z)\, dz = 0.$$

Since $\det \nabla y > 0$ a.e., we deduce that $|y^{-1}(E)| = 0$.

We shall show that

$$N(y, \Omega, z) \leq 1 \quad a.e.$$

Let us take $\phi \in C_0(\mathbb{R}^3)$ such that $\phi \geq 0$. Obviously, we obtain since y_n is one-to-one a.e. that

$$\int_\Omega \phi(y_n(x)) \det \nabla y_n(x)\, dx = \int_{y_n(\Omega)} \phi(z)\, dz \leq \int_{\mathbb{R}^3} \phi(z)\, dz.$$

Hence,

$$\int_\Omega \phi(y(x)) \det \nabla y(x)\, dx \leq \int_{\mathbb{R}^3} \phi(z)\, dz.$$

Furthermore, this inequality implies that $N(y, \Omega, z) \leq 1$ a.e. That is, y is one-to-one a.e. as desired. ∎

Our next task is to prove the convergence of the sequence of magnetization vectors $\{m_n\}$.

LEMMA 3.6. *If $\{y_n, m_n\} \subset \mathcal{A}$ is a $\mathcal{A}$-bounded sequence, then there exists $m \in W^{1,2}(\mathcal{O}(y); \mathbb{R}^3)$ such that a subsequence (not relabeled) converges weakly to (y, m).*

Proof. We shall apply again the technique of distribution functions. We define the set

$$D_t^n = \{x \in \Omega : |\nabla_z m_n \circ y_n(x)| > t\}.$$

We claim that

$$|D_t^n| \leq K c_L^{-\frac{1}{q+1}} t^{-2\frac{q}{q+1}}. \tag{3.9}$$

Indeed, by the Schwartz inequality

$$\begin{aligned}
|D_t^n| &= \int_{D_t^n} 1\,dx = \int_{D_t^n} \left(\frac{|\nabla m_n| \det{}^{\frac{1}{2}} \nabla y_n}{|\nabla m_n| \det{}^{\frac{1}{2}} \nabla y_n} \right) dx \\
&\le \left(\int_{D_t^n} |\nabla m_n|^2 \det \nabla y_n \, dx \right)^{1/2} \left(\int_{D_t^n} |\nabla m_n|^{-2} (\det \nabla y_n)^{-1} \, dx \right)^{1/2} \\
&\le \frac{1}{t} K^{1/2} \left(\int_{D_t^n} \frac{1}{\det \nabla y_n} \, dx \right)^{1/2} \\
&\le \frac{1}{t} K^{1/2} \left(\int_{D_t^n} \frac{1}{(\det \nabla y_n)^q} \, dx \right)^{\frac{1}{2q}} |D_t^n|^{\frac{1}{2}(1-\frac{1}{q})} \\
&\le \frac{1}{t} K^{1/2} \left(\int_{D_t^n} c_L^{-1} \psi \left(\det \nabla y_n \right) \, dx \right)^{\frac{1}{2q}} |D_t^n|^{\frac{1}{2}(1-\frac{1}{q})} \\
&\le \frac{1}{t} c_L^{-\frac{1}{2q}} K^{\frac{1}{2}+\frac{1}{2q}} |D_t^n|^{\frac{1}{2}(1-\frac{1}{q})}.
\end{aligned}$$

Hence, (3.9) follows.

We recall that $\mathcal{O}(y) = y(\bar{\Omega}) \backslash y(\partial \bar{\Omega})$ is open. We define the two families of sets $\mathcal{O}_\varepsilon(y) = \{z \in \mathcal{O}(y) : \operatorname{dist}(z, \partial \mathcal{O}(y)) > \varepsilon\}$ and $\mathcal{O}^\epsilon(y) = \{z \in \mathbb{R}^3 : \ \operatorname{dist}(z, \mathcal{O}(y)) \le \epsilon\}$.

We note that

$$\bigcup_{\varepsilon > 0} \mathcal{O}_\varepsilon(y) = \mathcal{O}(y).$$

For $\varepsilon > 0$ fixed, we have that $\mathcal{O}_\varepsilon(y) \subset \mathcal{O}(y_n)$ for sufficiently large n because of the uniform convergence of y_n to y. We now claim that for any $\varepsilon > 0$ there exists a subsequence (not relabeled) such that

$$m_n \rightharpoonup m \quad \text{in} \quad W^{1,2}(\mathcal{O}_\varepsilon(y); \mathbb{R}^3). \tag{3.10}$$

We have for any $\varepsilon > 0$ that

$$\int_{\mathcal{O}_\varepsilon(y)} |\nabla_z m_n(z)|^2 \, dz \le \int_{\mathcal{O}(y_n)} |\nabla_z m_n(z)|^2 \, dz = \int_\Omega |\nabla_z m_n|^2 \det \nabla y_n \, dx \le K.$$

Due to the constraint (2.2), we can see that

$$\begin{aligned}
\int_{\mathcal{O}(y_n)} |m_n(z)|^2 \, dz &= \int_\Omega |m_n(y_n(x))|^2 \det \nabla y_n(x) \, dx \\
&= \tau^2 \int_\Omega (\det \nabla y_n(x))^{-1} \, dx \\
&\le \tau^2 \left(\int_\Omega (\det \nabla y_n(x))^{-q} \, dx \right)^{\frac{1}{q}} |\Omega|^{1-\frac{1}{q}} \\
&\le \tau^2 c_L^{-\frac{1}{q}} \left(\int_\Omega \psi(\det \nabla y_n(x)) \, dx \right)^{\frac{1}{q}} |\Omega|^{1-\frac{1}{q}} \le K_1.
\end{aligned}$$

This immediately implies that there exists a subsequence (not relabeled) such that

$$m_n \rightharpoonup m \quad \text{in} \quad W^{1,2}(\mathcal{O}_\varepsilon(y);\mathbb{R}^3).$$

It thus follows that there exists a further subsequence (not relabeled) such that

$$m_n \to m \quad \text{in} \quad L^2(\mathcal{O}_\varepsilon(y);\mathbb{R}^3). \tag{3.11}$$

We shall now show that

$$\chi_{\mathcal{O}(y_n)} m_n \to \chi_{\mathcal{O}(y)} m \text{ in } L^2(\mathbb{R}^3;\mathbb{R}^3) \text{ and } \chi_{\mathcal{O}(y_n)} \nabla_z m_n \rightharpoonup \chi_{\mathcal{O}(y)} \nabla_z m \text{ in } L^2(\mathbb{R}^3;\mathbb{R}^{3\times 3}).$$

We first estimate $\chi_{\mathcal{O}(y_n)} m_n - \chi_{\mathcal{O}(y)} m$ in $L^2(\mathbb{R}^3;\mathbb{R}^3)$ by observing that
$\chi_{\mathcal{O}(y_n)} m_n - \chi_{\mathcal{O}(y)} m = (\chi_{\mathcal{O}(y_n)} - \chi_{\mathcal{O}_\epsilon(y)}) m_n + \chi_{\mathcal{O}_\epsilon(y)}(m_n - m) + (\chi_{\mathcal{O}_\epsilon(y)} - \chi_{\mathcal{O}(y)}) m.$

Thus,

$$\begin{aligned}
&\|\chi_{\mathcal{O}(y_n)} m_n - \chi_{\mathcal{O}(y)} m\|_{L^2(\mathbb{R}^3)} \\
&\quad \le \|m_n\|_{L^2(\mathcal{O}(y_n)\Delta\mathcal{O}_\epsilon(y))} + \|m_n - m\|_{L^2(\mathcal{O}_\epsilon)} + \|m\|_{L^2(\mathcal{O}(y)\setminus\mathcal{O}_\epsilon(y))} \\
&\quad = I + II + III.
\end{aligned}$$

To estimate I, we set $\Omega_n^\epsilon := y_n^{-1}(\mathcal{O}_\epsilon(y))$. We can see using (2.2) and (2.7) that for $t < 1$

$$\begin{aligned}
I^2 &= \int_\Omega (1 - \chi_{\Omega_n^\epsilon})^2 |m_n|^2 \det \nabla y_n \, dx \\
&= \left(\int_{A_t^n} + \int_{\Omega \setminus A_t^n} \right) \frac{(1 - \chi_{\Omega_n^\epsilon})^2 \tau^2}{\det \nabla y_n} \, dx \\
&\le \tau^2 \left(\int_{A_t^n} \left(\frac{1}{\det \nabla y_n} \right)^q dx \right)^{\frac{1}{q}} |A_t^n|^{1-\frac{1}{q}} + \frac{\tau^2}{t} |\Omega \setminus (\Omega_n^\epsilon \cup A_t^n)| \\
&\le c t^{q-1} + \frac{\tau^2}{t} |\Omega \setminus (\Omega_n^\epsilon \cup A_t^n)|.
\end{aligned}$$

We first choose $t < 1$ to make the first term small, that is, less than $\frac{1}{2}(\delta/3)^2$. We then show that we can select ϵ so that the second term is less than $\frac{1}{2}(\delta/3)^2$. This would imply that $I < \delta/3$, as desired. We can do so because

$$\begin{aligned}
|\mathcal{O}^\epsilon(y) \setminus \mathcal{O}_\epsilon(y)| &\ge |y_n(\Omega \setminus \Omega_n^\epsilon)| \ge |y_n(\Omega \setminus (\Omega_n^\epsilon \cup A_t^n))| \\
&= \int_{\Omega \setminus (\Omega_n^\epsilon \cup A_t^n)} \det \nabla y_n \ge t |\Omega \setminus (\Omega_n^\epsilon \cup A_t^n)|.
\end{aligned}$$

Our claim follows because $|\mathcal{O}^\epsilon(y) \setminus \mathcal{O}_\epsilon(y)|$ can be made arbitrarily small, for fixed t.

For fixed $\epsilon > 0$ and for sufficiently large n, we have that $II < \delta/3$ because of (3.11). Finally, for given $\delta > 0$ one can find $\epsilon > 0$ for which $III < \delta/3$ because $|\mathcal{O}(y) \setminus \mathcal{O}_\epsilon(y)|$ can be made arbitrarily small, and integration is absolutely continuous with respect to the set of integration.

The weak convergence is slightly easier. For $\varphi \in L^2(\mathbb{R}^3;\mathbb{R}^{3\times 3})$, we consider

$$\begin{aligned}
&\int_{\mathbb{R}^3} (\chi_{\mathcal{O}(y_n)}\nabla_z m_n - \chi_{\mathcal{O}(y)}\nabla_z m)\cdot\varphi\,dz \\
&\quad = \int_{\mathbb{R}^3} \Big[(\chi_{\mathcal{O}(y_n)} - \chi_{\mathcal{O}_\epsilon(y)})\nabla_z m_n\cdot\varphi \\
&\qquad\qquad + \chi_{\mathcal{O}_\epsilon(y)}(\nabla_z m_n - \nabla_z m)\cdot\varphi + (\chi_{\mathcal{O}_\epsilon(y)} - \chi_{\mathcal{O}(y)})\nabla_z m\cdot\varphi\Big]\,dz \\
&\quad = J + JJ + JJJ.
\end{aligned}$$

We can see that

$$|J| \le \|\nabla_z m_n\|_{L^2(\mathcal{O}(y_n))}\|\varphi\|_{L^2(\mathcal{O}_(y_n)\setminus\mathcal{O}_\epsilon(y))}.$$

We can make $|J| < \delta/3$ by taking $\epsilon > 0$ small enough. The second term can be made small due to (3.10). Moreover, it is clear that $|JJJ| < \delta/3$ for sufficiently small $\epsilon > 0$. ∎

Finally, we shall demonstrate that the term $\int_\Omega |\nabla m_n|^2 \det\nabla y_n\,dx$ is lower semicontinuous.

LEMMA 3.7. *If $\{(y_m, m_n)\} \subset \mathcal{A}$ is an $\mathcal{A}$-bounded sequence, then*

$$\int_\Omega |\nabla_z m|^2 \det\nabla y\,dx \le \liminf_{n\to\infty}\int_\Omega |\nabla_z m_n|^2\det\nabla y_n\,dx.$$

Proof. If we take any $\delta > 0$, then there exists $\varepsilon > 0$ such that

$$\int_{\mathcal{O}_\varepsilon(y)} |\nabla_z m|^2\,dz \ge \int_{\mathcal{O}(y)} |\nabla_z m|^2\,dz - \delta.$$

It then follows from (3.10) that

$$\begin{aligned}
&\liminf_{n\to\infty}\int_\Omega |\nabla_z m_n|^2\det\nabla y_n\,dx \\
&\quad = \liminf_{n\to\infty}\int_{\mathcal{O}(y_n)} |\nabla_z m_n|^2\,dz \ge \liminf_{n\to\infty}\int_{\mathcal{O}_\varepsilon(y)} |\nabla_z m_n|^2\,dz \\
&\quad \ge \int_{\mathcal{O}_\varepsilon(y)} |\nabla_z m|^2\,dz \ge \int_{\mathcal{O}(y)} |\nabla_z m|^2\,dz - \delta = \int_\Omega |\nabla_z m|^2\det\nabla y\,dx - \delta.
\end{aligned}$$

The proof follows since δ was arbitrary. ∎

We are now ready for the proof of Theorem 3.1.

Proof of Theorem 3.1. If $\{(y_n, m_n)\} \subset \mathcal{A}$ is an $\mathcal{A}$-bounded sequence (3.1) with bound K, then by Lemmas 3.2–3.7 its subsequence converges weakly to an element (y, m) in $\mathcal{A}$ and

$$\int_\Omega \left\{|D^2 y|^2 + |\nabla y|^2 + \psi(\det\nabla y)\right\}\,dx + \int_{\mathcal{O}(y)} |\nabla_z m|^2\,dz \le K. \qquad ∎$$

4. The existence of an energy minimizer. We are going to demonstrate a lower semicontinuity property for the energy $\mathcal{E}$. We begin with a simple observation on minimizing sequences.

LEMMA 4.1. *If $\{(y_n, m_n)\} \subset \mathcal{A}$ is a minimizing sequence, then it is $\mathcal{A}$-bounded.*

Proof. We can estimate the magnetic interaction energy term, which is the only nonpositive expression in $\mathcal{E}$, by

$$\begin{aligned} J_n &= \left| \int_{\mathcal{O}(y_n)} h(z) \cdot m_n(z)\, dz \right| = \left| \int_\Omega h(y_n(x)) \cdot m_n(y_n(x)) \det \nabla y_n(x)\, dx \right| \\ &\le C_\varepsilon \int_{\mathcal{O}(y_n)} h^2(z)\, dz + \varepsilon \int_\Omega |m_n(y_n(x))|^2 \det \nabla y_n\, dx \\ &\le C_\varepsilon \,\|h\|^2_{L^2(\mathbb{R}^3)} + \varepsilon\tau^2 \int_\Omega (\det \nabla y_n)^{-1}\, dx \\ &\le C_\varepsilon \,\|h\|^2_{L^2(\mathbb{R}^3)} + \varepsilon\tau^2 c_L^{-\frac{1}{q}} \left(\int_\Omega \psi(\det \nabla y_n)\, dx \right)^{\frac{1}{q}} |\Omega|^{\frac{q-1}{q}}. \end{aligned}$$

By Young's inequality, we have that

$$J_n \le C_\varepsilon \,\|h\|^2_{L^2(\mathbb{R}^3)} + \varepsilon\tau^2 c_L^{-\frac{1}{q}} \left(\frac{1}{q} \int_\Omega \psi(\det \nabla y_n)\, dx + \frac{q-1}{q} |\Omega| \right).$$

Hence, due to (2.8) we have that

$$\begin{aligned} &\int_\Omega \left[\kappa |D^2 y_n|^2 + |\nabla y_n|^2 + \psi(\det \nabla y_n) + |\nabla_z m_n \circ y_n|^2 \det \nabla y_n \right] dx \\ &\quad \le \max\left\{1, \frac{1}{C_L}\right\} \mathcal{E}(y_n, m_n) + J_n + |\Omega| \\ &\quad \le K + |\Omega| + C_\varepsilon \,\|h\|^2_{L^2(\mathbb{R}^3)} + \varepsilon\tau^2 c_L^{-\frac{1}{q}} \left(\frac{1}{q} \int_\Omega \psi(\det \nabla y_n)\, dx + \frac{q-1}{q} |\Omega| \right). \end{aligned}$$

Since ε can be made arbitrarily small and since we have assumed that $h \in L^2(\mathbb{R}^3)$, our claim follows. ∎

We are thus ready to prove the main result of this paper, the existence of a solution to (2.9).

THEOREM 4.2. *Suppose that the magnetostrictive free energy $\mathcal{E}$ is given by* (2.3) *and that the growth assumptions* (2.7) *and* (2.8) *hold. Then the minimum free energy satisfying the saturation constraint* (2.2) *is attained.*

Proof. We consider a minimizing sequence $\{(y_n, m_n)\} \subset \mathcal{A}$. By Lemma 4.1, $\{(y_m, m_n)\}$ is an $\mathcal{A}$-bounded sequence. Hence, by Theorem 3.1 there exists $(y, m) \in \mathcal{A}$ such that a subsequence $\{(y_{n_k}, m_{n_k})\}$ converges weakly to (y, m). It is sufficient for us to show that

$$\mathcal{E}(y, m) \le \liminf_{n\to\infty} \mathcal{E}(y_m, m_n).$$

We shall treat each term in $\mathcal{E}$ separately, because after choosing a suitable subsequence we may replace lim inf with lim.

Due to the lower semicontinuity of the norm, we see for the elastic surface energy that

$$\kappa \int_\Omega |D^2 y|^2\, dx \le \liminf_{n\to\infty} \kappa \int_\Omega |D^2 y_n|^2\, dx.$$

We recall from (2.6) that the anisotropy energy density $\Phi(\nabla y, m)$ is the sum $\Phi(\nabla y, m) = W(\nabla y, m) + \psi(\det \nabla y)$. We first show that

$$\lim_{n\to\infty} \int_\Omega W(\nabla y_n, m_n)\, dx = \int_\Omega W(\nabla y, m)\, dx. \tag{4.1}$$

Indeed, we know that $\nabla y_n \to \nabla y$ in $L^p(\Omega : \mathbb{R}^{3\times 3})$ for $p < 6$ and $m_n \circ y_n \to m \circ y$ a.e. Moreover, due to Lemma 3.4 we have that

$$|m_n \circ y_n| = \frac{\tau}{\det \nabla y_n} \to \frac{\tau}{\det \nabla y} = |m \circ y| \quad \text{in } L^1(\Omega).$$

Thus, by [3, Theorem 1], (there is no need to extract another subsequence) we conclude that $m_n \circ y_n \to m \circ y$ in $L^1(\Omega; \mathbb{R}^3)$. Thus, by the continuity of the Nemytskii operator [24]

$$L^r(\Omega) \times L^1(\Omega) \ni (\nabla y, m) \to W(\nabla y, m) \in L^1(\Omega),$$

where $r < 6$ is the growth factor for $W(F, m)$ given by (2.8), we deduce (4.1). We finally recall that we have proved that

$$\int_\Omega \psi(\det \nabla y)\, dx \le \liminf_{n\to\infty} \int_\Omega \psi(\det \nabla y_n)\, dx.$$

For the magnetic exchange energy, we have from Lemma 3.7 that

$$\alpha \int_\Omega |\nabla m|^2 \det \nabla y\, dx \le \liminf_{n\to\infty} \alpha \int_\Omega |\nabla m_n|^2 \det \nabla y_n\, dx.$$

We now turn to the magnetic interaction energy and observe that

$$\lim_{n\to\infty} \int_\Omega (h \circ y_n \cdot m_n \circ y_n) \det \nabla y_n\, dx = \int_\Omega (h \circ y \cdot m \circ y) \det \nabla y\, dx.$$

We recall that

$$\int_\Omega (h \circ y_n \cdot m_n \circ y_n) \det \nabla y_n\, dx = \int_{\mathcal{O}(y_n)} m_n(z) \cdot h(z)\, dz = \int_{\mathbb{R}^3} \chi_{y_n(\Omega)} m_n(z) \cdot h(z)\, dz.$$

Since $\chi_{y_n(\Omega)} m_n$ converges to $\chi_{y(\Omega)} m$ in $L^2(\mathbb{R}^3; \mathbb{R}^3)$, our claim follows.

We finally have to show the convergence of the magnetization energy $e_{mag}(y, m)$ given by (2.4). We note that for given y and m, the weak solution $\zeta \in \mathcal{H}(\mathbb{R}^3)$ of the magnetostatic equation (2.5) satisfies

$$\int_{\mathbb{R}^3} (-\nabla_z \zeta + \chi_{\mathcal{O}(y)} m) \nabla_z \eta\, dz = 0 \quad \text{for all } \eta \in \mathcal{H}(\mathbb{R}^3), \tag{4.2}$$

where

$$\mathcal{H}(\mathbb{R}^3) = \left\{ \zeta \in \mathcal{D}'(\mathbb{R}^3) :\ \nabla \zeta \in L^2,\ \int_{\mathbb{R}^3} \zeta(z)\, dz = 0 \right\}.$$

Since $\nabla_z \zeta$ for $\zeta \in \mathcal{H}(\mathbb{R}^3)$ is an $L^2(\mathbb{R}^3; \mathbb{R}^3)$ projection of $\chi_{\mathcal{O}(y)} m(z)$, we have that

$$\|\nabla_z \zeta\|_{L^2} \le \|\chi_{\mathcal{O}(y)} m\|_{L^2};$$

and since $\chi_{\mathcal{O}(y_n)} m_n \to \chi_{\mathcal{O}(y)} m$ in $L^2(\mathbb{R}^3; \mathbb{R}^3)$, we have

$$\lim_{n\to\infty} \frac{1}{2} \int_{\mathbb{R}^3} |\nabla_z \zeta_n|^2\, dz = \frac{1}{2} \int_{\mathbb{R}^3} |\nabla_z \zeta|^2\, dz.$$

We have to check that weak limits also satisfy the pointwise magnetic saturation constraint (2.2). If $(y_n, m_n) \subset \mathcal{A}$ is a minimizing sequence, then the magnetic saturation constraint (2.2) follows since we showed L^2 convergence of $m_n \circ y_n$ and $\det \nabla y_n$. Namely, we have that

$$\begin{aligned}\int_\Omega \left|(m_n \circ y_n) \det \nabla y_n - (m \circ y) \det \nabla y\right| dx \\ \leq \| m_n - m \|_{L^2} \| \det \nabla y_n \|_{L^2} + \| m \|_{L^2} \| \det \nabla y_n - \det \nabla y \|_{L^2}.\end{aligned}$$

Finally, combining all of the above results we conclude that

$$\mathcal{E}(y, m) \leq \lim_{n \to \infty} \mathcal{E}(y_n, m_n),$$

that is, $(y, m) \in \mathcal{A}$ is the desired minimum of $\mathcal{E}$. □

REFERENCES

[1] R. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] K. Bhattacharya and R. D. James, *A theory of thin films of martensitic materials with applications to microactuators*, J. Mech. Phys. Solids, 47 (1999), pp. 531–576.

[3] H. Brezis and E. Lieb, *A relation between pointwise convergence of functions and convergence of functionals*, Proc. Amer. Math. Soc., 88 (1983), pp. 486–490.

[4] W. F. Brown, Jr., *Micromagnetics*, Robert E. Krieger Publishing Co. Inc., Huntington, New York, 1978.

[5] P. Bělík, T. Brule, and M. Luskin, *On the numerical modeling of deformations of pressurized martensitic thin films*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 525–548.

[6] P. Bělík and M. Luskin, *A computational model for the indentation and phase transformation of a martensitic thin film*, J. Mech. Phys. Solids, 50 (2002), pp. 1789–1815.

[7] P. Bělík and M. Luskin, *A total-variation surface energy model for thin films of martensitic crystals*, Interfaces Free Bound., 4 (2002), pp. 71–88.

[8] A. De Simone and P. Podio-Guidugli, *Inertial and self-interactions in structured continua: liquid crystals and magnetostrictive solids*, Meccanica, 30 (1995), pp. 629–640.

[9] A. De Simone, *Energy minimizers for large ferromagnetic bodies*, Arch. Rational Mech. Anal., 125 (1993), pp. 99–143.

[10] A. De Simone and G. Dolzmann, *Existence of minimizers for a variational problem in two-dimensional nonlinear magnetoelasticity*, Arch. Rational Mech. Anal., 144 (1998), pp. 107–120.

[11] J. Dong, J. Xie, J. Lu, C. Adelmann, C. Palmstrom, J. Cui, Q. Pan, T. Shield, R. James, and S. McKernan, *Shape memory and ferromagnetic shape memory effects in single-crystal Ni_2MnGa thin films*, J. Appl. Phys., (2004).

[12] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, SIAM, Philadelphia, 1999.

[13] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, Studies in Advanced Mathematics, CRC Press, Boca Raton, FL, 1992.

[14] I. Fonseca and W. Gangbo, *Degree Theory in Analysis and Applications*, Clarendon Press, New York, 1995.

[15] M. Giaquita, G. Modica, and J. Souček, *Cartesian currents in the calculus of variations*, vol. I, II, Springer-Verlag, Berlin, 1998.

[16] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, New York, 1998.

[17] G. Gioia and R. D. James, *Micromagnetics of very thin films*, Proc. Roy. Soc. Lond., 453 (1997), pp. 213–223.

[18] M. E. Gurtin, *Topics in Finite Elasticity*, SIAM, Philadelphia, 1981.

[19] R. D. James and D. Kinderlehrer, *Frustration in ferromagnetic materials*, Contin. Mech. Thermodyn., 2 (1990), pp. 215–239.

[20] R. D. James and D. Kinderlehrer, *Theory of magnetostriction with application to $Tb_xDy_{1-x}Fe_2$*, Phil. Mag. B, 68 (1993), pp. 237–274.

[21] R. D. James and M. Wuttig, *Magnetostriction of martensite*, Phil. Mag. A, 77 (1998), pp. 1273–1299.

[22] D. Kinderlehrer, *Magnetoelastic interactions*, in Variational Methods for Discontinuous Structures (Como, 1994), Progr. Nonlinear Differential Equations Appl., 25, Birkhäuser, Basel, Switzerland, 1996, pp. 177–189.

[23] R. Kohn and S. Müller, *Surface energy and microstructure in coherent phase transitions*, Comm. Pure Appl. Math., 47 (1994), pp. 405–435.

[24] M. Krasnoselskii, P. Zabreiko, E. Pustylnik, and P. Sobolevskii, *Integral Operators in Spaces of Summable Functions*, Noordhoff, Leiden, The Netherlands, 1976.

[25] P. Krulevitch, A. P. Lee, P. B. Ramsey, J. C. Trevino, J. Hamilton, and M. A. Northrup, *Thin film shape memory microactuators*, J. MEMS, 5 (1996), pp. 270–282.

[26] M. Luskin, *On the computation of crystalline micorstructure.*, Acta Numer., 1996, Cambridge University Press, Cambridge, pp. 191–257.

[27] S. Müller and S. J. Spector, *An existence theory for nonlinear elasticity that allows for cavitation*, Arch. Rational Mech. Anal., 131 (1995), pp. 1–66.

[28] J. Sivaloganathan and S. J. Spector, *On the existence of minimizers with prescribed singular points in nonlinear elasticity*, J. Elasticity, 59 (2000), pp. 83–113.

SIAM J. MATH. ANAL.
Vol. 36, No. 6, pp. 2020–2032

A PREY-PREDATOR MODEL WITH HYSTERESIS EFFECT*

TOYOHIKO AIKI† AND EMIL MINCHEV‡

Abstract. This paper provides mathematical analysis of a system of nonlinear PDEs which describes a prey-predator model with hysteresis effect. Existence and uniqueness of solutions for the system under consideration are proved.

Key words. hysteresis, nonlinear PDEs, prey-predator model, subdifferential

AMS subject classifications. 35R70, 35K50, 37N25, 47J40

DOI. 10.1137/S0036141004440186

1. Introduction and biological motivation.

The present paper deals with a class of prey-predator models which could take into account the diffusive as well as the hysteresis effects in the evolution of the populations and is described by the following system of PDEs:

$$\sigma_t - (\lambda(u))_t - \kappa\Delta\sigma + \partial I_{u,v}(\sigma) \ni F(\sigma, u, v) \quad \text{in } Q, \tag{1}$$

$$u_t - \Delta u = h(\sigma, u, v) \quad \text{in } Q, \tag{2}$$

$$v_t - \Delta v = g(\sigma, u, v) \quad \text{in } Q, \tag{3}$$

where $T > 0$, $\Omega \subset R^N$ is a bounded domain with smooth boundary $\partial\Omega$, $Q = (0,T)\times\Omega$; $\kappa \geq 0$ is a constant; $\lambda : R \to R$, $F, h, g : R^3 \to R$, $f_*, f^* : R^2 \to R$ are given functions. We assume that $f_*, f^* \in C^2(R^2)$, $0 \leq f_* \leq f^* \leq 1$ on R^2, and all partial derivatives of first and second order of f_* and f^* are bounded on R^2. We denote by $I_{u,v}(\cdot)$ the indicator function of the interval $[f_*(u,v), f^*(u,v)]$, and $\partial I_{u,v}(\cdot)$ denotes the subdifferential of $I_{u,v}(\cdot)$. The subdifferential $\partial I_{u,v}(\sigma)$ is a set-valued mapping in our statement of the problem

$$\partial I_{u,v}(\sigma) = \begin{cases} \emptyset & \text{if } \sigma > f^*(u,v) \text{ or } \sigma < f_*(u,v), \\ [0, +\infty) & \text{if } \sigma = f^*(u,v) > f_*(u,v), \\ \{0\} & \text{if } f_*(u,v) < \sigma < f^*(u,v), \\ (-\infty, 0] & \text{if } \sigma = f_*(u,v) < f^*(u,v), \\ R & \text{if } \sigma = f^*(u,v) = f_*(u,v). \end{cases} \tag{4}$$

Equation (1) corresponds to the kinetics of the density of the quantity of food σ (for the prey), (2), (3) describe the evolution of the prey and evolution of the predator, respectively; here u and v are the densities of the prey and predator, respectively. A typical example from the population dynamics is the following system:

$$\sigma_t + au_t - \kappa\Delta\sigma + \partial I_{u,v}(\sigma) \ni 0 \quad \text{in } Q, \tag{5}$$

*Received by the editors January 27, 2004; accepted for publication (in revised form) August 30, 2004; published electronically July 18, 2005.
http://www.siam.org/journals/sima/36-6/44018.html

†Department of Mathematics, Faculty of Education, Gifu University, Yanagido 1-1, Gifu 501-1193, Japan (aiki@cc.gifu-u.ac.jp).

‡Department of Applied Physics, School of Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan (eminchev@hotmail.com, iac04002@kurenai.waseda.jp).
This author's research was supported by grant P04050 of the Japan Society for the Promotion of Science.

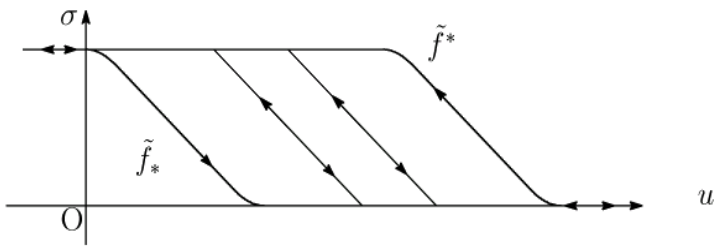

FIG. 1.

$$u_t - \Delta u = a_0 \sigma u - b_0 u v \quad \text{in } Q, \tag{6}$$

$$v_t - \Delta v = -c_0 v + d_0(1-\sigma)v \quad \text{in } Q, \tag{7}$$

where a, a_0, b_0, c_0, d_0 are positive constants. The system (5)–(7) could be considered a generalization of the classical prey-predator model allowing hysteresis relation between the prey/predator densities u, v and the density of the food quantity σ for the prey with vector input (u, v) and output σ. Our model originates from a prey-predator model of the type

$$\sigma = \lambda(u) \quad \text{in } Q, \tag{8}$$

$$u_t - \Delta u = h(\sigma, u, v) \quad \text{in } Q, \tag{9}$$

$$v_t - \Delta v = g(\sigma, u, v) \quad \text{in } Q, \tag{10}$$

in which the density of the food σ is determined by relation (8). We generalize this model in order to allow this relation to possibly depend also on the previous evolution data. More precisely, in our model the speed of change of density of food when the density of the prey decreases is different from the speed when the density of the prey increases. This situation can be described by the generalized stop operator shown in Figure 1, where $\tilde{f}^*$ and $\tilde{f}_*$ are upper and lower curves, respectively, of the hysteresis loop and a is the slope of the line in the loop. As a biologically consistent example for the constraint functions $\tilde{f}_*, \tilde{f}^*$ we can consider, for instance,

$$\tilde{f}^*(u) = \begin{cases} 1 & \text{if } u < 2, \\ -2.5u^2 + 10u - 9 & \text{if } 2 \le u < 2.2, \\ -u + 3.1 & \text{if } 2.2 \le u < 3, \\ 2.5u^2 - 16u + 25.6 & \text{if } 3 \le u < 3.2, \\ 0 & \text{if } 3.2 \le u, \end{cases} \quad \tilde{f}_*(u) = \begin{cases} 1 & \text{if } u < 0, \\ -2.5u^2 + 1 & \text{if } 0 \le u < 0.2, \\ -u + 1.1 & \text{if } 0.2 \le u < 1, \\ 2.5u^2 - 6u + 3.6 & \text{if } 1 \le u < 1.2, \\ 0 & \text{if } 1.2 \le u. \end{cases}$$

The biologically relevant assumption on the constraint functions reflected in the above example is that $\tilde{f}_*(0) = \tilde{f}^*(0)$ $(= 1)$ and $\tilde{f}_* = \tilde{f}^* = 0$ on $[u_c, \infty)$, which is due to the fact that the food σ should be constant $(= 1)$ if $u = 0$, and σ should keep zero value if the density u is bigger than some critical value $u_c > 0$.

Let us note that although there are indications for the existence of hysteresis in various biological problems (see, for example, [10], [13]), the mathematical treatment of biological problems with hysteresis has been considered only in a few papers; see

[7], as well as the survey paper [14]. The paper [7] seems to be the first to treat hysteresis phenomena in a biological problem. The authors of [7] treated bacterial growth in a petri dish modeled by a hysteresis operator of relay type which describes the relation between the rate of the growth of the bacterial population and the pH of the surrounding acid-buffer mix. The survey paper [14] treats applications of hysteresis in various natural phenomena. One of the chapters of [14] is devoted to the applications in biological problems and the authors of [14] also note the necessity of developing new models describing hysteresis effect in biological processes. The lack of papers dealing with hysteresis in biological phenomena could be explained with the diversity of the processes involved in mathematical biology. Let us note that many biological problems involve a fold catastrophe regime which, as is shown in [11], could be replaced by hysteresis model. Also there are various biological processes whose state variables change due to change of parameters in such a way that when the parameters go back to the old values the system does not follow its steps in return and thus a hysteresis loop is formed.

It is known that some types of hysteresis operators can be represented by ordinary differential inclusion containing subdifferential of the indicator function of a closed interval (whose length could possibly depend on the unknown variables). This fact was already pointed out by Visintin in [16]. Now we give a brief explanation of the fact that a function σ is determined by the hysteresis operator in Figure 1 if and only if σ is the solution of the differential inclusion

$$(*) \qquad \sigma_t + au_t + \partial I_u(\sigma) \ni 0 \quad \text{in } Q,$$

where u is some given function, $\tilde{f}_*$ and $\tilde{f}^*$ are Lipschitz continuous functions on R and $0 \le \tilde{f}_* \le \tilde{f}^* \le 1$ on R, and $-a \le \tilde{f}_*' \le 0$ and $-a \le \tilde{f}^{*'} \le 0$ on R. Indeed, since the solution of the inclusion is unique it is sufficient to show that σ defined by Figure 1 is a solution of (*). If $\tilde{f}_* < \sigma < \tilde{f}^*$, then $\sigma_t = au_t$ so that (*) holds. If $\tilde{f}_*(u) = \sigma$, then $u_t \ge 0$ and $\sigma_t = u_t \tilde{f}_*'(u) \ge -au_t$. This implies that $-\sigma_t - au_t \le 0$ and (*) holds. It is easy to see that (*) holds also if $\tilde{f}^*(u) = \sigma$.

This characterization of hysteresis operators was used for analysis of many nonlinear phenomena; for example, a real-time control problem (see [8]), solid-liquid phase transition (see [6], [9]), and shape memory alloy (see [1], [2], [3]). Also, in [9] it is explained that by using the inclusion (*) various types of hysteresis operators can be described. However, to the best of our knowledge this approach to hysteresis phenomena has not been used to problems from population dynamics.

In the model under consideration we allow also a small diffusive effect ($0 \le \kappa \ll 1$) for the food of the prey σ (from a biological point of view the diffusive effect for σ (for example, a plant occupying the domain Ω) is almost negligible with respect to other terms).

In the present paper we obtain results for positivity, boundedness, existence, and uniqueness of solutions of the prey-predator model with hysteresis effect (1)–(3). Using the method of Yosida approximation combined with derivation of appropriate uniform bounds, we prove that there exists at least one solution of the system under consideration (1)–(3). Furthermore, uniqueness of solutions is obtained in the case when $N \le 3$.

2. Preliminary notes. Denote by H the Hilbert space $L^2(\Omega)$ with the usual scalar product $(\cdot, \cdot)$ and norm $|\cdot|_H$. Denote by V the Sobolev space $H^1(\Omega)$ equipped with the norm $|u|_V = (u,u)_V^{1/2}$, where $(u,v)_V = (u,v) + a(u,v)$, $a(u,v) = \int_\Omega \nabla u(x) \cdot \nabla v(x) dx$, $u, v \in V$. Let $A : V \to V'$ be a linear continuous operator defined by

$\langle Au, v\rangle = a(u, v)$, $u, v \in V$, where V' is the dual space of V and $\langle\cdot,\cdot\rangle$ stands for the duality pairing between V' and V. For simplicity, in what follows we will denote the supremum of a bounded function by $|\cdot|_\infty$. Define the operator $-\Delta_N : D(-\Delta_N)(\subset H) \to H$ by the restriction of A to the elements $w \in V$ such that $Aw \in H$, i.e., $D(-\Delta_N) = \{w \in H^2(\Omega) : \ \frac{\partial w}{\partial n} = 0 \text{ in } H^{1/2}(\partial\Omega)\}$ and $-\Delta_N w = -\Delta w$ for all $w \in D(-\Delta_N)$, where $\frac{\partial}{\partial n}$ is the outward normal derivative on $\partial\Omega$.

DEFINITION 2.1. *Let $\kappa \geq 0$. A triplet of functions $\{\sigma, u, v\}$ is called a solution of the system* (1)–(3) *if*

(i) $\sigma \in W^{1,2}(0,T;H)\cap L^\infty(0,T;V)\cap L^2(0,T;H^2(\Omega))$ if $\kappa > 0$ and $\sigma \in W^{1,2}(0,T;H)$ if $\kappa = 0$.

(ii) $u, v \in W^{1,2}(0,T;H) \cap L^\infty(0,T;V) \cap L^2(0,T;H^2(\Omega))$.

(iii) $\sigma' - (\lambda(u))' - \kappa\Delta_N\sigma + \partial I_{u,v}(\sigma) \ni F(\sigma, u, v)$ in H a.e. in $(0,T)$.

(iv) $u' - \Delta_N u = h(\sigma, u, v)$ in H a.e. in $(0,T)$.

(v) $v' - \Delta_N v = g(\sigma, u, v)$ in H a.e. in $(0,T)$.

(vi) $\sigma(0) = \sigma_0, \ \ u(0) = u_0, \ \ v(0) = v_0$.

For simplicity, we denote, respectively, by σ', u', and v' the time-derivatives σ_t, u_t, and v_t of σ, u, and v. Note that the inclusion (iii) implies the following:

(iii)(a) $f_*(u, v) \leq \sigma \leq f^*(u, v)$ a.e. in Q.

(iii)(b) $(\sigma'(t) - (\lambda(u))'(t) - \kappa\Delta\sigma(t) - F(\sigma(t), u(t), v(t)), \sigma(t) - z) \leq 0$ for all $z \in H$ with $f_*(u(t), v(t)) \leq z \leq f^*(u(t), v(t))$ a.e. in Ω for a.e. $t \in (0,T)$.

(iii)(c) $\frac{\partial\sigma(t)}{\partial n} = 0$ a.e. on $\partial\Omega$ for a.e. $t \in (0,T)$ if $\kappa > 0$.

Throughout the paper we suppose that the following assumptions hold:

H1. $\kappa \geq 0$ is a given constant; $\lambda \in C^2(R)$, λ', and λ'' are bounded functions on R.

H2. $f_*, f^* \in C^2(R^2)$, $0 \leq f_* \leq f^* \leq 1$ on R^2, and all partial derivatives of first and second order of f_* and f^* are bounded on R^2. We put $C_0 = \max\{|f_*|_{W^{2,\infty}(R^2)}, |f^*|_{W^{2,\infty}(R^2)}\}$.

H3. F, h, and g are Lipschitz continuous functions on R^3 (with a common Lipschitz constant M), and $h(\sigma, 0, v) = 0$ for $\sigma \in [0,1], v \in R$, $g(\sigma, u, 0) = 0$ for $\sigma \in [0,1], u \in R$.

H4. $\sigma_0, u_0, v_0 \in L^\infty(\Omega) \cap V$ and $u_0 \geq 0$, $v_0 \geq 0$, $f_*(u_0, v_0) \leq \sigma_0 \leq f^*(u_0, v_0)$ a.e. in Ω.

3. Main results.

3.1. Nonnegativity of solutions.

THEOREM 3.1. *Any solution $\{\sigma, u, v\}$ of* (1)–(3) *satisfies the estimate*

$$\sigma \geq 0, \ u \geq 0, \ v \geq 0 \quad \text{a.e. in } Q. \tag{11}$$

Proof. The estimate for σ follows from the constraint $0 \leq f_*(u, v) \leq \sigma \leq f^*(u, v) \leq 1$ a.e. in Q. Now we prove the estimate for u. We multiply both sides of (2) by $[-u]^+$ (the positive part of $-u$) and using the Lipschitz continuity of h, we obtain that

$$\frac{1}{2}\frac{d}{dt}\left|[-u]^+\right|_H^2 + a([-u]^+, [-u]^+) \leq M\left|[-u]^+\right|_H^2 \quad \text{a.e. in } (0,T).$$

Therefore, by integration and application of the Gronwall inequality we conclude that $u(t) \geq 0$ a.e. in Q. Analogously it can be proved that $v(t) \geq 0$ a.e. in Q. ∎

3.2. Boundedness of solutions.

THEOREM 3.2. *Any solution* $\{\sigma, u, v\}$ *of* (1)–(3) *satisfies the estimate*

$$|\sigma|_\infty, |u|_\infty, |v|_\infty \le M_0, \tag{12}$$

where $M_0 = \max\{1, k_1 e^{MT}\}$, $k_1 = \max\{|u_0|_\infty, |v_0|_\infty\}$.

Proof. First let us note that $|\sigma|_\infty \le 1$ by the definition. Now we prove the estimate for u. Define $p(t) = k_1 e^{Mt}$ for $t \in [0,T]$. We have in view of Theorem 3.1 that

$$(u-p)' - \Delta_N(u-p) = h(\sigma,u,v) - Mp \le M(u-p) \text{ a.e. in } (0,T)\times\Omega. \tag{13}$$

Multiplying both sides of inequality (13) by $[u-p]^+$, we obtain that

$$\frac{1}{2}\frac{d}{dt}\left|[u-p]^+\right|_H^2 \le M\left|[u-p]^+\right|_H^2 \quad \text{a.e. in } (0,T).$$

Thus integrating and applying Gronwall inequality, we conclude that $u(t) \le p(t) \le k_1 e^{MT} \le M_0$ a.e. in Q. Similarly, it can be proved that $v \le M_0$, a.e. in $(0,T)\times\Omega$. □

Remark 3.1. From Theorems 3.1 and 3.2 it follows that by restricting ourselves to the set $\{0 \le \sigma \le M_0,\ 0 \le u \le M_0,\ 0 \le v \le M_0\}$ (if necessary), we can assume without loss of generality that the functions F, h, and g are bounded and Lipschitz continuous on R^3.

Remark 3.2. The proofs of Theorems 3.1 and 3.2 could be easily adapted to the system (5)–(7) using the boundedness of σ.

3.3. Existence of solutions.

3.3.1. Approximate solutions. For $\sigma, u, v \in R$ denote by $\partial I^\mu_{u,v}$ the Yosida regularization of the subdifferential graph $\partial I_{u,v}$,

$$\partial I^\mu_{u,v}(\sigma) = \frac{1}{\mu}[\sigma - f^*(u,v)]^+ - \frac{1}{\mu}[f_*(u,v) - \sigma]^+.$$

Consider the following approximate system of PDEs:

$$\sigma_t - (\lambda(u))_t - \kappa\Delta\sigma + \partial I^\mu_{u,v}(\sigma) = F(\sigma,u,v) \quad \text{in } Q, \tag{14}$$

$$u_t - \Delta u = h(\sigma,u,v) \quad \text{in } Q, \tag{15}$$

$$v_t - \Delta v = g(\sigma,u,v) \quad \text{in } Q. \tag{16}$$

DEFINITION 3.3. *Let* $\kappa \ge 0$. *The triplet of functions* $\{\sigma_\mu, u_\mu, v_\mu\}$ *is said to be a solution of the system* (14)–(16) *if*

(i) $\sigma_\mu \in W^{1,2}(0,T;H) \cap L^\infty(0,T;V) \cap L^2(0,T;H^2(\Omega))$ if $\kappa > 0$ and $\sigma_\mu \in W^{1,2}(0,T;H)$ if $\kappa = 0$.

(ii) $u_\mu, v_\mu \in W^{1,2}(0,T;H) \cap L^\infty(0,T;V) \cap L^2(0,T;H^2(\Omega))$.

(iii) $\sigma_\mu' - (\lambda(u_\mu))' - \kappa\Delta_N\sigma_\mu + \partial I^\mu_{u_\mu,v_\mu}(\sigma_\mu) = F(\sigma_\mu,u_\mu,v_\mu)$ in H a.e. in $(0,T)$.

(iv) $u_\mu' - \Delta_N u_\mu = h(\sigma_\mu,u_\mu,v_\mu)$ in H a.e. in $(0,T)$.

(v) $v_\mu' - \Delta_N v_\mu = g(\sigma_\mu,u_\mu,v_\mu)$ in H a.e. in $(0,T)$.

(vi) $\sigma_\mu(0) = \sigma_0$, $u_\mu(0) = u_0$, $v_\mu(0) = v_0$.

LEMMA 3.4. *Let* $\{\sigma_\mu, u_\mu, v_\mu\}$ *be a solution of* (14)–(16). *Then*

$$u_\mu \geq 0,\ v_\mu \geq 0 \quad \textit{a.e. in } Q. \tag{17}$$

The proof of Lemma 3.4 is similar to the proof of Theorem 3.1.

LEMMA 3.5. *Let* $\{\sigma_\mu, u_\mu, v_\mu\}$ *be a solution of* (14)–(16). *Then*

$$|u_\mu|_\infty,\ |v_\mu|_\infty \leq M_0,$$

for the same constant M_0 *as in Theorem* 3.2.

The proof of Lemma 3.5 is similar to the proof of Theorem 3.2.

THEOREM 3.6. *There exists a constant* $\kappa_0 > 0$ *such that for* $0 < \kappa < \kappa_0$ *there exists at least one solution of the system* (1)–(3).

3.3.2. Proof of Theorem 3.6. By a result of Colli and Hoffmann [5], it follows that the approximate system (14)–(16) possesses a unique solution $\{\sigma_\mu, u_\mu, v_\mu\}$ for each $\mu > 0$. Indeed, the function $A(\sigma, u) = \sigma - \lambda(u)$ is Lipschitz continuous as well as the function $F(\sigma, u, v) - \partial I^\mu_{u,v}(\sigma)$ and $(A(\sigma_1, u) - A(\sigma_2, u), \sigma_1 - \sigma_2) \geq |\sigma_1 - \sigma_2|^2$ for all $\sigma_1, \sigma_2, u \in R$. Moreover, define

$$\Phi(\mathbf{U}) = \begin{cases} \frac{1}{2}a(u,u) + \frac{1}{2}a(v,v) & \text{if } u, v \in V, \\ +\infty & \text{otherwise,} \end{cases} \quad \text{for } \mathbf{U} = \begin{pmatrix} u \\ v \end{pmatrix} \in H \times H.$$

We have that Φ is a proper convex l.s.c. function on $H \times H$, and its subdifferential is $\begin{pmatrix} -\Delta_N u \\ -\Delta_N v \end{pmatrix}$. Thus, choosing $X = H \times H$ in Theorem 1 of [5], we conclude that there exists a unique solution of the approximate system (14)–(16).

Now, we will prove some uniform bounds for the triplets $\{\sigma_\mu, u_\mu, v_\mu\},\ \mu > 0$, that solve the equations

$$\sigma'_\mu - (\lambda(u_\mu))' - \kappa\Delta_N\sigma_\mu + \partial I^\mu_{u_\mu, v_\mu}(\sigma_\mu) = F(\sigma_\mu, u_\mu, v_\mu) \quad \text{in } H \text{ a.e. in } (0,T), \tag{18}$$

$$u'_\mu - \Delta_N u_\mu = h(\sigma_\mu, u_\mu, v_\mu) \quad \text{in } H \text{ a.e. in } (0,T), \tag{19}$$

$$v'_\mu - \Delta_N v_\mu = g(\sigma_\mu, u_\mu, v_\mu) \quad \text{in } H \text{ a.e. in } (0,T). \tag{20}$$

To this end we derive certain energy inequalities. Again for simplicity of the notation we will write $\{\sigma, u, v\}$ instead of $\{\sigma_\mu, u_\mu, v_\mu\}$. Now, we multiply (19) by u' and (20) by v'; adding together and applying Young's inequality, we obtain that

$$|u'|^2_H + |v'|^2_H + \frac{d}{dt}|\nabla u|^2_H + \frac{d}{dt}|\nabla v|^2_H \leq C_1 \quad \text{a.e. in } (0,T), \tag{21}$$

where $C_1 = (|h|^2_\infty + |g|^2_\infty)|\Omega|$, $|\Omega|$ denotes the Lebesgue measure of the set Ω. Multiply (19) by $-\Delta u$ and (20) by $-\Delta v$, and adding together, we conclude that

$$\frac{d}{dt}|\nabla u|^2_H + \frac{d}{dt}|\nabla v|^2_H + |\Delta u|^2_H + |\Delta v|^2_H \leq C_1 \quad \text{a.e. in } (0,T). \tag{22}$$

LEMMA 3.7. *Let* $\{\sigma, u, v\}$ *be a solution of* (14)–(16). *Then the function*

$$(I^\mu_{u,v}(\sigma))(t) = \frac{1}{2\mu}\left|[\sigma - f^*(u,v)]^+\right|^2_H + \frac{1}{2\mu}\left|[f_*(u,v) - \sigma]^+\right|^2_H$$

is absolutely continuous on $[0,T]$ *and*

$$\frac{d}{dt}I^{\mu}_{u,v}(\sigma) \le (\partial I^{\mu}_{u,v}(\sigma),\sigma') + C_0|\partial I^{\mu}_{u,v}(\sigma)|_H(|u'|_H+|v'|_H) \quad a.e.\ in\ (0,T).$$

The proof of Lemma 3.7 is similar to the proof of Lemma 4.1 of [6].

Now, we multiply (18) by σ' and using Lemma 3.7, we obtain that

(23)
$$|\sigma'|^2_H + \kappa\frac{d}{dt}|\nabla\sigma|^2_H + 2\frac{d}{dt}I^{\mu}_{u,v}(\sigma) \le C_2\left(|u'|^2_H+|v'|^2_H+\kappa^2|\Delta\sigma|^2_H+1\right) \text{ a.e. in } (0,T),$$

where $C_2=\max\{4C_0^2+C_0+3C(\lambda),2C_0|F|^2_\infty|\Omega|\}$, $C(\lambda)=\max\{|\lambda'|_\infty,|\lambda''|_\infty\}$.
Since $f_*(u,v), f^*(u,v)\in H^2(\Omega)$ a.e. in $(0,T)$, we have that

$$\begin{aligned}
&(\partial I^{\mu}_{u,v}(\sigma),-\Delta\sigma)\\
&=\left(\frac{1}{\mu}[\sigma-f^*(u,v)]^+,-\Delta(\sigma-f^*(u,v))\right)+\left(\frac{1}{\mu}[\sigma-f^*(u,v)]^+,-\Delta f^*(u,v)\right)\\
&+\left(\frac{1}{\mu}[f_*(u,v)-\sigma]^+,-\Delta(f_*(u,v)-\sigma)\right)+\left(\frac{1}{\mu}[f_*(u,v)-\sigma]^+,\Delta f_*(u,v)\right)\\
&=\frac{1}{\mu}|\nabla[\sigma-f^*(u,v)]^+|^2_H+\left(\frac{1}{\mu}[\sigma-f^*(u,v)]^+,-\Delta f^*(u,v)\right)\\
&+\frac{1}{\mu}|\nabla[f_*(u,v)-\sigma]^+|^2_H+\left(\frac{1}{\mu}[f_*(u,v)-\sigma]^+,\Delta f_*(u,v)\right)\\
&\ge -\frac{1}{4\mu^2}\left\{|[\sigma-f^*(u,v)]^+|^2_H+|[f_*(u,v)-\sigma]^+|^2_H\right\}-|\Delta f^*(u,v)|^2_H-|\Delta f_*(u,v)|^2_H
\end{aligned}$$

$$\ge -\frac{1}{4}|\partial I^{\mu}_{u,v}(\sigma)|^2_H-|\Delta f^*(u,v)|^2_H-|\Delta f_*(u,v)|^2_H. \tag{24}$$

Also

$$(F(\sigma,u,v),-\Delta\sigma)\le C^*_F(|\nabla\sigma|^2_H+|\nabla u|^2_H+|\nabla v|^2_H), \tag{25}$$

where $C^*_F=\left|\frac{\partial F}{\partial\sigma}\right|_\infty+\left|\frac{\partial F}{\partial u}\right|_\infty+\left|\frac{\partial F}{\partial v}\right|_\infty$.
Now, multiplying (18) by $-\Delta\sigma$, we get in view of (24) and (25) that

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}|\nabla\sigma|^2_H+\kappa|\Delta\sigma|^2_H+((\lambda(u))',\Delta\sigma)\\
&\le\frac{1}{4}|\partial I^{\mu}_{u,v}(\sigma)|^2_H+|\Delta f^*(u,v)|^2_H+|\Delta f_*(u,v)|^2_H+C^*_F(|\nabla\sigma|^2_H+|\nabla u|^2_H+|\nabla v|^2_H).
\end{aligned}$$

Since $((\lambda(u))',\Delta\sigma)=-\frac{d}{dt}(\lambda'(u)\nabla u,\nabla\sigma)-(\Delta\lambda(u),\sigma')$, we conclude that

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}|\nabla\sigma|^2_H+\kappa|\Delta\sigma|^2_H-\frac{d}{dt}(\lambda'(u)\nabla u,\nabla\sigma)&\le\frac{1}{4}|\partial I^{\mu}_{u,v}(\sigma)|^2_H+|\Delta f^*(u,v)|^2_H+|\Delta f_*(u,v)|^2_H\\
&+\frac{1}{2}|\Delta\lambda(u)|^2_H+\frac{1}{2}|\sigma'|^2_H+C^*_F(|\nabla\sigma|^2_H+|\nabla u|^2_H+|\nabla v|^2_H).
\end{aligned} \tag{26}$$

Note that $\Delta\lambda(u)=\lambda''(u)|\nabla u|^2+\lambda'(u)\Delta u$ and consequently,

$$|\Delta\lambda(u)|^2_H\le 2C(\lambda)^2(|\nabla u|^4_{L^4}+|\Delta u|^2_H).$$

By the Gagliardo–Nirenberg inequality (cf. [17]), we have that

$$|\nabla u|^4_{L^4(\Omega)} \le C(\Omega)|u|^2_{H^2(\Omega)}|u|^2_\infty \le C(\Omega)(|u|^2_H + |\Delta u|^2_H)|u|^2_\infty \le C_3(1+|\Delta u|^2_H), \tag{27}$$

where the constant C_3 depends on M_0 (cf. Lemma 3.5). Thus, we obtain that

$$|\Delta\lambda(u)|^2_H \le C_4(1+|\Delta u|^2_H), \tag{28}$$

with $C_4 = 2C(\lambda)^2(C_3+1)$. Hence we have that

$$\frac{1}{2}\frac{d}{dt}|\nabla\sigma|^2_H + \kappa|\Delta\sigma|^2_H - \frac{d}{dt}(\lambda'(u)\nabla u, \nabla\sigma) \le \frac{1}{4}|\partial I^\mu_{u,v}(\sigma)|^2_H + |\Delta f^*(u,v)|^2_H + |\Delta f_*(u,v)|^2_H$$

$$+ \frac{1}{2}C_4(1+|\Delta u|^2_H) + \frac{1}{2}|\sigma'|^2_H + C^*_F(|\nabla\sigma|^2_H + |\nabla u|^2_H + |\nabla v|^2_H). \tag{29}$$

Now, we multiply (18) by $\partial I^\mu_{u,v}(\sigma)$ and using Lemma 3.7, we obtain the estimate

$$\frac{d}{dt}I^\mu_{u,v}(\sigma) + \frac{1}{2}|\partial I^\mu_{u,v}(\sigma)|^2_H \le C_5(|u'|^2_H + |v'|^2_H)$$

$$+ \frac{\kappa}{4}|\partial I^\mu_{u,v}(\sigma)|^2_H + \kappa|\Delta f^*(u,v)|^2_H + \kappa|\Delta f_*(u,v)|^2_H + C_6, \tag{30}$$

where $C_5 = (C_0+1)(C_0+C(\lambda)^2)$, $C_6 = (C_0+1)|F|^2_\infty|\Omega|$.

Noting that $|\Delta f_*(u,v)|^2_H + |\Delta f^*(u,v)|^2_H \le 2C_7(|\nabla u|^4_{L^4} + |\nabla v|^4_{L^4} + |\Delta u|^2_H + |\Delta v|^2_H)$, where $C_7 = 32C_0^2$, we conclude in view of (27) that

$$|\Delta f_*(u,v)|^2_H + |\Delta f^*(u,v)|^2_H \le C_8(1+|\Delta u|^2_H + |\Delta v|^2_H), \tag{31}$$

with $C_8 = 2C_7(C_3+1)$. Adding (29) and (30), we get in view of (31) that

$$\frac{d}{dt}\left\{I^\mu_{u,v}(\sigma) + \frac{1}{2}|\nabla\sigma|^2_H - (\lambda'(u)\nabla u, \nabla\sigma)\right\} + \kappa|\Delta\sigma|^2_H + \frac{1-\kappa}{4}|\partial I^\mu_{u,v}(\sigma)|^2_H$$

$$\le C_9(1 + |\nabla\sigma|^2_H + |\nabla u|^2_H + |\nabla v|^2_H + |\sigma'|^2_H$$

$$+ |u'|^2_H + |v'|^2_H + (1+\kappa)(1+|\Delta u|^2_H + |\Delta v|^2_H)), \tag{32}$$

where $C_9 = \max\left\{\frac{C_4}{2} + C_6, C^*_F, \frac{1}{2}, C_5, \frac{C_4}{2} + C_8\right\}$.

Let $\varepsilon_1, \varepsilon_2, \varepsilon_3$ be positive numbers to be specified later. Calculate (21) + $\varepsilon_1 \times$ (22) + $\varepsilon_2 \times$ (23) + $\varepsilon_3 \times$ (32). We have that

$$(1-\varepsilon_2C_2-\varepsilon_3C_9)|u'|^2_H + (1-\varepsilon_2C_2-\varepsilon_3C_9)|v'|^2_H$$

$$+ (\varepsilon_2-\varepsilon_3C_9)|\sigma'|^2_H + (\varepsilon_1-\varepsilon_3C_9(1+\kappa))|\Delta u|^2_H + (\varepsilon_1-\varepsilon_3C_9(1+\kappa))|\Delta v|^2_H$$

$$+ \kappa(\varepsilon_3-\varepsilon_2C_2\kappa)|\Delta\sigma|^2_H + \varepsilon_3\frac{1-\kappa}{4}|\partial I^\mu_{u,v}(\sigma)|^2_H$$

$$+ \frac{d}{dt}\{(1+\varepsilon_1)|\nabla u|^2_H + (1+\varepsilon_1)|\nabla v|^2_H + \left(\varepsilon_2\kappa + \frac{\varepsilon_3}{2}\right)|\nabla\sigma|^2_H$$

$$- \varepsilon_3(\lambda'(u)\nabla u, \nabla\sigma) + (2\varepsilon_2 + \varepsilon_3) I^{\mu}_{u,v}(\sigma)\}$$

$$\leq C_{10} + \varepsilon_3 C_9 \left(2 + \kappa + |\nabla u|_H^2 + |\nabla v|_H^2 + |\nabla\sigma|_H^2\right), \tag{33}$$

where $C_{10} = C_1 + \varepsilon_1 C_1 + \varepsilon_2 C_2$. Now, we will fix ε_i, $i = 1, 2, 3$, as well the constant κ_0 in the statement of the theorem, so that the coefficients in the first three lines of (33), i.e., $1 - \varepsilon_2 C_2 - \varepsilon_3 C_9$, $\varepsilon_2 - \varepsilon_3 C_9$, $\varepsilon_1 - \varepsilon_3 C_9(1+\kappa)$, $\kappa(\varepsilon_3 - \varepsilon_2 C_2 \kappa)$, $\varepsilon_3 \frac{1-\kappa}{4}$ will be all positive whenever $\kappa \in (0, \kappa_0)$. For instance, we can take $\varepsilon_2 = \frac{1}{4C_2}$, $\varepsilon_3 = \min\{\frac{1}{4C_9}\min\{1, \frac{1}{2C_2}\}, \frac{1}{C(\lambda)^2}\}$, and then $\varepsilon_1 = \min\{\min\{1, \frac{1}{2C_2}\}, \frac{4C_9}{C(\lambda)^2}\}$ with $\kappa_0 = \min\{\frac{1}{2}, \frac{1}{2C_9}\min\{1, \frac{1}{2C_2}\}, \frac{2}{C(\lambda)^2}\}$. We note that in this case

$$1 - \varepsilon_2 C_2 - \varepsilon_3 C_9 \geq \frac{1}{2}, \tag{34}$$

$$\varepsilon_2 - \varepsilon_3 C_9 \geq \frac{\varepsilon_2}{2} \tag{35}$$

along with

$$\varepsilon_1 - \varepsilon_3 C_9(1+\kappa) \geq \frac{\varepsilon_1}{2} \qquad \text{for all } \kappa \in (0, \kappa_0), \tag{36}$$

so that the above coefficients are all bounded from below uniformly with respect to κ. Moreover, concerning the coefficients in the third line of (33), we have that $\kappa(\varepsilon_3 - \varepsilon_2 C_2 \kappa) \geq \kappa\frac{\varepsilon_3}{2}$, $\varepsilon_3\frac{1-\kappa}{4} \geq \frac{\varepsilon_3}{8}$ for all $\kappa \in (0, \kappa_0)$.

Consequently, from (33) we can deduce uniform estimates for $\sigma_\mu = \sigma$, $u_\mu = u$, and $v_\mu = v$ with respect to the parameter μ. We have that $\{\sigma_\mu\}_\mu, \{u_\mu\}_\mu$, and $\{v_\mu\}_\mu$ are bounded in $W^{1,2}(0,T;H) \cap L^\infty(0,T;V) \cap L^2(0,T;H^2(\Omega))$, $\{\partial I^{\mu}_{u_\mu,v_\mu}(\sigma_\mu)\}_\mu$ is bounded in $L^2(0,T;H)$, and $\{I^{\mu}_{u_\mu,v_\mu}(\sigma_\mu)\}_\mu$ is bounded in $L^\infty(0,T)$.

Therefore (cf. [6], [8]), possibly by extracting a subsequence $\mu_n \searrow 0$, we conclude that $\sigma_{\mu_n} \to \sigma$, $u_{\mu_n} \to u$, and $v_{\mu_n} \to v$ weakly in $W^{1,2}(0,T;H) \cap L^2(0,T;H^2(\Omega))$ and weakly star in $L^\infty(0,T;V)$ to a triplet $\{\sigma, u, v\}$, which is a solution of the system (1)–(3). Let us note also that $f_*(u,v) \leq \sigma \leq f^*(u,v)$ a.e. in Q. □

Remark 3.3. Without loss of generality, in the proof of Theorem 3.6 it could be assumed that σ is also bounded. Thus, in view of Remarks 3.1 and 3.2 it follows that the proof of Theorem 3.6 can be easily adapted to the system (5)–(7) as well.

3.4. Uniqueness of solutions. The purpose of this section is to discuss the uniqueness of solutions. Before the proof of uniqueness we give an estimate for a solution of a parabolic equation. The estimate in the following lemma will play a very important role in the proof of uniqueness.

LEMMA 3.8 (cf. [12, Theorem 3.7.1]). *Let $\mu_0 > 0$ and θ be a solution of the following initial boundary value problem:*

$$\theta' - \mu_0 \Delta\theta = f \quad \text{in } Q, \tag{37}$$

$$\frac{\partial\theta}{\partial n} = 0 \quad \text{on } (0,T)\times\partial\Omega, \quad \theta(0) = \theta_0, \tag{38}$$

where f and θ_0 are given functions. If $f \in L^r(0,T;L^q(\Omega))$ with $\frac{1}{r} + \frac{N}{2q} < 1$ for $q, r \geq 1$ and $\theta_0 \in L^\infty(\Omega)$, then there exists a positive constant C_* depending only on Ω, μ_0, q, r, and N such that

$$|\theta|_{L^\infty(0,t;L^\infty(\Omega))} \leq C_*(|f|_{L^r(0,t;L^q(\Omega))} + |\theta_0|_{L^\infty(\Omega)}) \quad \text{for } 0 \leq t \leq T.$$

Proof. First, we assume that $|f|_{L^r(0,T;L^q(\Omega))} + |\theta_0|_{L^\infty(\Omega)} \le 1$. Let θ be a solution of (37)–(38). Then Theorem 3.7.1 of [12] implies that

$$|\theta|_{L^\infty(0,t;L^\infty(\Omega))} \le C_* \quad \text{for } 0 \le t \le T,$$

where C_* is a positive constant. Next, for general f and θ_0 we put $\hat{\theta} = \frac{\theta}{\ell_0}$, where θ is a solution of (37)–(38) and $\ell_0 = |f|_{L^r(0,t;L^q(\Omega))} + |\theta_0|_{L^\infty(\Omega)}$. Immediately, $\hat{\theta}$ is a solution of (37)–(38) with $\hat{f} = \frac{f}{\ell_0}$ and $\hat{\theta}_0 = \frac{\theta_0}{\ell_0}$ so that $|\hat{\theta}|_{L^\infty(0,t;L^\infty(\Omega))} \le C_*$ for $0 \le t \le T$ because $|\hat{f}|_{L^r(0,T;L^q(\Omega))} + |\hat{\theta}_0|_{L^\infty(\Omega)} \le 1$. Thus we can prove this lemma. □

In order to prove uniqueness we assume that $\Omega \subset R^3$.

THEOREM 3.9. *Let $\Omega \subset R^3$. Then the system* (1)–(3) *admits at most one solution.*

The main idea of the proof is due to Kenmochi, Koyama, and Meyer [8].

Proof. Let $\{\sigma_1, u_1, v_1\}$ and $\{\sigma_2, u_2, v_2\}$ be solutions of (1)–(3) in the sense of Definition 2.1. Also, we put $\sigma = \sigma_1 - \sigma_2$, $u = u_1 - u_2$, $v = v_1 - v_2$. For $s \in (0, T]$ we define

$$L(s) = \max\{|f_*(u_1, v_1) - f_*(u_2, v_2)|_{L^\infty(0,s;L^\infty(\Omega))}, |f^*(u_1, v_1) - f^*(u_2, v_2)|_{L^\infty(0,s;L^\infty(\Omega))}\}.$$

The proof is rather long. So, we divide it into several steps.

1*st step.*

$$\begin{aligned} &|[\sigma(t) - L(s)]^+|_H^2 + |[-\sigma(t) - L(s)]^+|_H^2 \\ &\le K_1 \exp\left\{\int_0^t (3 + |u_1'(\tau)|_H^2 + |u_2'(\tau)|_H^2) d\tau\right\} \\ &\quad \times \int_0^t (|u(\tau)|_H^2 + |v(\tau)|_H^2 + |\sigma(\tau)|_H^2 + |u(\tau)|_{L^\infty(\Omega)}^2 + |u'(\tau)|_H^2) d\tau \text{ for } t \in [0, s], \end{aligned} \tag{39}$$

where K_1 is a positive constant.

Proof of 1*st step.* We put $\tilde{\sigma}_1 = \sigma_1 - [\sigma - L(s)]^+, \tilde{\sigma}_2 = \sigma_2 + [\sigma - L(s)]^+$ a.e. on $(0, s) \times \Omega$. Easily, we have $f_*(u_i, v_i) \le \tilde{\sigma}_i \le f^*(u_i, v_i)$ for $i = 1, 2$, a.e. on $(0, s) \times \Omega$. Then, Definition 2.1(iii)(b) implies that

$$\begin{aligned} &(\sigma_1'(t), [\sigma(t) - L(s)]^+) + \kappa a(\sigma_1(t), [\sigma(t) - L(s)]^+) \\ &\le (F(u_1(t), v_1(t), \sigma_1(t)) + (\lambda(u_1))'(t), [\sigma(t) - L(s)]^+) \quad \text{for a.e. } t \in [0, s]. \end{aligned} \tag{40}$$

Also, we have

$$\begin{aligned} &-(\sigma_2'(t), [\sigma(t) - L(s)]^+) - \kappa a(\sigma_2(t), [\sigma(t) - L(s)]^+) \\ &\le -(F(u_2(t), v_2(t), \sigma_2(t)) + (\lambda(u_2))'(t), [\sigma(t) - L(s)]^+) \quad \text{for a.e. } t \in [0, s]. \end{aligned} \tag{41}$$

By adding (40) and (41), we obtain

$$\begin{aligned} &(\sigma'(t), [\sigma(t) - L(s)]^+) + \kappa a(\sigma(t), [\sigma(t) - L(s)]^+) \\ &\le (F(u_1(t), v_1(t), \sigma_1(t)) - F(u_2(t), v_2(t), \sigma_2(t)), [\sigma(t) - L(s)]^+) \\ &\quad + ((\lambda(u_1))'(t) - (\lambda(u_2))'(t), [\sigma(t) - L(s)]^+) \quad \text{for a.e. } t \in [0, s] \end{aligned}$$

so that

(42)
$$\begin{aligned}
&\frac{d}{dt}|[\sigma(t)-L(s)]^+|_H^2\\
&\le M(|u(t)|_H+|v(t)|_H+|\sigma(t)|_H)|[\sigma(t)-L(s)]^+|_H\\
&\quad+((\lambda'(u_1)(t)-\lambda'(u_2)(t))u_1'(t),[\sigma(t)-L(s)]^+)+(\lambda'(u_2)u'(t),[\sigma(t)-L(s)]^+)\\
&\le K_2(|u(t)|_H^2+|v(t)|_H^2+|\sigma(t)|_H^2+|u(t)|_{L^\infty(\Omega)}^2+|u'(t)|_H^2)\\
&\quad+(3+|u_1'(t)|_H^2)|[\sigma(t)-L(s)]^+|_H^2 \quad \text{for a.e. } t\in[0,s],
\end{aligned}$$

where $C(\lambda)=\max\{|\lambda''|_{L^\infty(R)},|\lambda'|_{L^\infty(R)}\}$ and $K_2=M^2+2C(\lambda)^2$. Applying the Gronwall inequality to (42) it holds that

$$\begin{aligned}
&|[\sigma(t)-L(s)]^+|_H^2\le K_2\exp\left\{\int_0^t(3+|u_1'(\tau)|_H^2)d\tau\right\}\\
&\times\int_0^t(|u(\tau)|_H^2+|v(\tau)|_H^2+|\sigma(\tau)|_H^2+|u(\tau)|_{L^\infty(\Omega)}^2+|u'(\tau)|_H^2)d\tau \quad \text{for } t\in[0,s].
\end{aligned}$$

We can obtain similar estimate for $[-\sigma(t)-L(s)]^+$. Hence, we get (39). □

2nd step. It holds that

$$\frac{1}{2}(|u(t)|_H^2+|v(t)|_H^2)\le 4M\int_0^t(|u(\tau)|_H^2+|v(\tau)|_H^2+|\sigma(\tau)|_H^2)d\tau,\quad t\in[0,T], \tag{43}$$

$$\int_0^t|u'(\tau)|_H^2d\tau\le 4M\int_0^t(|u(\tau)|_H^2+|v(\tau)|_H^2+|\sigma(\tau)|_H^2)d\tau,\quad t\in[0,T]. \tag{44}$$

The proof of this step is omitted since it is quite standard.

3rd step.

$$\begin{aligned}
&|u|_{L^\infty(0,t;L^\infty(\Omega))}\le C_*M(|u|_{L^s(0,t;H)}+|v|_{L^s(0,t;H)}+|\sigma|_{L^s(0,t;H)}),\\
&|v|_{L^\infty(0,t;L^\infty(\Omega))}\le C_*M(|u|_{L^s(0,t;H)}+|v|_{L^s(0,t;H)}+|\sigma|_{L^s(0,t;H)}) \text{ for } 0\le t\le T,
\end{aligned}$$

where C_* is a positive constant given by Lemma 3.8.

Proof of 3rd step. Lemma 3.8 guarantees that

$$|u|_{L^\infty(0,t;L^\infty(\Omega))}\le C_*|h(\sigma_1,u_1,v_1)-h(\sigma_2,u_2,v_2)|_{L^s(0,t;H)} \text{ for } 0\le t\le T.$$

Therefore, by using H3, we conclude that the assertion of the 3rd step is true. □

4th step. There exists a positive constant K_3 such that

$$|\sigma|_{L^\infty(0,s;H)}^2\le K_3s^{1/4}(|u|_{L^\infty(0,s;H)}^2+|v|_{L^\infty(0,s;H)}^2+|\sigma|_{L^\infty(0,s;H)}^2) \text{ for } 0\le s\le T. \tag{45}$$

Proof of 4th step. It is easy to see that

$$\begin{aligned}
&|\sigma|\le|[\sigma-L(s)]^+-[-\sigma-L(s)]^+-\sigma|+|[\sigma-L(s)]^+|+|[-\sigma-L(s)]^+|\\
&\le L(s)+|[\sigma-L(s)]^+|+|[-\sigma-L(s)]^+| \quad \text{a.e. on } (0,s)\times\Omega.
\end{aligned} \tag{46}$$

It follows from the definition of $L(s)$ and the 3rd step that

$$\begin{aligned}
&L(s)^2\le 4M_0^2(|u|_{L^\infty(0,s;L^\infty(\Omega))}^2+|v|_{L^\infty(0,s;L^\infty(\Omega))}^2)\\
&\le 24M_0^2C_*^2M^2s^{1/4}(|u|_{L^\infty(0,s;H)}^2+|v|_{L^\infty(0,s;H)}^2+|\sigma|_{L^\infty(0,s;H)}^2).
\end{aligned} \tag{47}$$

On account of (39), (46), (47), the 3rd step, and (44), we observe that

$$\begin{aligned}
&|\sigma|^2_{L^\infty(0,s;H)} \le K_3 s^{1/4}(|u|^2_{L^\infty(0,s;H)} + |v|^2_{L^\infty(0,s;H)} + |\sigma|^2_{L^\infty(0,s;H)})\\
&+K_4\int_0^s (|u(t)|^2_H + |v(t)|^2_H + |\sigma(t)|^2_H)dt\\
&+K_4\int_0^s (|u(t)|^2_{L^\infty(\Omega)} + |u'(t)|^2_H)dt\\
&\le K_4(1+T^{3/4}+C_*MT+4MT^{3/4})s^{1/4}\\
&\quad\times(|u|^2_{L^\infty(0,s;H)} + |v|^2_{L^\infty(0,s;H)} + |\sigma|^2_{L^\infty(0,s;H)}),
\end{aligned}$$

where $K_4 = 24M_0^2C_*^2M^2 + K_1\exp\{\int_0^T(3+|u_1'(\tau)|^2_H + |u_2'(\tau)|^2_H)d\tau\}$. Thus we have proved the 4th step. □

Proof of the uniqueness. By (43) we have

$$|u|^2_{L^\infty(0,s;H)}+|v|^2_{L^\infty(0,s;H)} \le 8Ms(|u|^2_{L^\infty(0,s;H)}+|v|^2_{L^\infty(0,s;H)}+|\sigma|^2_{L^\infty(0,s;H)}) \text{ for } 0\le s\le T.$$

This inequality together with (45) implies that

$$\begin{aligned}
&|u|^2_{L^\infty(0,s;H)} + |v|^2_{L^\infty(0,s;H)} + |\sigma|^2_{L^\infty(0,s;H)}\\
&\le (8MT^{3/4}+K_3)s^{1/4}(|u|^2_{L^\infty(0,s;H)} + |v|^2_{L^\infty(0,s;H)} + |\sigma|^2_{L^\infty(0,s;H)}) \text{ for } 0\le s\le T.
\end{aligned}$$

Here, we take $s_1 > 0$ satisfying $(8MT^{3/4}+K_3)s_1^{1/4} \le \frac{1}{2}$. Then we see that

$$|u|^2_{L^\infty(0,s_1;H)} + |v|^2_{L^\infty(0,s_1;H)} + |\sigma|^2_{L^\infty(0,s_1;H)} = 0,$$

that is, $u = v = \sigma = 0$ a.e. on $(0,s_1)\times\Omega$. The choice of s_1 is independent of initial values. Therefore, we can obtain the uniqueness of the solution. □

3.5. Existence and uniqueness in the case when $\Omega \subset R^3$. In this section we present an existence and uniqueness result for the case when $\Omega \subset R^3$.

THEOREM 3.10. *Let $\Omega \subset R^3$. Then* (i) *the system* (1)–(3) *with $\kappa = 0$ possesses a unique solution;* (ii) *there exists a constant $\kappa_0 > 0$ such that the system* (1)–(3) *possesses a unique solution for any κ satisfying $0 < \kappa < \kappa_0$.*

Proof. (i) Let us note that the validity of the estimates for the approximate solutions from section 3.3.2 extends to the constructed solution $\{\sigma^{(\kappa)}, u^{(\kappa)}, v^{(\kappa)}\}$ of the system (1)–(3) whenever $0 < \kappa < \kappa_0$ (cf. (33)–(36)). Therefore, in view of Theorem 3.9 and arguing as in [8], it could be shown that $\{\sigma^{(\kappa)}, u^{(\kappa)}, v^{(\kappa)}\}$ converges in a suitable sense to the unique solution $\{\sigma, u, v\}$ of the system (1)–(3) with $\kappa = 0$. Moreover, σ satisfies $\sigma \in L^\infty(0,T;V)$. (ii) is a corollary of Theorems 3.6 and 3.9. □

Acknowledgment. The authors express their gratitude to the referees for their valuable suggestions and comments.

REFERENCES

[1] T. AIKI, *One-dimensional shape memory alloy problems including a hysteresis operator*, Funkcial. Ekvac., 46 (2003), pp. 441–469.

[2] T. AIKI AND N. KENMOCHI, *Some models for shape memory alloys*, in Mathematical Aspects of Modeling Structure Formation Phenomena, GAKUTO Internat. Ser. Math. Sci. Appl. 17, Gakkotosho, Tokyo, 2002, pp. 144–162.

[3] T. Aiki and N. Kenmochi, *Models for shape memory alloys described by subdifferentials of indicator functions*, in Elliptic and Parabolic Problems (Rolduc and Gaeta 2001), World Scientific, River Edge, NJ, 2002, pp. 1–10.
[4] H. Brézis, *Operateurs Maximaux Monotones et Semi-Groupes de Contractions dans les Espaces de Hilbert*, North–Holland/American Elsevier, Amsterdam, London, New York, 1973.
[5] P. Colli and K. H. Hoffmann, *A nonlinear evolution problem describing multi-component phase changes with dissipation*, Numer. Funct. Anal. Optim., 14 (1993), pp. 275–297.
[6] P. Colli, N. Kenmochi, and M. Kubo, *A phase field model with temperature dependent constraint*, J. Math. Anal. Appl., 256 (2001), pp. 668–685.
[7] F. C. Hoppensteadt, W. Jäger, and C. Pöppe, *A hysteresis model for bacterial growth patterns,* in Modelling of Patterns in Space and Time, Lecture Notes in Biomath. 55, W. Jäger and J. D. Murray, eds., Springer-Verlag, Berlin, 1984, pp. 123–134.
[8] N. Kenmochi, T. Koyama, and G. H. Meyer, *Parabolic PDEs with hysteresis and quasivariational inequalities*, Nonlinear Anal., 34 (1998), pp. 665–686.
[9] N. Kenmochi, E. Minchev, and T. Okazaki, *Ordinary differential systems describing hysteresis effects and numerical simulations*, Abstr. Appl. Anal., 7 (2002), pp. 563–583.
[10] J.-P. Kernevez, G. Joly, M.-C. Duban, B. Bunow, and D. Thomas, *Hysteresis, oscillations, and pattern formation in realistic immobilized enzyme systems*, J. Math. Biol., 7 (1979), pp. 41–56.
[11] M. A. Krasnosel'skii and A. V. Pokrovskii, *Systems with Hysteresis*, Springer-Verlag, Berlin, 1989.
[12] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasi-Linear Equations of Parabolic Type*, Transl. Math. Monograph 23, AMS, Providence, RI, 1967.
[13] M. Landau, P. Lorente, J. Henry, and S. Canu, *Hysteresis phenomena between periodic and stationary solutions in a model of pacemaker and nonpacemaker coupled cardiac cells*, J. Math. Biol., 25 (1987), pp. 491–509.
[14] J. W. Macki, P. Nistri, and P. Zecca, *Mathematical models for hysteresis*, SIAM Rev., 35 (1993), pp. 94–123.
[15] I. D. Mayergoyz, *Mathematical Models of Hysteresis*, Springer-Verlag, New York, 1991.
[16] A. Visintin, *Differential Models of Hysteresis*, Springer-Verlag, Berlin, 1994.
[17] E. Zeidler, *Nonlinear Functional Analysis and its Applications.* II *A Linear Monotone Operators*, Springer-Verlag, New York, Berlin, 1990.

SIAM J. MATH. ANAL.
Vol. 36, No. 6, pp. 2033–2048

A UNIFORM ANALYSIS OF NONSYMMETRIC AND COERCIVE LINEAR OPERATORS*

GIANCARLO SANGALLI†

Abstract. In this work, we show how to construct, by means of the function space interpolation theory, a *natural* norm $|\!|\!| \cdot |\!|\!|$ for a generic linear *coercive* and *nonsymmetric* operator $\mathcal{L}$. The natural norm $|\!|\!| \cdot |\!|\!|$ allows for *continuity* and *inf-sup* conditions which hold independently of $\mathcal{L}$. In particular we will consider the convection-diffusion-reaction operator, for which we obtain continuity and inf-sup conditions that are uniform with respect to the operator coefficients. In this case, our results give some insight for the analysis of the singular perturbed behavior of the operator, occurring when the diffusivity coefficient is small. Furthermore, our analysis is preliminary to applying some recent numerical methodologies (such as *least-squares* and *adaptive wavelet* methods) to this class of operators, and more generally to analyzing any numerical method within the classical framework [I. Babuška and A. K. Aziz, in *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations*, Academic Press, New York, 1972, pp. 1–359].

Key words. advection-diffusion, singular perturbation, interpolation theory

AMS subject classifications. 35J20, 46B70

DOI. 10.1137/S0036141003434996

1. Introduction.

Consider the convection-diffusion-reaction linear operator

$$w \mapsto \mathcal{L}w := -\kappa \Delta w + \beta \cdot \nabla w + \rho w, \tag{1.1}$$

where the argument w is a function on the domain $\Omega \subset \mathbb{R}^n$, κ is a constant positive diffusion coefficient, $\beta : \Omega \to \mathbb{R}^n$ is a velocity field, and $\rho : \Omega \to \mathbb{R}$ is a reaction coefficient. Under suitable assumptions on the coefficients, e.g., $\rho - 1/2\,\mathrm{div}\beta \geq 0$, the operator $\mathcal{L}$ is an isomorphism from $V := H_0^1(\Omega)$ into $V^* := H^{-1}(\Omega)$. In fact, given a source term $f \in V^*$, the boundary value problem

$$\begin{cases} \mathcal{L}u = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega \end{cases} \tag{1.2}$$

admits a unique solution $u \in V$. Nevertheless, the norm of $\mathcal{L}$, as linear operator from $H_0^1(\Omega)$ into $H^{-1}(\Omega)$,

$$\|\mathcal{L}\|_{H_0^1(\Omega)\to H^{-1}(\Omega)} := \sup_{w\in H_0^1(\Omega)} \frac{\|\mathcal{L}w\|_{H^{-1}(\Omega)}}{\|w\|_{H_0^1(\Omega)}} = \sup_{w\in H_0^1(\Omega)} \sup_{v\in H_0^1(\Omega)} \frac{\langle \mathcal{L}w, v\rangle}{\|w\|_{H_0^1(\Omega)}\|v\|_{H_0^1(\Omega)}},$$

and the norm of its inverse $\mathcal{L}^{-1}$

$$\begin{aligned} \|\mathcal{L}^{-1}\|_{H^{-1}(\Omega)\to H_0^1(\Omega)} &:= \sup_{w\in H_0^1(\Omega)} \frac{\|w\|_{H_0^1(\Omega)}}{\|\mathcal{L}w\|_{H^{-1}(\Omega)}} \\ &= \left(\inf_{w\in H_0^1(\Omega)} \sup_{v\in H_0^1(\Omega)} \frac{\langle \mathcal{L}w, v\rangle}{\|w\|_{H_0^1(\Omega)}\|v\|_{H_0^1(\Omega)}} \right)^{-1} \end{aligned}$$

*Received by the editors September 22, 2003; accepted for publication (in revised form) September 30, 2004; published electronically August 9, 2005.
http://www.siam.org/journals/sima/36-6/43499.html
†Istituto di Matematica Applicata e Tecnologie Informatiche C.N.R., Via Ferrata 1, 27100 Pavia, Italy (sangalli@imati.cnr.it).

depend on the coefficients κ, β, and ρ.

Our analysis encompasses any linear and *coercive* operator $\mathcal{L}$, of which (1.1) is a model case. Given such an operator $\mathcal{L}$, we construct a norm $\|\!|\cdot|\!\|$ on its domain V such that the *continuity*

$$\sup_{w\in V}\sup_{v\in V}\frac{\langle \mathcal{L}w, v\rangle}{|\!|\!|w|\!|\!|\,|\!|\!|v|\!|\!|} \leq \mathcal{C}_c < +\infty \tag{1.3}$$

and the *inf-sup* condition

$$\inf_{w\in V}\sup_{v\in V}\frac{\langle \mathcal{L}w, v\rangle}{|\!|\!|w|\!|\!|\,|\!|\!|v|\!|\!|} \geq \mathcal{C}_{is} > 0 \tag{1.4}$$

hold true with constants $\mathcal{C}_c$ and $\mathcal{C}_{is}$ independent of $\mathcal{L}$. Therefore, for the example (1.1), $\mathcal{C}_c$ and $\mathcal{C}_{is}$ will be independent of the coefficients κ, β, and ρ.

If $\mathcal{L}$ were symmetric, besides coercive, then conditions (1.3)–(1.4) would hold true for the so-called energy norm, i.e., by setting $|\!|\!|w|\!|\!| := \langle \mathcal{L}w, w\rangle^{1/2}$, with $\mathcal{C}_c = \mathcal{C}_{is} = 1$. Our aim is to extend this trivial result to the nonsymmetric case, obtaining a suitable $\|\!|\cdot|\!\|$ by means of the function space interpolation.

The norm $\|\!|\cdot|\!\|$, for which (1.3)–(1.4) hold true, depends on $\mathcal{L}$ and gives the *natural* topology for $\mathcal{L}$. For the example (1.1)–(1.2), given a source term f and a perturbed source term $f + \delta f$, denoting by u and $u + \delta u$ the solutions of $\mathcal{L}u = f$ and $\mathcal{L}(u + \delta u) = f + \delta f$, respectively, one easily gets from (1.3)–(1.4)

$$\frac{|\!|\!|\delta u|\!|\!|}{|\!|\!|u|\!|\!|} \leq \frac{\mathcal{C}_c}{\mathcal{C}_{is}}\frac{|\!|\!|\delta f|\!|\!|_*}{|\!|\!|f|\!|\!|_*};$$

i.e., the relative perturbation of the solution of (1.2) is uniformly bounded by the relative perturbation of the source term, $|\!|\!|\cdot|\!|\!|_*$ being the dual of norm of $|\!|\!|\cdot|\!|\!|$. This is the proper framework to understand the behavior of (1.1)–(1.2) for small values of the diffusivity κ, when the higher order term $-\kappa\Delta$ acts as a singular perturbation on the lower order term $\beta\cdot\nabla + \rho\,\mathrm{Id}$.

Conditions (1.3)–(1.4) are also the proper framework for using some recent numerical methodologies for solving (1.1)–(1.2). Particularly, we are thinking of the *least-squares* formulations in the context of finite element methods [6] or in the context of wavelet methods [12], and of *adaptive wavelet methods* [11] (see also [10, 2, 5]).

More generally, (1.3)–(1.4) are the starting point for the classical analysis of numerical methods devoted to (1.1)–(1.2). When the continuity and inf-sup conditions are known for an operator $\mathcal{L}$, then *ideal* numerical methods should preserve them at the discrete level. This happens, for example, with symmetric and coercive operators (see [9]) or with some indefinite problems (as in mixed formulations; see [7]), and it is in general the key property for the classical error theory (see, e.g., [1]). Even though there are very effective numerical methods for solving (1.1)–(1.2), such as the *streamline-upwind Petrov-Galerkin* (SUPG) finite element method (see [8] and [17]), the error analysis of them typically does not follow the classical argument mentioned above and it is not completely satisfactory (see [18]). We hope this paper could give some insights for a deeper theoretical understanding of numerical methods devoted to (1.1)–(1.2) (we refer to [18, section 4], [19], and [3, section 2.1] for a further discussion on the topic).

This work is an extension of our previous analysis proposed in [18], where the convection-diffusion operator, without the reaction term, is considered. Different

estimates for (1.1)–(1.2) have been obtained by other authors; see, for example, the analysis by Bertoluzza, Canuto, and Tabacco in [4, section 2.1] or the paper by Dörfler [13]. The peculiarity of our paper is that both conditions (1.3)–(1.4) are obtained for (1.1)–(1.2).

The outline of this paper is as follows: in section 2 we present our methodology for obtaining (1.3)–(1.4) in the case of a generic nonsymmetric and coercive operator $\mathcal{L}$. Then we apply the theory—first, in section 3.1, to the very simple one-dimensional ($n = 1$) convection-diffusion-reaction model problem, and then, in section 3.2, to the multidimensional ($n > 1$) case—and discuss the results.

2. The abstract framework. In this section, we present our idea for obtaining uniform continuity and inf-sup conditions (1.3)–(1.4).

Let V be a Hilbert space, and let V^* be its dual. In the present section we consider a generic coercive isomorphism $\mathcal{L} : V \to V^*$ and the associated bilinear form

$$a(w, v) := {}_{V^*}\langle \mathcal{L}w, v\rangle_V \quad \forall w, v \in V. \tag{2.1}$$

The abstract variational problem which corresponds to (1.2) is

$$\text{find } u \in V \text{ such that } a(u, v) = {}_{V^*}\langle f, v\rangle_V \quad \forall v \in V. \tag{2.2}$$

We also assume that $\|\cdot\|_V$, the norm of V, is the *energy norm* for $\mathcal{L}$, i.e.,

$$a(w, w) = \|w\|_V^2 \quad \forall w \in V. \tag{2.3}$$

We split $\mathcal{L} = \mathcal{L}_{\text{sym}} + \mathcal{L}_{\text{skew}}$, and introduce the bilinear forms $a_{\text{sym}}(\cdot,\cdot)$ and $a_{\text{skew}}(\cdot,\cdot)$ on $V \times V$ such that

$$\begin{aligned} {}_{V^*}\langle \mathcal{L}_{\text{sym}}w, v\rangle_V &:= a_{\text{sym}}(w, v) := \frac{1}{2}\left(a(w, v) + a(v, w)\right) \quad \forall w, v \in V, \\ {}_{V^*}\langle \mathcal{L}_{\text{skew}}w, v\rangle_V &:= a_{\text{skew}}(w, v) := \frac{1}{2}\left(a(w, v) - a(v, w)\right) \quad \forall w, v \in V. \end{aligned} \tag{2.4}$$

In other words $\mathcal{L}_{\text{sym}}$ is the symmetric part of $\mathcal{L}$ (i.e., $a_{\text{sym}}(w, v) = a_{\text{sym}}(v, w)\ \forall w, v \in V$), and we have

$$\begin{aligned} a_{\text{sym}}(w, w) &= \|w\|_V^2 && \forall w \in V, \\ a_{\text{sym}}(w, v) &\le \|w\|_V \|v\|_V && \forall w, v \in V, \end{aligned} \tag{2.5}$$

while $\mathcal{L}_{\text{skew}}$ is the skew-symmetric part of $\mathcal{L}$ (i.e., $a_{\text{skew}}(w, v) = -a_{\text{skew}}(v, w)\ \forall w, v \in V$).

Finally, we define

$$\begin{aligned} \|w\|_{A_0}^2 &:= \|w\|_V^2 \quad \forall w \in V, \\ \|w\|_{A_1}^2 &:= \|w\|_V^2 + \|\mathcal{L}_{\text{skew}}w\|_{V^*}^2 \quad \forall w \in V, \end{aligned} \tag{2.6}$$

where

$$\|\mathcal{L}_{\text{skew}}w\|_{V^*} = \sup_{v \in V} \frac{a_{\text{skew}}(w, v)}{\|v\|_V}.$$

We also set $A_0 = A_1 = V$ from the algebraic standpoint; in other words A_0 and A_1 are the same space with the same topology, but the two norms $\|\cdot\|_{A_0}$ and $\|\cdot\|_{A_1}$ are different (even though equivalent, up to constants depending on L).

The following lemma states two basic estimates; we explicitly compute the constants appearing in the estimates to highlight their independence of $\mathcal{L}$.

LEMMA 2.1. *We have*

$$a(w,v) \le 2^{1/2}\|w\|_{A_i}\|v\|_{A_{1-i}} \quad \forall w,v \in V, \tag{2.7}$$

$$\sup_{v\in V}\frac{a(w,v)}{\|v\|_{A_{1-i}}} \ge 5^{-1/2}\|w\|_{A_i} \quad \forall w \in V \tag{2.8}$$

for $i=0$ or $i=1$.

Proof. Let v and w be two generic elements of V.

By using the Cauchy–Schwarz inequality we easily get

$$\begin{aligned} a(w,v) &= a_{\mathrm{sym}}(w,v) + a_{\mathrm{skew}}(w,v) \\ &\le \|w\|_V\|v\|_V + \|\mathcal{L}_{\mathrm{skew}}w\|_{V^*}\|v\|_V \\ &\le 2^{1/2}\|w\|_{A_1}\|v\|_{A_0}. \end{aligned}$$

Similarly, since $a_{\mathrm{skew}}(w,v) = -a_{\mathrm{skew}}(v,w)$, we also get $a(w,v) \le 2^{1/2}\|w\|_{A_0}\|v\|_{A_1}$, and then (2.7) follows.

Recalling (2.3) and (2.5), we have

$$\|w\|_V \le \sup_{v\in V}\frac{a(w,v)}{\|v\|_V} \tag{2.9}$$

and

$$\sup_{v\in V}\frac{a_{\mathrm{sym}}(w,v)}{\|v\|_V} = \|w\|_V \le \sup_{v\in V}\frac{a(w,v)}{\|v\|_V}. \tag{2.10}$$

Then, we get

$$\begin{aligned} \|\mathcal{L}_{\mathrm{skew}}w\|_{V^*} &= \sup_{v\in V}\frac{a_{\mathrm{skew}}(w,v)}{\|v\|_V} \\ &\le \sup_{v\in V}\frac{a(w,v)}{\|v\|_V} + \sup_{v\in V}\frac{a_{\mathrm{sym}}(w,v)}{\|v\|_V} \\ &\le 2\sup_{v\in V}\frac{a(w,v)}{\|v\|_V}, \end{aligned} \tag{2.11}$$

and collecting (2.9) and (2.11), we get

$$\|w\|_{A_1} \le 5^{1/2}\sup_{v\in V}\frac{a(w,v)}{\|v\|_{A_0}}, \tag{2.12}$$

which is (2.8) for $i=1$. We are left to show that

$$\|w\|_{A_0} \le 5^{1/2}\sup_{v\in V}\frac{a(w,v)}{\|v\|_{A_1}}; \tag{2.13}$$

for that purpose, we make use of a duality argument. Reasoning as for (2.12) we obtain

$$\|\tilde{w}\|_{A_1} \le 5^{1/2}\sup_{v\in V}\frac{a(v,\tilde{w})}{\|v\|_{A_0}} \tag{2.14}$$

for any $\tilde{w} \in V$. Given a generic $w \in V$, we associate with it $\tilde{w} \in V$ such that $a(v, \tilde{w}) = a_{\mathrm{sym}}(v, w)\ \forall v \in V$; thanks to (2.14) we have

$$\|\tilde{w}\|_{A_1} \leq 5^{1/2} \sup_{v \in V} \frac{a(v, \tilde{w})}{\|v\|_{A_0}} = 5^{1/2} \sup_{v \in V} \frac{a_{\mathrm{sym}}(v, w)}{\|v\|_{A_0}} = 5^{1/2} \|w\|_{A_0},$$

whence

$$\begin{aligned} \|w\|_{A_0}^2 &= a_{\mathrm{sym}}(w, w) = a(w, \tilde{w}) \\ &\leq \sup_{v \in V} \frac{a(w, v)}{\|v\|_{A_1}} \cdot \|\tilde{w}\|_{A_1} \\ &\leq 5^{1/2} \sup_{v \in V} \frac{a(w, v)}{\|v\|_{A_1}} \cdot \|w\|_{A_0}, \end{aligned}$$

which completes the proof. ∎

From Lemma 2.1 we can obtain a family of intermediate estimates by means of the function spaces interpolation. We follow the notation and the definitions of [20]; for the reader's convenience, we recall the fundamental definition of *interpolated norm*, according to the *K-method*: given $0 < \theta < 1$ and $1 \leq p \leq +\infty$, we define

$$\|w\|_{(A_0, A_1)_{\theta, p}} := \left[\int_0^{+\infty} \inf_{\substack{w_0 \in A_0, w_1 \in A_1, \\ w_0 + w_1 = w}} \left(t^{-\theta} \|w_0\|_{A_0} + t^{1-\theta} \|w_1\|_{A_1} \right)^p \frac{dt}{t} \right]^{1/p}. \tag{2.15}$$

Generally $(A_0, A_1)_{\theta,p}$ is the space of functions $w \in A_0 + A_1$ such that $\|w\|_{(A_0,A_1)_{\theta,p}} < +\infty$. In our particular case, A_0 and A_1 are the same space from the algebraic standpoint ($A_0 \equiv A_1 \equiv V$), and $\|\cdot\|_{(A_0,A_1)_{\theta,p}}$ simply is a new norm on V.

LEMMA 2.2. *Given θ, p, and p' such that $0 < \theta < 1$, $1 \leq p \leq +\infty$, and $1/p + 1/p' = 1$, we have*

$$a(w, v) \leq 2^{1/2} \|w\|_{(A_0,A_1)_{\theta,p}} \|v\|_{(A_0,A_1)_{1-\theta,p'}} \quad \forall w, v \in V, \tag{2.16}$$

$$\sup_{v \in V} \frac{a(w, v)}{\|v\|_{(A_0,A_1)_{1-\theta,p'}}} \geq 5^{-1/2} \|w\|_{(A_0,A_1)_{\theta,p}} \quad \forall w \in V. \tag{2.17}$$

Proof. Typically interpolation theorems are stated in terms of linear operators instead of bilinear forms. Then it is more convenient to rephrase (2.7) as

$$\begin{aligned} \|\mathcal{L} w\|_{A_1^*} &\leq 2^{1/2} \|w\|_{A_0}, \\ \|\mathcal{L} w\|_{A_0^*} &\leq 2^{1/2} \|w\|_{A_1} \end{aligned} \tag{2.18}$$

and (2.8) as

$$\begin{aligned} \|w\|_{A_0} &\leq 5^{1/2} \|\mathcal{L} w\|_{A_1^*}, \\ \|w\|_{A_1} &\leq 5^{1/2} \|\mathcal{L} w\|_{A_0^*} \end{aligned} \tag{2.19}$$

$\forall\ w \in V$.

From (2.18) and thanks to the theorems in [20, sections 1.3.3 and 1.11.2], we get (2.16). Proceeding similarly for $\mathcal{L}^{-1}$, from (2.19) we obtain

$$\|\mathcal{L}^{-1} \phi\|_{(A_0,A_1)^*_{1-\theta,p'}} \leq 5^{1/2} \|\phi\|_{(A_0,A_1)_{\theta,p}}$$

for any $\phi \in V^*$, and that gives (2.17). ∎

Thanks to (2.5), $\mathcal{L}_{\rm sym}$ is an isomorphism from V into $V^* \equiv \mathcal{L}_{\rm sym}(V)$; henceforth, we also assume that $\mathcal{L}_{\rm skew}$ is injective. Then we introduce the two Hilbert spaces C_0 and C_1:

$$\begin{aligned} &C_0 := \mathcal{L}_{\rm skew}(V) \quad \text{with } \|\phi\|_{C_0} := \|\mathcal{L}_{\rm skew}^{-1}\phi\|_V, \\ &C_1 := \mathcal{L}_{\rm sym}(V) \quad \text{with } \|\phi\|_{C_1} := \|\mathcal{L}_{\rm sym}^{-1}\phi\|_V = \|\phi\|_{V^*}. \end{aligned} \tag{2.20}$$

In the next lemma we analyze the structure of $\|\cdot\|_{(A_0,A_1)_{\theta,p}}$.

LEMMA 2.3. *Given θ, p, and p' such that $0 < \theta < 1$, $1 \le p \le +\infty$, and $1/p + 1/p' = 1$, we have*

$$1/10\,\|w\|^2_{(A_0,A_1)_{\theta,p}} \le \|w\|^2_V + \|\mathcal{L}_{\rm skew}w\|^2_{(C_0,C_1)_{\theta,p}} \le 2\|w\|^2_{(A_0,A_1)_{\theta,p}} \quad \forall w \in V. \tag{2.21}$$

Proof. Since $\|w\|_V \le \|w\|_{A_i}$ with $i = 0, 1$, then $\|w\|_V \le \|w\|_{(A_0,A_1)_{\theta,p}}$ follows by a straightforward application of the interpolation theorem (e.g., [20, section 1.3.3]). We also have

$$\begin{aligned} &\|\mathcal{L}_{\rm skew}w\|_{C_0} = \|w\|_{A_0}, \\ &\|\mathcal{L}_{\rm skew}w\|_{C_1} \le \|w\|_{A_1}, \end{aligned}$$

which gives $\|\mathcal{L}_{\rm skew}w\|_{(C_0,C_1)_{\theta,p}} \le \mathcal{C}\|w\|_{(A_0,A_1)_{\theta,p}}$, whence $\|w\|^2_V + \|\mathcal{L}_{\rm skew}w\|^2_{(C_0,C_1)_{\theta,p}} \le 2\|w\|^2_{(A_0,A_1)_{\theta,p}}$.

In order to complete the proof, we deal directly with the definition of interpolated norm (2.15). For any $t > 0$ consider the two splittings

$$\begin{aligned} &w = \tilde{w}_0(t) + \tilde{w}_1(t) \text{ with } \tilde{w}_i(t) \in V, \quad i = 1, 2, \\ &w = \hat{w}_0(t) + \hat{w}_1(t) \text{ with } \hat{w}_i(t) \in V, \quad i = 1, 2. \end{aligned} \tag{2.22}$$

Then define $w_0(t) \in V$ and $w_1(t) \in V$ such that $\mathcal{L}w_i(t) = \mathcal{L}_{\rm sym}\tilde{w}_i(t) + \mathcal{L}_{\rm skew}\hat{w}_i(t)$, i.e.,

$$a(w_i(t), v) = a_{\rm sym}(\tilde{w}_i(t), v) + a_{\rm skew}(\hat{w}_i(t), v) \quad \forall v \in V, \quad i = 0, 1, \tag{2.23}$$

whence $w = w_0(t) + w_1(t)$ $\forall t > 0$.

Thanks to (2.8) and to the properties of $a_{\rm sym}(\cdot,\cdot)$ and $a_{\rm skew}(\cdot,\cdot)$ we have

$$\begin{aligned} \|w_0(t)\|_{A_0} &\le 5^{1/2} \sup_{v\in V} \frac{a(w_0(t), v)}{\|v\|_{A_1}} \\ &\le 5^{1/2} \left(\sup_{v\in V} \frac{a_{\rm sym}(\tilde{w}_0(t), v) - a_{\rm skew}(v, \hat{w}_0(t))}{\|v\|_{A_1}} \right) \\ &\le 5^{1/2} \left(\sup_{v\in V} \frac{a_{\rm sym}(\tilde{w}_0(t), v)}{\|v\|_V} + \sup_{v\in V} \frac{a_{\rm skew}(v, \hat{w}_0(t))}{\|\mathcal{L}_{\rm skew}v\|_{V^*}} \right) \\ &\le 5^{1/2} \left(\|\tilde{w}_0(t)\|_V + \|\hat{w}_0(t)\|_V \right). \end{aligned} \tag{2.24}$$

In a similar way, we have

$$
\begin{aligned}
\|w_1(t)\|_{A_1} &\le 5^{1/2} \sup_{v\in V} \frac{a(w_1(t),v)}{\|v\|_{A_0}} \\
&\le 5^{1/2} \left(\sup_{v\in V} \frac{a_{\mathrm{sym}}(\tilde{w}_1(t),v) + a_{\mathrm{skew}}(\hat{w}_1(t),v)}{\|v\|_{A_0}} \right) \\
&\le 5^{1/2} \left(\sup_{v\in V} \frac{a_{\mathrm{sym}}(\tilde{w}_1(t),v)}{\|v\|_V} + \sup_{v\in V} \frac{a_{\mathrm{skew}}(\hat{w}_1(t),v)}{\|v\|_V} \right) \\
&\le 5^{1/2} \left(\|\tilde{w}_1(t)\|_V + \|\mathcal{L}_{\mathrm{skew}} \hat{w}_1(t)\|_{V^*} \right).
\end{aligned} \tag{2.25}
$$

From (2.15), by the triangle inequality and using (2.24)–(2.25), we have

$$
\begin{aligned}
\|w\|_{(A_0,A_1)_{\theta,p}} &\le \left[\int_0^{+\infty} \left(t^{-\theta} \|w_0(t)\|_{A_0} + t^{1-\theta} \|w_1(t)\|_{A_1} \right)^p \frac{dt}{t} \right]^{1/p} \\
&\le 5^{1/2} \Bigg[\int_0^{+\infty} \Big(t^{-\theta} \|\tilde{w}_0(t)\|_V + t^{-\theta} \|\hat{w}_0(t)\|_V \\
&\qquad + t^{1-\theta} \|\tilde{w}_1(t)\|_V + t^{1-\theta} \|\mathcal{L}_{\mathrm{skew}} \hat{w}_1(t)\|_{V^*} \Big)^p \frac{dt}{t} \Bigg]^{1/p} \\
&\le 5^{1/2} \left[\int_0^{+\infty} \left(t^{-\theta} \|\tilde{w}_0(t)\|_V + t^{1-\theta} \|\tilde{w}_1(t)\|_V \right)^p \frac{dt}{t} \right]^{1/p} \\
&\quad + \left[\int_0^{+\infty} \left(t^{-\theta} \|\mathcal{L}_{\mathrm{skew}} \hat{w}_0(t)\|_{C_0} + t^{1-\theta} \|\mathcal{L}_{\mathrm{skew}} \hat{w}_1(t)\|_{C_1} \right)^p \frac{dt}{t} \right]^{1/p}.
\end{aligned}
$$

Finally, taking the infimum over all $\tilde{w}_0 \in V$, and $\tilde{w}_1 = w - \tilde{w}_0 \in V$, $\hat{w}_0 \in V$ and $\hat{w}_1 = w - \hat{w}_0 \in V$, and using [20, 1.3.3.(f)], we finally get $\|w\|_{(A_0,A_1)_{\theta,p}} \le 5^{1/2} \left(\|w\|_V + \|\mathcal{L}_{\mathrm{skew}} w\|_{(C_0,C_1)_{\theta,p}} \right)$, completing the proof of (2.21). ∎

When $p = p' = 2$ and $\theta = 1-\theta = 1/2$, Lemma 2.2 gives the continuity and inf-sup conditions for $\mathcal{L}$, as stated in section 1, where $|\!|\!| \cdot |\!|\!| = \| \cdot \|_{(A_0,A_1)_{1/2,2}}$. In particular, under the hypotheses of Lemma 2.3, we have the following obvious corollary.

COROLLARY 2.4. *Under the assumption of Lemma* 2.3 *and setting*

$$
|\!|\!| \cdot |\!|\!| := \left(\| \cdot \|_V^2 + \| \mathcal{L}_{\mathrm{skew}} \cdot \|^2_{(C_0,C_1)_{1/2,2}} \right)^{1/2}, \tag{2.26}
$$

we have the continuity and inf-sup conditions (1.3)–(1.4) *for* $\mathcal{L}$*, with constants* $\mathcal{C}_c$ *and* $\mathcal{C}_{is}$ *independent of* $\mathcal{L}$*.*

The particular case considered in Corollary 2.4 is of interest from the numerical analysis standpoint, as discussed in [18, section 4].

3. The convection-diffusion-reaction operator. We now apply the results of the previous section to the convection-diffusion-reaction operator. In Lemmas 2.1–2.3 we have explicitly computed the constants involved into the estimates, in order to emphasize that the estimates do not depend on $\mathcal{L}$; henceforth, for the sake of simplicity, we will use generic constants denoted by $\mathcal{C}$, $\mathcal{C}_1$, $\mathcal{C}_2$, which are independent on the operator coefficients κ, β, and ρ and on the domain Ω.

3.1. The one-dimensional case. We start with the analysis of the very simple one-dimensional operator, with constant coefficients $\kappa > 0$ and $\rho \geq 0$ and unitary velocity. Then, for this subsection only, we will consider a special case of (1.1), which is

$$w \mapsto \mathcal{L}w := -\kappa w'' + w' + \rho w, \tag{3.1}$$

where the argument w is a function of the interval $\Omega = [0,1]$.

We consider first, and with particular emphasis, the ordinary differential equation with homogeneous Dirichlet boundary conditions (1.2). The variational formulation (2.2) reads as

$$\text{find } u \in V \text{ such that } a(u,v) = \int_0^1 fv \quad \forall v \in V,$$

where

$$\begin{aligned} V &= H_0^1(0,1) \text{ with } \|\cdot\|_V^2 = \kappa|\cdot|_{H^1}^2 + \rho\|\cdot\|_{L^2}^2, \\ a(w,v) &= \kappa\int_0^1 w'v' + \int_0^1 w'v + \rho\int_0^1 wv. \end{aligned} \tag{3.2}$$

Then $\mathcal{L}_{\text{sym}}w = -\kappa w'' + \rho w$, $\mathcal{L}_{\text{skew}}w = w'$, $a_{\text{sym}}(w,v) = \kappa\int_0^1 w'v' + \rho\int_0^1 wv$, and $a_{\text{skew}}(w,v) = \int_0^1 w'v$. From the algebraic standpoint, $C_0 = L_0^2(0,1)$ and $C_1 = H^{-1}(0,1)$, where L_0^2 is the subspace of L^2 of zero mean value functions, its natural norm is $\|\cdot\|_{L_0^2} := \|\cdot\|_{L^2}$, and H^{-1} is the dual of H_0^1, endowed with the dual norm $\|\cdot\|_{H^{-1}} = \sup_{v\in H_0^1(0,1)} \langle\cdot, v\rangle / |v|_{H^1}$ (we recall that $|\cdot|_{H^1} := [\int_0^1 (w')^2]^{1/2}$ is a norm on H_0^1). It is easy to see that L_0^2 is a dense subspace of H^{-1}. From Corollary 2.4 we immediately have the following result.

THEOREM 3.1. *For the case* (3.1)–(3.2)*, uniform continuity and inf-sup conditions* (1.3)–(1.4) *hold true with respect to the norm*

$$w \mapsto |\!|\!| w |\!|\!| = \left(\kappa|w|_{H^1}^2 + \|w'\|_{(C_0,C_1)_{1/2,2}}^2 + \rho\|w\|_{L^2}^2\right)^{1/2}. \tag{3.3}$$

Now we focus our attention on $|\!|\!|\cdot|\!|\!|$ in (3.3) in order to better understand its structure. Roughly speaking, the term $\|w'\|_{(C_0,C_1)_{1/2,2}}$ is related to the skew-symmetric part of $\mathcal{L}$, which is the *first* order derivative. Then we expect $w \mapsto \|w'\|_{(C_0,C_1)_{1/2,2}}$ to act as a 1/2-order norm uniformly on the operator coefficients κ and ρ. That is in fact stated in the next theorem: we show that $\|w'\|_{(C_0,C_1)_{1/2,2}}$ stays between the $H^{1/2}$-seminorm and $H_{00}^{1/2}$-norm, where $H^{1/2} := (L^2, H^1)_{1/2,2}$ and $H_{00}^{1/2} := (L^2, H_0^1)_{1/2,2}$ are the two usual Hilbert spaces of order 1/2, endowed with the usual norms given by interpolation (see [15]), and $|w|_{H^{1/2}}$ is the seminorm $\|w - \Pi_0 w\|_{H^{1/2}}$, $\Pi_0\cdot$ denoting the mean value of its argument.

THEOREM 3.2. *For the case* (3.1)–(3.2)*, we have*

$$\mathcal{C}_1|w|_{H^{1/2}} \leq \|w'\|_{(C_0,C_1)_{1/2,2}} \leq \mathcal{C}_2\|w\|_{H_{00}^{1/2}} \quad \forall w \in V. \tag{3.4}$$

Proof. When $\rho = 0$, (3.4) follows from (3.18); we assume henceforth that $\rho > 0$.

We consider first the left inequality in (3.4), i.e.,

$$\mathcal{C}|w|_{H^{1/2}} \leq \|w'\|_{(C_0,C_1)_{1/2,2}} \quad \forall w \in V. \tag{3.5}$$

It is easy to see that $\|z'\|_{L^2} \simeq \|z\|_{H^1}$ and $\|z'\|_{H^{-1}} \simeq \|z\|_{L^2}$ for any $z \in H^1 \cap L^2_0$; then, thanks to the theorems in [20, sections 1.3.3, 1.11.2, and 1.17.1], the first order derivative is a topological isomorphism from $H^{1/2} \cap L^2_0$ into $(H^{-1}, L^2)_{1/2,2}$, which means

$$|w|_{H^{1/2}} = \|w - \Pi_0 w\|_{H^{1/2}} \simeq \|w'\|_{(H^{-1},L^2)_{1/2,2}}. \tag{3.6}$$

We introduce now the new space $\widetilde{C}_0$: from the algebraic standpoint we set $\widetilde{C}_0 := L^2$, and we define $\|\cdot\|_{\widetilde{C}_0} := \left(\kappa\|\cdot\|^2_{L^2} + \rho\|\cdot\|^2_{H^{-1}}\right)^{1/2}$. Our next step is to show that

$$\|\phi\|_{(H^{-1},L^2)_{1/2,2}} \le \mathcal{C}\|\phi\|_{(\widetilde{C}_0,C_1)_{1/2,2}} \quad \forall \phi \in L^2. \tag{3.7}$$

For that purpose we split a generic $\phi \in L^2$ into

$$\phi = \phi_{\text{high}} + \phi_{\text{low}}, \tag{3.8}$$

where $\phi_{\text{high}}, \phi_{\text{low}} \in L^2$ are, roughly speaking, the high frequency part and the low frequency part of ϕ, respectively, in such a way that

$$\kappa^{1/2}\|\phi_{\text{high}}\|_{L^2} + \rho^{1/2}\|\phi_{\text{low}}\|_{H^{-1}} \le \mathcal{C}\|\phi\|_{\widetilde{C}_0}, \tag{3.9}$$

$$\kappa^{-1/2}\|\phi_{\text{high}}\|_{H^{-1}} + \rho^{-1/2}\|\phi_{\text{low}}\|_{L^2} \le \mathcal{C}\|\phi\|_{C_1}. \tag{3.10}$$

For that purpose, we introduce an auxiliary problem: let $\psi \in H^1_0$ be the solution of

$$\mathcal{L}_{\text{sym}}\psi = \phi \qquad \text{in } (0,1) \tag{3.11}$$

and let $\phi_{\text{high}} := -\kappa\psi''$ and $\phi_{\text{low}} := \rho\psi$.

Multiplying both members of the differential equation (3.11) by $-\psi''$, integrating over $(0,1)$, and integrating by parts we get

$$\kappa\|\psi''\|^2_{L^2} + \rho\|\psi'\|^2_{L^2} = -\int_0^1 \phi\psi'';$$

then, thanks to the Cauchy–Schwarz inequality, we have

$$\|\phi_{\text{high}}\|_{L^2} = \|\kappa\psi''\|_{L^2} \le \|\phi\|_{L^2}. \tag{3.12}$$

Integrating (3.11) we have

$$-\kappa\psi' + \kappa\psi'(0) + \rho\Psi = \Phi,$$

where $\Psi(x) = \int_0^x \psi(t)\,dt$ and analogously $\Phi(x) = \int_0^x \phi(t)\,dt$. After multiplying both members by $\Psi - \Pi_0\Psi$, integrating over $(0,1)$, and integrating by parts, we obtain

$$\kappa\|\psi\|^2_{L^2} + \rho\|\Psi - \Pi_0\Psi\|^2_{L^2} = \int_0^1 \Phi(\Psi - \Pi_0\Psi),$$

whence now

$$\|\phi_{\text{low}}\|_{H^{-1}} = \rho\|\Psi - \Pi_0\Psi\|_{L^2} \le \|\Phi - \Pi_0\Phi\|_{L^2} = \|\phi\|_{H^{-1}}. \tag{3.13}$$

Collecting (3.12)–(3.13) we obtain (3.9). From (3.11) it is also easy to obtain the estimate $(\kappa\|\psi'\|^2_{L^2} + \rho\|\psi\|^2_{L^2})^{1/2} \le \|\phi\|_{V^*} = \|\phi\|_{C_1}$, which gives (3.10) straightforwardly.

Consider now the linear operator $\phi \mapsto (\phi_{\text{high}}, \phi_{\text{low}})$ from L^2 into $L^2 \times L^2$, with ϕ_{high}, ϕ_{low} as defined earlier: by interpolation from the two continuity estimates (3.9) and (3.10) we get

$$\|\phi_{\text{high}}\|_{(L^2,H^{-1})_{1/2,2}} + \|\phi_{\text{low}}\|_{(H^{-1},L^2)_{1/2,2}} \leq \mathcal{C}\|\phi\|_{(\widetilde{C}_0,C_1)_{1/2,2}}, \tag{3.14}$$

whence, by using the triangle inequality and since $\|\cdot\|_{(L^2,H^{-1})_{1/2,2}} = \|\cdot\|_{(H^{-1},L^2)_{1/2,2}}$, we obtain (3.7). Finally (3.6) and (3.7) gives (3.5).

Now we consider the right equivalence in (3.4), which is

$$\|w'\|_{(C_0,C_1)_{1/2,2}} \leq \mathcal{C}\|w\|_{H^{1/2}_{00}} \quad \forall w \in V. \tag{3.15}$$

Given $w \in H^1_0$ it is easy to see that

$$\|w'\|_{C_0} = \|w\|_V = \|w\|_{C_1^*}$$

and

$$\|w'\|_{C_1} = \|w'\|_{V^*} \leq \|w\|_{\widetilde{C}_0^*},$$

whence (thanks to the theorem in [20, section 1.11.2])

$$\|w'\|_{(C_0,C_1)_{1/2,2}} \leq \|w\|_{(C_1^*,\widetilde{C}_0^*)_{1/2,2}} = \|w\|_{(\widetilde{C}_0,C_1)^*_{1/2,2}}. \tag{3.16}$$

Moreover, passing to the duals in (3.7), still using the theorem in [20, section 1.11.2], we also have

$$\|w\|_{(\widetilde{C}_0,C_1)^*_{1/2,2}} \leq \|w\|_{(H^{-1},L^2)^*_{1/2,2}} = \|w\|_{(H^1_0,L^2)_{1/2,2}} = \|w\|_{(H^{1/2}_{00})}. \tag{3.17}$$

Inequalities (3.16)–(3.17) give (3.15). □

It is worth noting that Theorems 3.1 amd 3.2 allow for $\rho = 0$ as well; in that case we have $\|w'\|_{(C_0,C_1)_{1/2,2}} = \|w'\|_{(H^{-1},L^2_0)_{1/2,2}}$, since the coefficient κ easily cancels when interpolating. Let $H^1_\#$ be the subspace of H^1 of functions w such that $w(0) = w(1)$ endowed with the $\|\cdot\|_{H^1_\#} := \|\cdot\|_{H^1}$, and $H^{1/2}_\# := (L^2, H^1_\#)_{1/2,2}$ endowed with the norm given by interpolation. Given $z \in H^1_\# \cap L^2_0$, one has $\|z'\|_{L^2_0} \simeq \|z\|_{H^1_\#}$ and $\|z'\|_{H^{-1}} \simeq \|z\|_{L^2}$, whence (making use of the theorems in [20, sections 1.3.3, 1.11.2, and 1.17.1], for example) we have $\|z'\|_{(H^{-1},L^2_0)_{1/2,2}} \simeq \|z\|_{(L^2,H^1_\#)_{1/2,2}}$ and therefore $\|w'\|_{(H^{-1},L^2_0)_{1/2,2}} \simeq \|w - \Pi_0 w\|_{(L^2,H^1_\#)_{1/2,2}}$, for any $w \in H^1_0$; this means that we have the following characterization:

$$\rho = 0 \;\Rightarrow\; |w|_{H^{1/2}_\#} := \|w - \Pi_0 w\|_{(L^2,H^1_\#)_{1/2,2}} = \|w'\|_{(C_0,C_1)_{1/2,2}} \quad \forall w \in V. \tag{3.18}$$

We may also deal with different kinds of boundary conditions. Consider the example

$$\begin{cases} \mathcal{L}u = f & \text{in } (0,1), \\ u(0) = u'(1) = 0, \end{cases} \tag{3.19}$$

where $\mathcal{L}$ is still formally given by (3.1). The variational formulation (2.2) now requires

$$V = \left\{ v \in H^1(0,1) \text{ such that } v(0) = 0 \right\},$$
$$a(w,v) = \kappa \int_0^1 w'v' + \int_0^1 w'v + \rho \int_0^1 wv.$$

The key point is that the bilinear form $a(\cdot,\cdot)$ is coercive on V; accordingly, we define $\|\cdot\|_V$ as

$$\|w\|_V^2 := a(w,w) = \kappa|w|_{H^1}^2 + \rho\|w\|_{L^2}^2 + \frac{1}{2}w(1)^2,$$

and we now have

$$a_{\rm sym}(w,v) = \kappa\int_0^1 w'v' + \rho\int_0^1 wv + \frac{1}{2}w(1)v(1),$$
$$a_{\rm skew}(w,v) = \int_0^1 w'v - \frac{1}{2}w(1)v(1).$$

Then we can still make use of the theory of section 2 and obtain uniform inf-sup and continuity conditions from Corollary 2.4.

When the bilinear form $a(\cdot,\cdot)$ is not coercive, we cannot use the results of section 2. This is the case of

$$\begin{cases} -\kappa u'' + u' = f \text{ in } (0,1), \\ u'(0) = u(1) = 0, \end{cases} \tag{3.20}$$

i.e., when $\rho = 0$ and we prescribe a Neumann boundary condition at the *inflow* $x = 0$; then $V = \left\{v \in H^1(0,1) \text{ such that } v(1) = 0\right\}$ and

$$a(w,w) = \kappa|w|_{H^1}^2 + \rho\|w\|_{L^2}^2 - \frac{1}{2}w(1)^2,$$

which is not positive in general, when κ and ρ are small enough. However, when $f = 1$ the solution of (3.20) is $u(x) = \kappa\left(\exp(1/\kappa) - \exp(x/\kappa)\right) + x - 1$; for $\kappa \to 0$ we have $\|u\|_{L^2} \approx \kappa\exp(1/\kappa)$, whence we see that (3.20) is not uniformly well posed with respect to κ.

3.2. The multidimensional case. In this section, we analyze the multidimensional convection-diffusion-reaction operator with Dirichlet homogeneous boundary conditions (1.1)–(1.2), and the associated bilinear form

$$a(w,v) = \kappa\int_\Omega \nabla w\cdot\nabla v + \int_\Omega \beta\cdot\nabla w\, v + \int_\Omega \rho wv,$$

which is defined on $H_0^1(\Omega)\times H_0^1(\Omega)$ (see, e.g., [15]). Under the assumption

$$\rho - \frac{1}{2}{\rm div}(\beta) \geq 0 \tag{3.21}$$

the bilinear form $a(\cdot,\cdot)$ is coercive, whence we set

$$V = H_0^1(\Omega),$$
$$\|w\|_V^2 = a(w,w) = \kappa|w|_{H^1}^2 + \left(\rho - \frac{1}{2}{\rm div}(\beta)\right)\|w\|_{L^2}^2. \tag{3.22}$$

The decomposition (2.4) gives

$$a_{\rm sym}(w,v) = \kappa\int_\Omega \nabla w\cdot\nabla v + \int_\Omega \left(\rho - \frac{1}{2}{\rm div}(\beta)\right)wv,$$
$$a_{\rm skew}(w,v) = \int_\Omega \beta\cdot\nabla w\, v + \frac{1}{2}\int_\Omega {\rm div}(\beta)\, wv. \tag{3.23}$$

For the sake of simplicity, we shall consider henceforth the case

$$\operatorname{div}(\beta) = 0. \tag{3.24}$$

In order to apply Corollary 2.4 to this case, we need $\mathcal{L}_{\text{skew}} = \beta \cdot \nabla$ to be injective on V. This is assured, for example, by the assumption

$$\text{there exists a smooth } \phi : \Omega \to \mathbb{R} \text{ such that } \nabla\phi \cdot \beta \geq \mathcal{C} > 0; \tag{3.25}$$

we refer to [14] for further details. Definition (2.20) says that, from the algebraic standpoint, C_0 is the space of the streamline derivatives $\beta \cdot \nabla w$ of functions $w \in H^1_0$, while C_1 is H^{-1}. Corollary 2.4 then gives the following result.

THEOREM 3.3. *For the case* (3.22), (3.24)–(3.25)*, the uniform continuity and inf-sup conditions* (1.3)–(1.4) *hold true with respect to the norm*

$$w \mapsto |\!|\!| w |\!|\!| = \left(\kappa |w|^2_{H^1} + \|\beta \cdot \nabla w\|^2_{(C_0,C_1)_{1/2,2}} + \rho \|w\|^2_{L^2} \right)^{1/2}. \tag{3.26}$$

Roughly speaking, we expect $\|\beta \cdot \nabla w\|_{(C_0,C_1)_{1/2,2}}$ to be of order $1/2$ in the direction of β and of order 0 in the directions orthogonal to β (this can be more easily seen for the case $\rho = 0$), but a rigorous analysis of the structure of $\|\beta \cdot \nabla w\|_{(C_0,C_1)_{1/2,2}}$ is more difficult now than for the simpler one-dimensional case considered in section 3.1. The next result shows that $\|\beta \cdot \nabla w\|_{(C_0,C_1)_{1/2,1}}$ has some uniform bounds independent of κ and ρ (though the anisotropy is not investigated). Then we end with a comparison between $\|\beta \cdot \nabla w\|_{(C_0,C_1)_{1/2,1}}$ and $\|\beta \cdot \nabla w\|_{(C_0,C_1)_{1/2,2}}$.

PROPOSITION 3.4. *For the case* (3.22), (3.24)–(3.25)*, we have*

$$\begin{aligned} \mathcal{C}_p \|\beta\|^{1/2}_{L^\infty} \mathit{diam}(\Omega)^{-1/2} \|w\|_{L^2} &\leq \|\beta \cdot \nabla w\|_{(C_0,C_1)_{1/2,1}} \\ &\leq \mathcal{C} \|\beta\|^{1/2}_{L^\infty} \|w\|_{(L^2,H^1_0)_{1/2,1}} \quad \forall w \in V, \end{aligned} \tag{3.27}$$

where the constant $\mathcal{C}_p$ of the Poincaré-like inequality depends on $\beta/\|\beta\|_{L^\infty}$ and (the shape of) Ω.

Proof. Let η be the solution of $\beta \cdot \nabla \eta = \|\beta\|_{L^\infty}$ with $\eta = 0$ on $\partial\Omega^- := \{\mathbf{x} \in \partial\Omega | \beta(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) < 0\}$, where $\mathbf{n}$ denotes the outward normal unit vector defined on $\partial\Omega$. The existence of η is guaranteed by (3.25). Given $w \in H^1_0$, integrating by parts, and using the Cauchy–Schwarz inequality and (3.24), we have

$$\begin{aligned} \|\beta\|_{L^\infty} \|w\|^2_{L^2} &= \int_\Omega \beta \cdot \nabla \eta \, w^2 \\ &= -2 \int_\Omega \eta w \, \beta \cdot \nabla w \\ &\leq 2 \|\eta w\|_V \|\beta \cdot \nabla w\|_{V^*}. \end{aligned} \tag{3.28}$$

We have

$$\|\eta w\|_{L^2} \leq \|\eta\|_{L^\infty} \|w\|_{L^2}, \tag{3.29}$$

and using the classical Poincaré inequality, it is easy to get

$$\begin{aligned} |\eta w|_{H^1} &\leq \mathcal{C}(\|\eta\|_{L^\infty} |w|_{H^1} + \|\nabla \eta\|_{(L^\infty)^2} \|w\|_{L^2}) \\ &\leq \mathcal{C}(\|\eta\|_{L^\infty} + \operatorname{diam}(\Omega) \|\nabla \eta\|_{(L^\infty)^2}) |w|_{H^1}. \end{aligned} \tag{3.30}$$

Moreover, thanks to (3.25), we have $\widetilde{\mathcal{C}_p} := \mathrm{diam}(\Omega)^{-1}\|\eta\|_{L^\infty} + \|\nabla\eta\|_{(L^\infty)^2} < +\infty$ (e.g., see [14, Theorem 3.2]), where $\widetilde{\mathcal{C}_p}$ depends on η, i.e., on $\beta/\|\beta\|_{L^\infty}$ and on (the shape of) Ω. Then

$$\|\eta w\|_V \le \mathcal{C}\widetilde{\mathcal{C}_p}\mathrm{diam}(\Omega)\|w\|_V; \tag{3.31}$$

substituting back in (3.28),

$$\begin{aligned}\|\beta\|_{L^\infty}\|w\|_{L^2}^2 &\le \mathcal{C}\widetilde{\mathcal{C}_p}\mathrm{diam}(\Omega)\|w\|_V\|\beta\cdot\nabla w\|_{V^*}\\ &= \mathcal{C}\widetilde{\mathcal{C}_p}\mathrm{diam}(\Omega)\|\beta\cdot\nabla w\|_{C_0}\|\beta\cdot\nabla w\|_{C_1}.\end{aligned} \tag{3.32}$$

It has been proven in [16] (see also Lemma (a) in [20, section 1.10.1]) that when a linear operator $L : C_0 \cap C_1 \to E$ satisfies $\|L\phi\|_E \le \|\phi\|_{C_0}^{1/2}\|\phi\|_{C_1}^{1/2}$, for all $\phi \in C_0 \cap C_1$, it also satisfies $\|L\phi\|_E \le \|\phi\|_{(C_0,C_1)_{1/2,1}}$; using this in (3.32), with $L = (\beta\cdot\nabla)^{-1}$ and $\phi = \beta\cdot\nabla w$, we get

$$\mathcal{C}_p\|\beta\|_{L^\infty}^{1/2}\mathrm{diam}(\Omega)^{-1/2}\|w\|_{L^2} \le \|\beta\cdot\nabla w\|_{(C_0,C_1)_{1/2,1}} \quad \forall w \in V, \tag{3.33}$$

which is the left inequality of (3.27).

We have, thanks to the theorem in [20, section 1.3.3],

$$\begin{aligned}\|\beta\cdot\nabla w\|_{(C_0,C_1)_{1/2,1}}^2 &\le \|\beta\cdot\nabla w\|_{C_0}\|\beta\cdot\nabla w\|_{C_1}\\ &\le \kappa^{1/2}|w|_{H^1}\|\beta\cdot\nabla w\|_{V^*}\\ &\quad + \rho^{1/2}\|w\|_{L^2}\|\beta\cdot\nabla w\|_{V^*}\end{aligned} \tag{3.34}$$

and

$$\begin{aligned}\|\beta\cdot\nabla w\|_{V^*} \le \kappa^{-1/2}\|\beta\cdot\nabla w\|_{H^{-1}} \le \kappa^{-1/2}\|\beta\|_{L^\infty}\|w\|_{L^2},\\ \|\beta\cdot\nabla w\|_{V^*} \le \rho^{-1/2}\|\beta\cdot\nabla w\|_{L^2} \le \rho^{-1/2}\|\beta\|_{L^\infty}|w|_{H^1};\end{aligned} \tag{3.35}$$

from (3.34)–(3.35) we get

$$\|\beta\cdot\nabla w\|_{(C_0,C_1)_{1/2,1}}^2 \le 2\|\beta\|_{L^\infty}|w|_{H^1}\|w\|_{L^2}.$$

Still using the result of [16] mentioned above, for (3.32)–(3.33) we get

$$\|\beta\cdot\nabla w\|_{(C_0,C_1)_{1/2,1}} \le \mathcal{C}\|\beta\|_{L^\infty}^{1/2}\|w\|_{(L^2,H_0^1)_{1/2,1}} \quad \forall w \in V,$$

which concludes the proof of (3.27). ∎

In the previous proposition, we have shown uniform bounds (with respect to the operator coefficients) for $\|\beta\cdot\nabla w\|_{(C_0,C_1)_{1/2,1}}$. As a general result of the interpolation theory (see, e.g., [20, 1.3.3.d]), we have

$$\|\beta\cdot\nabla w\|_{(C_0,C_1)_{1/2,2}} \le \|\beta\cdot\nabla w\|_{(C_0,C_1)_{1/2,1}} \quad \forall w \in V, \tag{3.36}$$

and similarly

$$\|w\|_{(A_0,A_1)_{1/2,2}} \le \|w\|_{(A_0,A_1)_{1/2,1}} \quad \forall w \in V. \tag{3.37}$$

The converse inequality of (3.36), that is, $\|\beta\cdot\nabla w\|_{(C_0,C_1)_{1/2,1}} \le \mathcal{C}\|\beta\cdot\nabla w\|_{(C_0,C_1)_{1/2,2}}$, does not hold true; on the other hand the converse of (3.37) holds true, and it is,

roughly speaking, *almost* uniform in the sense that the constant in it only depends on a logarithm of the coefficients, as stated in the next proposition.

PROPOSITION 3.5. *Consider the case* (3.22), (3.24), *and* (3.25)*: let*

$$\alpha := \max\left\{\kappa^{1/2}\rho^{1/2}, \kappa\, diam(\Omega)\right\} / \|\beta\|_{L^\infty}. \tag{3.38}$$

When $\alpha \leq 1$ *we have*

$$\|w\|_{(A_0,A_1)_{1/2,1}} \leq \left(\mathcal{C} - \log^{1/2}(\alpha)\right) \|w\|_{(A_0,A_1)_{1/2,2}} \quad \forall w \in V, \tag{3.39}$$

while for $\alpha > 1$ *we have*

$$\|w\|_{(A_0,A_1)_{1/2,1}} \leq \mathcal{C}\|w\|_{(A_0,A_1)_{1/2,2}} \quad \forall w \in V. \tag{3.40}$$

Proof. We only consider here the case $\alpha \leq 1$, since when $\alpha > 1$ we can set $\alpha := 1$ instead of (3.38) and follow the proof. Recall that from the definition (2.6) we have

$$\begin{aligned} \|w\|_{A_0} &\leq \|w\|_{A_0} \quad \forall w \in V, \\ \|w\|_{A_0} &\leq \|w\|_{A_1} \quad \forall w \in V, \end{aligned} \tag{3.41}$$

and from (3.35) and the Poincaré inequality, we also have

$$\begin{aligned} \|w\|_{A_1} &\leq \|w\|_{A_1} \quad \forall w \in V, \\ \alpha\|w\|_{A_1} &\leq \mathcal{C}\|w\|_{A_0} \quad \forall w \in V. \end{aligned} \tag{3.42}$$

Then, by interpolation we get from (3.41)

$$\|w\|_{A_0} \leq \|w\|_{(A_0,A_1)_{1/2,2}} \quad \forall w \in V, \tag{3.43}$$

and from (3.42)

$$\alpha^{1/2}\|w\|_{A_1} \leq \mathcal{C}\|w\|_{(A_0,A_1)_{1/2,2}} \quad \forall w \in V. \tag{3.44}$$

By the definition (2.15) and by the triangle inequality we get

$$\begin{aligned} \|w\|_{(A_0,A_1)_{1/2,1}} &\leq \int_0^{+\infty} \left(t^{-1/2}\|w_0(t)\|_{A_0} + t^{1/2}\|w_1(t)\|_{A_1}\right) \frac{dt}{t} \\ &\leq \int_0^{\alpha} \left(t^{-1/2}\|w_0(t)\|_{A_0} + t^{1/2}\|w_1(t)\|_{A_1}\right) \frac{dt}{t} \\ &\quad + \int_{\alpha}^{1} \left(t^{-1/2}\|w_0(t)\|_{A_0} + t^{1/2}\|w_1(t)\|_{A_1}\right) \frac{dt}{t} \\ &\quad + \int_{1}^{+\infty} \left(t^{-1/2}\|w_0(t)\|_{A_0} + t^{1/2}\|w_1(t)\|_{A_1}\right) \frac{dt}{t} \\ &= I + II + III \end{aligned}$$

for any $w_0(t)$ and $w_1(t)$ with $w = w_0(t) + w_1(t)$, $w_i(t) \in V, i = 1, 2$ and $0 < t < +\infty$. Taking $w_0(t) = w$ and $w_1(t) = 0$ for $t \geq 1$ and using (3.43) we have

$$\begin{aligned} III &\leq \|w\|_{A_0} \int_1^{\infty} t^{-3/2} dt \\ &\leq 2\|w\|_{A_0} \\ &\leq 2\|w\|_{(A_0,A_1)_{1/2,2}}. \end{aligned}$$

In a very similar way, we deal with the first term, taking $w_1(t) = w$ and $w_0(t) = 0$ for $0 < t < \alpha$; thanks to (3.44) we obtain

$$\begin{aligned} I &\le \|w\|_{A_1} \int_0^\alpha t^{-1/2} dt \\ &\le 2\alpha^{1/2} \|w\|_{A_1} \\ &\le \mathcal{C} \|w\|_{(A_0,A_1)_{1/2,2}}. \end{aligned}$$

Thanks to the Cauchy–Schwarz inequality we have

$$\begin{aligned} &\int_\alpha^1 \left(t^{-1/2} \|w_0(t)\|_{A_0} + t^{1/2} \|w_1(t)\|_{A_1} \right) \frac{dt}{t} \\ &\quad \le \left[\int_\alpha^1 \frac{dt}{t} \right]^{1/2} \\ &\qquad \cdot \left[\int_\alpha^1 \left(t^{-1/2} \|w_0(t)\|_{A_0} + t^{1/2} \|w_1(t)\|_{A_1} \right)^2 \frac{dt}{t} \right]^{1/2} \\ &\quad \le [-\log(\alpha)]^{1/2} \\ &\qquad \cdot \left[\int_\alpha^1 \left(t^{-1/2} \|w_0(t)\|_{A_0} + t^{1/2} \|w_1(t)\|_{A_1} \right)^2 \frac{dt}{t} \right]^{1/2}, \end{aligned} \tag{3.45}$$

which holds true for any choice of $w_0(t)$ and $w_1(t)$ on $\alpha < t < 1$. Taking the infimum on w_0, w_1 we obtain

$$II \le [-\log(\alpha)]^{1/2} \|w\|_{(A_0,A_1)_{1/2,2}}.$$

Finally, (3.39) follows from the previous estimates on I, II, and III. ∎

From Propositions 3.4 and 3.5 we easily derive the next *almost* uniform bounds (up to a $\log(\alpha)^{1/2}$ factor, which is, roughly speaking, a *weak* loss of uniformity).

COROLLARY 3.6. *For the case* (3.22), (3.24)–(3.25), *given* α *from* (3.38), *we have*

$$\mathcal{C}_p \min\left\{1, |\log(\alpha)|^{-1/2}\right\} diam(\Omega)^{-1/2} \|\beta\|_{L^\infty}^{1/2} \|w\|_{L^2} \le |\!|\!| w |\!|\!| \quad \forall w \in V, \tag{3.46}$$

$$\|\beta \cdot \nabla w\|_{(C_0,C_1)_{1/2,2}} \le \mathcal{C} \|\beta\|_{L^\infty}^{1/2} \|w\|_{(L^2,H_0^1)_{1/2,1}} \quad \forall w \in V, \tag{3.47}$$

where $\mathcal{C}_p$ *depends on* $\beta/\|\beta\|_{L^\infty}$ *and (the shape of)* Ω.

Though (3.46)–(3.47) are not as sharp as the estimates we got in section 3.1 for the one-dimensional case, they put in evidence the relationship between the norm $|\!|\!| \cdot |\!|\!|$ defined in (3.26) and the skew-symmetric part $\mathcal{L}_{\text{skew}} = \beta \cdot \nabla$ of (1.1). Recall that $\max\{\kappa^{1/2} \text{diam}(\Omega)^{-1}, \rho^{1/2}\} \|w\|_{L^2} \le \mathcal{C} \|w\|_V \le \mathcal{C} |\!|\!| w |\!|\!|$, while (3.46) states the bound on the L^2-norm which is mainly due to $\|\beta \cdot \nabla w\|_{(C_0,C_1)_{1/2,2}}$. Then (3.46) becomes relevant when κ and ρ are small.

REFERENCES

[1] I. BABUŠKA AND A. K. AZIZ, *Survey lectures on the mathematical foundations of the finite element method*, in The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, Academic Press, New York, 1972, pp. 1–359.

[2] A. Barinka, T. Barsch, P. Charton, A. Cohen, S. Dahlke, W. Dahmen, and K. Urban, *Adaptive wavelet schemes for elliptic problems—implementation and numerical experiments*, SIAM J. Sci. Comput., 23 (2001), pp. 910–939.

[3] S. Berrone and C. Canuto, *Multilevel A Posteriori Error Analysis for Reaction-Convection-Diffusion Problems*, Tech. Report Preprint 18/2002, Dipartimento di Matematica, Politecnico di Torino, Turin, Italy, 2002.

[4] S. Bertoluzza, C. Canuto, and A. Tabacco, *Stable discretizations of convection-diffusion problems via computable negative-order inner products*, SIAM J. Numer. Anal., 38 (2000), pp. 1034–1055.

[5] S. Bertoluzza and M. Verani, *Convergence of a nonlinear wavelet algorithm for the solution of PDEs*, Appl. Math. Lett., 16 (2003), pp. 113–118.

[6] J. H. Bramble, R. D. Lazarov, and J. E. Pasciak, *Least-squares for second-order elliptic problems*, Comput. Methods Appl. Mech. Engrg., 152 (1998), pp. 195–210.

[7] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.

[8] A. N. Brooks and T. J. R. Hughes, *Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 32 (1982), pp. 199–259.

[9] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, Classics Appl. Math. 40, SIAM, Philadelphia, 2002.

[10] A. Cohen, W. Dahmen, and R. DeVore, *Adaptive wavelet methods for elliptic operator equations: convergence rates*, Math. Comp., 70 (2001), pp. 27–75.

[11] A. Cohen, W. Dahmen, and R. DeVore, *Adaptive wavelet methods.* II. *Beyond the elliptic case*, Found. Comput. Math., 2 (2002), pp. 203–245.

[12] W. Dahmen, A. Kunoth, and R. Schneider, *Wavelet least squares methods for boundary value problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1985–2013.

[13] W. Dörfler, *Uniform a priori estimates for singularly perturbed elliptic equations in multi-dimensions*, SIAM J. Numer. Anal., 36 (1999), pp. 1878–1900.

[14] H. Goering, A. Felgenhauer, G. Lube, H.-G. Roos, and L. Tobiska, *Singularly Perturbed Differential Equations*, Reihe Math. Research 13, Akademie-Verlag, Berlin, 1983.

[15] J.-L. Lions and E. Magenes, *Non-homogeneous Boundary Value Problems and Applications.* Vol. I, Springer-Verlag, New York, 1972.

[16] J.-L. Lions and J. Peetre, *Sur une classe d'espaces d'interpolation*, Inst. Hautes Études Sci. Publ. Math., 19 (1964), pp. 5–68.

[17] H.-G. Roos, M. Stynes, and L. Tobiska, *Numerical Methods for Singularly Perturbed Differential Equations*, Springer-Verlag, Berlin, 1996.

[18] G. Sangalli, *Analysis of the advection-diffusion operator using fractional order norms*, Numer. Math., 49 (2004), pp. 149–162.

[19] G. Sangalli, *Quasi optimality of the SUPG method for the one-dimensional advection-diffusion problem*, SIAM J. Numer. Anal., 41 (2003), pp. 1528–1542.

[20] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, 2nd ed., Johann Ambrosius Barth, Heidelberg, 1995.